Conversão de Voz Inter-Linguística

Anderson Fraiha Machado

Tese apresentada
ao
Instituto de Matemática e Estatística
da
Universidade de São Paulo
para
obtenção do título
de
Doutor em Ciências

Programa: Ciência da Computação

Orientador: Prof. Dr. Marcelo Queiroz

Coorientador: Prof. Dr. Antonio Bonafonte

Durante o desenvolvimento deste trabalho o autor recebeu auxílio financeiro da CAPES

São Paulo, maio de 2013

Conversão de Voz Inter-Linguística

Esta versão da tese contém as correções e alterações sugeridas pela Comissão Julgadora durante a defesa da versão original do trabalho, realizada em 21/05/2013. Uma cópia da versão original está disponível no Instituto de Matemática e Estatística da Universidade de São Paulo.

Comissão Julgadora:

- Prof. Dr. Marcelo Gomes de Queiroz (orientador) - IME-USP
- Prof. Dr. Fabio Kon - IME-USP
- Prof. Dr. Flávio Soares Corrêa da Silva - IME-USP
- Prof. Dr. Fernando Henrique de Oliveira Iazzetta - ECA-USP
- Prof. Dr. Miguel Arjona Ramirez - EP-USP

Agradecimentos

Toti Deo Gloria...

Gostaria de poder agradecer a todos que me ajudaram de forma direta e indireta neste trabalho. Mas infelizmente, neste caso, teríamos conteúdo suficiente para redigir outra tese como esta apresentada. Sendo assim, vou apenas destacar alguns personagens que marcaram minha trajetória como doutorando.

Primeiramente, gostaria de exaltar o nome do Deus Todo-Poderoso a quem sirvo e dedico esta obra. A Aquele que é, que foi, e que sempre será minha fonte de inspiração e motivo de transpiração. Muito obrigado pelas conquistas até aqui alcançadas!!!

Também, expresso minha profunda gratidão ao meu orientador, Marcelo Queiroz, pelo exemplo de inteligência e serenidade. Sentirei falta das conversas divertidas e do conteúdo altamente instrutivo apresentado por ele. Muito obrigado Marcelo por cada conselho e incentivo... por ter feito dessa jornada muito mais que um aprendizado técnico.

Igualmente importante, eu quero agradecer a hospitalidade demonstrada pelo meu coorientador Antônio Bonafonte, quem me recebeu em Barcelona como um membro de sua família. *Gracias Toni, por todas las enseñanzas: verdaderas lecciones de vida (además de las técnicas y métodos extraordinários). Agradezco a Dios por el ejemplo de mansedumbre y compañerismo. Que Dios te bendiga.* Também, agradeço à CAPES pela oportunidade de realizar meu doutorado-sanduíche de um ano na Espanha entre os anos de 2011 e 2012.

Não poderia deixar de agradecer minha preciosa família: os grandes exemplos de força e determinação de meu pai, Herculano Machado Neto, me incentivaram a lutar e correr atrás dos sonhos e projetos com determinação, garra e coragem, guardando sempre a honra, a honestidade, a humildade e a dignidade. Também não posso esquecer de minha mãe Ana Luiza Fraiha Machado, de quem herdei uma parte preciosa de minha personalidade e valores. Mulher corajosa, lutadora perpétua, um exemplo de sensibilidade e sabedoria, forjada nas muitas lutas e dificuldades intransponíveis aos olhos humanos. Muito obrigado mãe, por me fazer saber qual a verdadeira rocha sobre a qual devemos ser edificados. Muito obrigado meus irmãos Erick, Pri, Lely e Myle por todo apoio e carinho demonstrado a cada segundo de minha vida. Amo a todos vocês.

Igualmente, dedico esta tese a toda minha família estendida, em especial, à família Bonfim Córdoba, que escolheu me acolher e receber como filho amado aqui em São Paulo, não sendo eu merecedor de tal honra. Expresso na mesma medida, minha gratidão aos irmãos da Liber que me fizeram tão bem toda esta caminhada. Em especial, um abraço gigante para minha banda do coração, a banda EXCORDE, por me acolher tão gentilmente aqui em SAMPA.

Um agradecimento especial ao meu amigo-irmão Edwin Ferraz pelo tempo bom que passamos juntos em seu apartamento: *Valeu gurizão*!!! Aos amigos latinos-gaúchos de SANCA, de quem sinto

tanta saudade, manifesto minha grande estima. ;-)

Um agradecimento todo especial aos meus irmãos de Barcelona, por haver me aguentado um ano e me recebido tão bem. Em especial, um grande abraço às quatro famílias mais significantes para minha passagem por esta terra tão iluminada: Famílias Sousa, Bassols, Gonzalez e Oakley. Cada um de vocês ocupam um lugar todo especial dentro de mim. Muchas gracias my people.

Espero poder retribuir a cada um de vocês o carinho e dedicação a mim demonstrados durante todos estes anos.

Finalizo estas páginas de gratidão com as palavras do ilustríssimo pensador Antoine de Saint-Exupery: *"Aqueles que passam por nós não vão sós. Deixam um pouco de si, levam um pouco de nós."*

D+ ;P 42[1]

[1] Essa é pra você LL.

Resumo

A conversão de voz é um problema emergente em processamento de fala e voz com um crescente interesse comercial, tanto em aplicações como Tradução Fala para Fala (*Speech-to-Speech Translation* - SST) e em sistemas *Text-To-Speech* (TTS) personalizados. Um sistema de Conversão de Voz deve permitir o mapeamento de características acústicas de sentenças pronunciadas por um falante origem para valores correspondentes da voz do falante destino, de modo que a saída processada é percebida como uma sentença pronunciada pelo falante destino. Nas últimas duas décadas, o número de contribuições científicas relacionadas ao problema de conversão de voz tem crescido consideravelmente, e um panorama sólido do processo histórico, assim como de técnicas propostas são indispensáveis para contribuição neste campo. O objetivo deste trabalho é realizar um levantamento geral das técnicas utilizadas para resolver o problema, apontando vantagens e desvantagens de cada método, e a partir deste estudo, desenvolver novas ferramentas. Dentre as contribuições do trabalho, foram desenvolvidos um método para decomposição espectral em termos de bases radiais, mapas fonéticos artificiais, agrupamentos k-verossímeis, funções de empenamento em frequência entre outras, com o intuito de implementar um sistema de conversão de voz inter-linguístico independente de texto de alta qualidade.

Palavras-chave: Conversão de voz, Conversão Inter-linguística.

Abstract

Voice conversion is an emergent problem in voice and speech processing with increasing commercial interest, due to applications such as Speech-to-Speech Translation (SST) and personalized Text-To-Speech (TTS) systems. A Voice Conversion system should allow the mapping of acoustical features of sentences pronounced by a source speaker to values corresponding to the voice of a target speaker, in such a way that the processed output is perceived as a sentence uttered by the target speaker. In the last two decades the number of scientific contributions to the voice conversion problem has grown considerably, and a solid overview of the historical process as well as of the proposed techniques is indispensable for those willing to contribute to the field. The goal of this work is to provide a critical survey that combines historical presentation to technical discussion while pointing out advantages and drawbacks of each technique, and from this study, to develop new tools. Some contributions proposed in this work include a method for spectral decomposition in terms of radial basis functions, artificial phonetic map, warping functions among others, in order to implement a text-independent crosslingual voice conversion system of high quality.
Keywords: Voice Conversion, Cross-Lingual Voice Conversion.

Sumário

Lista de Abreviaturas xi

Lista de Símbolos xiii

Lista de Figuras xv

Lista de Tabelas xvii

1 Introdução 1

1.1 Considerações Preliminares . . . 1
1.1.1 Introdução à Conversão de Voz . . . 2
1.1.2 Motivação . . . 4
1.1.3 Aplicações . . . 4
1.2 Visão Geral de um Sistema de Conversão Típico . . . 5
1.2.1 Treinamento . . . 6
1.2.2 Transformação . . . 6
1.3 Estado da Arte em Conversão de Voz . . . 7
1.3.1 Visão Histórica das Técnicas de Conversão de Voz . . . 7
1.3.2 Classificação das Técnicas de Conversão de Voz . . . 8
1.4 Limitações e Desafios . . . 10
1.5 Objetivos . . . 13
1.6 Contribuições . . . 13
1.7 Organização do Trabalho . . . 14

2 Ferramentas e Métodos 15

2.1 Rudimentos do Processamento de Voz . . . 15
2.2 Modelos Clássicos do Sinal de Voz . . . 21
2.2.1 Modelo Fonte-Filtro . . . 21
2.2.2 Modelo Harmônico-Estocástico . . . 34
2.3 Tópicos em Conversão de Prosódia . . . 47
2.3.1 Modelagem da Prosódia . . . 48
2.3.2 Transformação da Prosódia . . . 50
2.4 Técnicas de Transformação Espectral . . . 53
2.4.1 Técnicas de Classificação . . . 54
2.4.2 Técnicas de Aproximação . . . 59

3 Conversor de Voz Inter-Linguístico **67**
3.1 Visão Estrutural do Sistema . . . 68
3.1.1 Fase de Treinamento . . . 69
3.1.2 Fase de Conversão . . . 71
3.2 Estágio I: Modelagem Harmônico-Estocástica . . . 74
3.2.1 Modelagem da Configuração Fásica . . . 76
3.2.2 Implementação . . . 79
3.3 Estágio II: Decomposição Paramétrica . . . 84
3.3.1 Estimação do Envelope Espectral . . . 88
3.3.2 Decomposição Espectral em Funções de Base . . . 91
3.4 Estágio III: Clusterização dos Dados . . . 101
3.4.1 Clusterização k-Verossímil . . . 104
3.4.2 Modelagem do Mapa Fonético Artificial . . . 107
3.5 Estágio IV: Alinhamento dos Corpora Não-Paralelos . . . 119
3.5.1 O Problema da Correspondência Fonética . . . 121
3.5.2 Mapeamento usando Bases da Teoria dos Grafos . . . 124
3.6 Estágio V: Transformação dos Parâmetros Acústicos . . . 128
3.6.1 Transformações Globais ao Nível da Sentença . . . 129
3.6.2 Transformações Locais ao Nível do Segmento . . . 131

4 Resultados Experimentais **141**
4.1 Definições Gerais . . . 141
4.2 Avaliação Objetiva . . . 143
4.2.1 Parametrização e Quantização . . . 143
4.2.2 Clusterização e Mapeamento . . . 146
4.2.3 Transformação Espectral . . . 151
4.3 Avaliação Subjetiva . . . 159
4.3.1 Teste Perceptual . . . 160
4.3.2 Discussão . . . 162

5 Conclusões **167**
5.1 Considerações Finais . . . 167
5.2 Trabalhos Aceitos em Anais de Congressos . . . 168
5.3 Sugestões para Pesquisas Futuras . . . 169

A Experimentos Preliminares **173**
A.1 Experimento I . . . 173
A.2 Experimento II . . . 174
A.3 Experimento III . . . 175

B Conceitos Básicos Complementares **177**
B.1 Linguística Aplicada à Localização dos Formantes . . . 177
B.2 Bases da Teoria de Grafos . . . 178
B.3 Bases da Morfologia Matemática . . . 179

Referências Bibliográficas **183**

Lista de Abreviaturas

ABX	Teste Perceptual de Múltiplas Escolhas
CFA	Classe Fonética Artificial
DEA	Distribuição Energética Acumulada
DFN	Distribuição em Frequência Normalizada
DFT	Transformada Discreta de Fourier (*Discrete Fourier Transform*)
FFT	Transformada de Fourier Rápida (*Fast Fourier Transform*)
GMM	Modelo de Misturas Gaussianas (*Gaussian Mixture Model*)
HMM	Modelo Oculto de Markov (*Hidden Markov Model*)
HSM	Modelo Harmônico-Estocástico (*Harmonic plus Stochastic Model*)
LPC	Coeficientes de Predição Linear (*Linear Predictive Coding*)
LT-DIAG	Transformação Linear usando somente matrizes de covariância Diagonal
LT-FULL	Transformação Linear usando matrizes de covariância completas
MQO	Mínimos Quadrados Ordinários
MFA	Análise Multi-Frames (*Multi-Frame Analysis*)
MFCC	Coeficientes Mel-Cepstrais (*Mel-Frequency Cepstral Coefficients*)
MOS	Teste de Opinião Média (*Mean Opinion Score*)
NFW	Deformação em Frequência Normalizada (*Normalized Frequency Warping*)
PSOLA	*Pitch Synchronous Overlap and Add*
RBF	Funções de Base Radial (*Radial Basis Functions*)
STASC	*Speaker Transformation Algorithm using Segmental Codebooks*
STRAIGHT	*Speech Transformation and Representation by Adaptive Interpolation of weiGHTed*
VTTF	Função de Transferência do Trato Vocal (*Vocal-Tract Transfer Function*)

Lista de Símbolos

ω	Frequência Angular
f_0	Frequência Fundamental
$[a, \mu, \sigma]$	Tupla composta pela Amplitude, Frequência Central e Largura de Banda de uma Base
Ψ_0	Frequência Fundamental
R	Taxa de Amostragem
N_w	Tamanho dos Frames
N_{adv}	Avanço da Janela de Segmentação
L^*	Ordem dos Vetores Quantizados
W_{hamming}	Janela de Segmentação
F_c	Frequência de Corte
ϵ	Menor Valor Considerável no Sistema
B_f	Largura de Banda dos Filtros usados no Mapa Fonético
M_f	Taxa de Amostragem do Mapa Fonético
M_c	Número de Classes usadas na Transformação
Ψ^k	Vetor de Características Acústicas do k-ésimo Segmento de Voz

Lista de Figuras

1.1 Conversão de voz entre falantes quaisquer. 3
1.2 Processo geral na conversão de voz. 6

2.1 O aparelho auditivo. 17
2.2 Parte superior do aparelho fonador. 19
2.3 Processo de produção do pulso glotal. 20
2.4 Modelo de geração do som irradiado pelo sistema fonador. 22
2.5 Modelagem do processo de produção da voz por um sistema fonte-filtro. 23
2.6 Visualização do envelope espectral da vogal [a]. 24
2.7 Dois exemplos de estimativa do envelope espectral utilizando coeficientes LPC. . . . 26
2.8 (a) Lugar dos zeros dos polinômios $H(z)$, $P(z)$ e $Q(z)$ na circunferência de raio unitário; (b) Módulo da resposta em frequência do filtro $H(z)^{-1}$ IIR com 6 polos e as posições angulares dos zeros de $P(z)$ e $Q(z)$. 27
2.9 Visualização dos formantes representados a partir de centroides e largura de banda dos polos. 28
2.10 Acima, o log do espectro de um sinal de voz; Abaixo, o cepstrum deste sinal de voz. 30
2.11 Diagrama de blocos para obtenção do MEL-cepstrum. 31
2.12 O fluxo glotal U_g e sua respectiva derivada dU_g. 33
2.13 O diagrama do Modelo Senoidal. 35
2.14 O diagrama do Modelo Harmônico. 37
2.15 Decomposição do segmento s em partes Harmônica (s_h) e Estocástica (s_s) no domínio temporal. 42
2.16 Decomposição espectral do segmento S em trem de pulsos Harmônicos (S_h) e espectro Estocástico (S_s). 44
2.17 As componentes Harmônica e Estocástica de um sinal de voz. 45
2.18 Ilustração do método Overlap-Add. 47
2.19 Contorno de pitch e energia do sinal de voz. 49
2.20 Dilatação (esquerda) e compressão (direita) na escala temporal de um sinal de voz. . 52
2.21 Diminuição (esquerda) e elevação (direita) do pitch de um sinal de voz. 52
2.22 As duas fases do processo de conversão de parâmetros acústicos. 53
2.23 Ajuste linear a partir de um conjunto de pontos no espaço Euclideano. 61
2.24 As duas fases do processo de conversão de parâmetros acústicos. 64

3.1 Visão do sistema sob o ponto de vista de fases de Treinamento e Conversão. 68
3.2 Visão do sistema sob o ponto de vista de fases de treinamento e transformação. . . . 69

3.3 Espalhamento espectral de algumas janelas de segmentação. 73
3.4 Visão do sistema dividido em módulos funcionais. 74
3.5 Módulo de Análise (**A**) e Síntese (**S**) 75
3.6 Evolução temporal dos segmentos sem o termo de fase linear. 79
3.7 O espectro de fases interpolado. 80
3.8 Fase de Parametrização (**P**) e sua respectiva Inversa ($\mathbf{P}^{-1}$) 84
3.9 Reconstrução da função constante $I(x) = 1$ usando diversos interpoladores radiais. . 87
3.10 Diversos interpoladores espectrais usando o algoritmo `interpolador`. 89
3.11 Análise multi-frame usando alguns do envelopadores propostos. 90
3.12 O espectro harmônico de amplitudes envelopado pelo Interpolador Gaussiano. 91
3.13 Exemplo de estimação harmônica com 8, 16, 24 e 32 bases (Nuttall) 98
3.14 Fase de Clusterização dos dados (**C**) em classes fonéticas artificiais 102
3.15 Posições do primeiro e segundo formantes (em Hertz) para diversas vogais. 107
3.16 Quantização do mapa fonético em 9×15 bandas. 108
3.17 Distribuição formântica (esq) e sua versão normalizada (dir) para cada vogal espanhola em um fragmento de voz. 111
3.18 Mapas de distribuição formântica extrapolados. 113
3.19 Exemplo de modelagem por desvio padrão lateral. 118
3.20 Fase de Mapeamento de Classes Fonéticas (**M**) 119
3.21 Mapeamento entre classes fonéticas incompatíveis. 122
3.22 Normalização do mapa fonético para fase de alinhamento. 123
3.23 Relação existente entre o problema do mapeamento ótimo e o emparelhamento perfeito. 124
3.24 O método de alinhamento aplicado ao alinhamento de sentenças paralelas. 126
3.25 O método de alinhamento aplicado ao alinhamento de sentenças paralelas. 128
3.26 Definição da função de Transformação (**T**) 129
3.27 Conversão de um segmento de voz usando apenas parâmetros globais. 131
3.28 Relação entre a função energética acumulada e o espectro original. 136
3.29 Conversão local usando deformação (empenamento) em frequência normalizada. 138

4.1 Distorções Espectrais e Taxas de Descontinuidade do Sinal de Entrada 145
4.2 Discriminantes de Fisher dos coeficientes espectrais e suas versões adaptadas à abordagem MFCC. 146
4.3 Principais regiões formânticas dos corpus origem 'ES_75' e destino 'ES_76'. 147
4.4 Principais regiões formânticas do corpus 'ES_79'. 149
4.5 Dados da quarta coluna da Tabela 4.2 exibidos em escala logarítmica. 155
4.6 Taxa de distorção espectral média da conversão entre cada par de corpora indicado. . 157
4.7 Visualização espectral das sentenças origem, destino e transformada. 158
4.8 Visualização das taxas de distorção espectral relacionando as sentenças origem, destino e transformada alinhadas no tempo. 159

A.1 Resultado perceptual dos testes de configuração fásica. 175

B.1 Principais operadores morfológicos. 180

Lista de Tabelas

1.1 Relação de trabalhos em conversão de voz típica, independente de texto* e inter-linguística◇. 11

2.1 O primeiro e a segundo formante das vogais comuns na língua portuguesa. 20
2.2 Larguras de banda do lóbulo central e secundários de janelas clássicas 39

3.1 Variáveis globais do sistema e valores default 73
3.2 *Alguns coeficientes d_m de cada função base.* 93

4.1 Indicadores de alinhamento entre classes fonéticas artificiais (CFA) de ambos os corpora origem e destino. 148
4.2 Taxas de erro do módulo de seleção considerando a função distância. 154
4.3 Resultado completo da entrevista perceptual (*MOS*). 161
4.4 Resultado médio da entrevista perceptual (*MOS*). 162
4.5 Relação de resultados experimentais de Opinião Média de Acerto – *MOS*. 163
4.6 Relação de resultados experimentais utilizando método *ABX*. 164

A.1 *A média do EMQ entre o envelope estimado e cada vetor de amplitude harmônica do conjunto fonético.* 173
A.2 Taxas de erro do módulo de seleção considerando o número de coeficientes M_c e o tipo de normalização. 176

B.1 Fonemas adaptados ao contexto da língua portuguesa, espanhola e inglesa. 179

Capítulo 1

Introdução

A **Fala** é uma ferramenta de comunicação inerente ao ser humano. Ela corresponde à capacidade de emitir sons de acordo com um padrão, a fim de permitir uma comunicação eficiente entre indivíduos distintos. Tal padrão, conhecido com língua, deve ser compreendido por ambas as partes comunicantes: o emissor e o receptor da fala. Os sons produzidos pelo emissor da fala são organizados em fonemas. Esses fonemas são invariantes da fala, isto é, são reconhecíveis por qualquer falante de uma respectiva língua no mundo e são essenciais na composição do conteúdo da informação comunicada, que são as **sentenças**.

Com a fala, podemos expressar não só sentenças, mas também emoções, sensações, ênfases e outros aspectos, como por exemplo melodias e sotaques regionais de cada indivíduo. Todos esses aspectos juntos contribuem para a formação de uma identidade sonora. A identidade sonora da fala, ou simplesmente identidade da fala, é um conjunto de parâmetros acústicos temporais e espectrais (tais como conteúdo formântico e prosódia) suficientes para caracterizar e distinguir um falante dentre vários outros. Este conjunto de parâmetros acústicos contribuem significativamente para caracterização do **timbre** de voz do indivíduo.

Como objeto de estudo deste trabalho, a **Conversão de Voz** corresponde essencialmente à transferência de parâmetros característicos da identidade da fala de um falante para outro. A Seção 1.1 introduz alguns conceitos relacionados e tem por objetivo apresentar conceitos gerais sobre conversão de voz; a Seção 1.2 se dedica a apresentar uma visão geral de um sistema de conversão de voz genérico; uma breve varredura da literatura sobre conversão de voz é realizada na Seção 1.3; a partir dos problemas observados na literatura (Seção 1.4), este capítulo segue apresentando os objetivos (Seção 1.5) da tese, bem como suas respectivas contribuições (Seção 1.6). Finalmente, uma apresentação geral de cada um dos demais capítulos deste texto é realizada na Seção 1.7.

1.1 Considerações Preliminares

O desenvolvimento de sistemas computacionais que processam a fala de vários modos é um desafio muito importante dentro da disciplina de Processamento de Sinais. Podemos de um modo geral definir a **conversão de voz** entre dois indivíduos como o processo de transformação da identidade da fala de um indivíduo para outro, ou seja, da conversão dos aspectos de timbre e prosódia entre os falantes envolvidos.

1.1.1 Introdução à Conversão de Voz

Na literatura os termos *Voice Transformation*, *Voice Conversion* e *Voice Morphing* aparecem frequentemente, e não devem de modo algum ser confundidos. A transformação da voz (*voice transformation*) engloba todas as tarefas e métodos que modificam qualquer um dos parâmetros da fala; por exemplo, transpor a altura musical de uma voz é um exemplo de transformação de voz. Em particular, as manipulações específicas denominadas *voice morphing* e *voice conversion* são métodos de transformação de voz.

Voice Morphing é um termo proveniente da disciplina de processamento de imagens, e é um caso especial da transformação de voz onde duas vozes são mescladas criando uma terceira voz (virtual), na qual usualmente dois falantes pronunciam ou cantam uma mesma coisa ao mesmo tempo.

Voice Conversion se refere ao processo de conversão de parâmetros acústicos que transformam sentenças pronunciadas por um falante origem (no Inglês, *source speaker*) em sentenças que pareçam terem sidos pronunciadas por um falante destino (no inglês, *target speaker*), conforme a formulação original do problema proposta em 1985 por Childers e co-autores [33].

O que diferencia substancialmente esses dois processos é o fato de que a conversão de voz (do inglês, *voice conversion*) é um método de mapeamento de parâmetros acústicos (timbre e prosódia) de um falante para outro, enquanto que o *voice morphing* produz uma sentença de fala correspondente à mistura das identidades sonoras de dois falantes. Exemplos populares destas técnicas podem ser encontrados em filmes como *"Farinelli"* (onde um soprano e um contratenor são combinados por *morphing* a fim de se obter uma voz equivalente à de um *castratto*) ou *"Alvin e os Esquilos"* (onde as vozes dos atores são moduladas em frequência e têm seus formantes alterados, sugerindo fantasiosamente a imitação de vozes de esquilos).

O principal objetivo de um Sistema de Conversão de Voz (SCV) é portanto modificar a voz de um falante origem de modo que o resultado da transformação fosse percebido como produzido pela voz do falante destino. O falante origem, denominado simplesmente **origem**, corresponde ao locutor que alimenta o sistema e o falante destino, denominado simplesmente **destino**, corresponde ao locutor referencial a quem se deseja relacionar a voz resultante da transformação.

Um sistema de conversão de voz leva em conta o timbre e a prosódia dos falantes origem e destino. Enquanto o timbre e a prosódia são aspectos facilmente reconhecíveis e de difícil definição em termos gerais, num contexto específico de conversão de voz, as características de timbre são usualmente associadas com a envoltória espectral dinâmica do sinal de voz, enquanto a prosódia está associada aos contornos melódicos e de energia, bem como à distribuição rítmica dos fonemas.

Tipos de Conversão

A fim de definir as transformações relacionadas ao timbre e prosódia, os sistemas de conversão de voz normalmente dependem de uma fase de treinamento, que pode ser dependente ou independente de texto. No primeiro caso, tanto o origem quanto o destino gravam as mesmas sentenças, que são então alinhadas no tempo (usando por exemplo o algoritmo *Dynamic Time Warping (DTW)* [15; 45; 60]), e posteriormente têm suas características acústicas mapeadas de modo síncrono.

Na abordagem independente de texto [48], as sentenças pronunciadas pelo falante origem não são necessariamente as mesmas pronunciadas pelo destino. As sentenças gravadas são, em geral, segmentadas em trechos de voz (*frames*) que são mapeados em um espaço de características e agru-

pados em conjuntos de segmentos similares, denominados *classes fonéticas artificiais*, que podem ou não coincidir com as classes fonéticas convencionais da respectiva língua. Os parâmetros acústicos das sentenças do origem são então mapeados para cada classe fonética, de acordo com a similaridade entre segmentos do origem e do destino.

Um outro aspecto distinto em sistemas de conversão de voz está relacionado ao conteúdo fonético das línguas utilizadas na fase de treinamento e na fase de conversão propriamente dita. Na conversão entre falantes de uma mesma língua [290], tanto o falante origem quanto o falante destino são fluentes numa determinada língua L, e tanto o treinamento quanto a transformação são realizados nesta mesma língua. Na Figura 1.1, este caso exige que todas as sentenças pronunciadas pelos falantes pertençam à mesma língua, ou seja, S_0, T_0, S_f e T_f são pronunciadas na língua L. Neste caso, pode-se optar por realizar tanto um treinamento dependente quanto independente de texto.

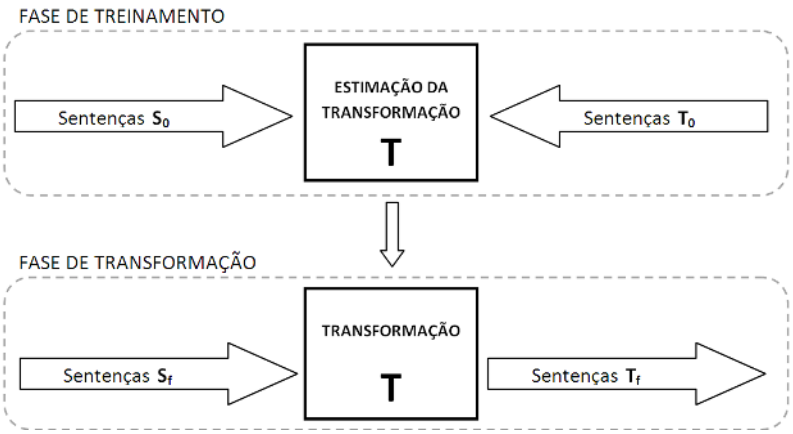

Figura 1.1: *Conversão de voz entre falantes quaisquer.*

Por outro lado, a conversão de voz inter-linguística (do inglês *Cross-lingual Voice Conversion* [49; 247; 274]) pressupõe que os sujeitos origem e destino falam línguas diferentes (L_0 e L_f, respectivamente), e uma sentença (na língua L_0) pronunciada pelo falante origem após a conversão deveria soar como a mesma sentença na língua L_0, porém com a identidade vocal do falante destino (que em geral não fala a língua L_0). É importante ressaltar o fato de que nenhum tipo de processamento simbólico-textual (como a tradução entre L_f e L_0) é pressuposto neste tipo conversão.

A Figura 1.1 contempla também a situação especial de falantes bilíngues. Se o falante origem é fluente tanto em sua língua L_0 (onde produz sentenças S_0) quanto na língua L_f, seu mapa fonético pode servir de ponte na transformação entre suas características fonéticas na língua L_0 original e as características fonéticas do destino na língua L_f. Se, por outro lado, o falante destino é fluente nas duas línguas, ele poderia em princípio registrar as mesmas frases (em L_0) usadas pelo falante origem na fase de treinamento, permitindo o alinhamento temporal dos segmentos para extração de parâmetros acústicos, numa abordagem conhecida como conversão de voz inter-linguística dependente de texto.

Entretanto, o que ocorre é que muitas vezes a fluência dos falantes na segunda língua não é tão boa quanto se espera a fim de se obter uma conversão de voz de alta qualidade. Espera-se desta forma, executar uma *conversão genuína*, onde cada falante pronuncia apenas sentenças em sua língua original. Neste caso, o problema torna-se mais complexo, pois o treinamento torna-se inevitavelmente independente de texto e, além disso, alimentado por sentenças pronunciadas em línguas distintas [241]. Neste caso, na Figura 1.1 as sentenças S_0, S_f e T_f pertencem à língua L_0,

enquanto somente as sentenças pronunciadas pelo falante destino na fase de treinamento (isto é, T_0) pertencem à língua L_f. Tal abordagem é conhecida como conversão de voz inter-linguística independente de texto [243; 244], e corresponde ao foco deste trabalho.

1.1.2 Motivação

O cinema é um ramo do mercado de entretenimento que movimenta bilhões de dólares todo ano. Existe um movimento de democratização na produção cinematográfica, que incentiva a lançar filmes no mercado de produtoras independentes de diversas nacionalidades. Com isso, o crescente lançamento de filmes estrangeiros é inevitável. Para a eficiente distribuição destas produções em um mercado mais amplo seria muito interessante se pudéssemos ouvir as narrativas do filme em diversos idiomas.

Atualmente, a oferta de dubladores não é suficientemente grande para cobrir a variedade de atores dos muitos filmes existentes no mercado. Fatalmente, decorre desta falta de cobertura a alocação de uma mesma voz para vários atores distintos. Além disso, corre-se o risco de um dublador possuir uma boa interpretação vocal, porém, com um timbre de voz muito distante do original. É óbvio que a qualidade do som influencia diretamente na qualidade do filme; assim, uma péssima dublagem desmotiva completamente o espectador a apreciar um filme, mesmo que este último seja de boa qualidade. Neste contexto, o uso de um software que transfere a identidade sonora do ator para o dublador seria de grande interesse. Neste caso, seria necessário o desenvolvimento de um sistema que adaptasse uma dublagem substituindo o timbre de voz de um dublador qualquer por outro timbre o mais próximo quanto possível do timbre de voz do correspondente ator original.

1.1.3 Aplicações

O interesse em resolver o problema de conversão de voz é muito grande, principalmente no ramo comercial. Seguem algumas aplicações para o uso de sistemas de conversão de voz.

Edição Geral de Voz com Caráter Imitativo

Em programas de comédia, é surpreendente a capacidade de algumas pessoas em imitar vozes de outras pessoas. Assimilar a identidade sonora e reproduzir sentenças de voz de outro indivíduo seria útil, principalmente em aplicações no ramo artístico. Dentre estas aplicações se destacam:

- Personificação de sistemas interativos de síntese de fala a partir do texto, os chamados sistemas TTS (do inglês, *Text-To-Speech*): Desenvolver um sistema TTS que é capaz de pronunciar sentenças com vozes de pessoas reais. Recentemente, Duxans [46] e Kain [107; 108] desenvolveram um trabalho sobre a personificação de sistemas TTS interativos;

- Desenvolvimento de um sistema de *karaokê* no qual a voz do cantor amador se converte na voz do cantor original [304];

- Reprodução de vozes de pessoas que já faleceram, permitindo a recriação de cenas a partir da voz de uma outra pessoa qualquer [277; 288].

Dentre os sistemas TTS, destaca-se o sistema FESTIVAL, que propõe a síntese de voz a partir de textos de várias línguas, produzindo sentenças em tempo real que soam com muita naturalidade

[19; 20; 264]. Aplicar uma conversão de voz em sistemas TTS como este aumentaria seu interesse comercial, uma vez que poderíamos personalizar tais sistemas a partir de amostras de vozes de pessoas reais.

Aplicações em Multimídia

Os sistemas de personalização de voz também seriam úteis em sistemas interativos, tais como:

- Ferramentas educacionais para aprendizado de língua estrangeira, que propiciem a demonstração de como seria se o próprio aluno estivesse pronunciado uma sentença em um idioma que ele ainda não fala [184];
- Intérpretes virtuais personalizados, capazes de simultaneamente traduzir uma sentença pronunciada de uma língua para outra, mantendo o timbre de voz do falante original [247].

Os Sistemas Intérpretes Virtuais Personalizados em geral utilizam um conjunto de sistemas responsáveis pelo **reconhecimento**, **tradução** e **síntese** da fala. O campo na computação que realiza pesquisas sobre sistemas de tradução entre línguas distintas é chamado de *Processamento de Linguagem Natural*. Recentemente, alguns projetos contribuíram para o avanço destes temas: o projeto **Verbmobil** [114] teve como objetivo realizar a tradução em tempo real entre os idiomas Alemão/Inglês e Alemão/Japonês[1] por um intérprete virtual em telefonia móvel; também destaca-se o projeto **TC-STAR** [23; 242] (http://tc-star.org), o qual viabilizou resultados contundentes em sistemas de tradução simultânea (*Speech-to-Speech Translation – SST*) entre diversas línguas. Vários outros trabalhos [47; 48; 49; 50; 175; 176; 233; 234; 244] foram desenvolvidos dentro do projeto TC-STAR em parceria com a empresa IBM.

Outras Aplicações

Por fim, existem ainda outras aplicações como:

- Desenvolvimento de sistemas que burlam sistemas de autenticação biométrico de voz, a fim de apontar fragilidades nestes sistemas e permitir melhorias [288];
- Desenvolvimento de sistemas de restauração de voz, destinados à restauração de vozes de pessoas que sofreram de alguma patologia que prejudica a emissão vocal, tais como vítimas de operações mal-sucedidas na laringe [222].

1.2 Visão Geral de um Sistema de Conversão Típico

A Figura 1.2 apresenta um esqueleto de um sistema de conversão de voz típico. O sistema recebe sentenças do falante origem (S_0) e do falante destino (T_0), que são usados na fase de treinamento para definir a transformação ($\mathbf{T}$) de características acústicas do origem (que podem ser locais ou globais) para características acústicas do destino. Após esta fase, o sistema recebe novas sentenças S_f do falante origem e sintetiza a sentença T_f, que deveria carregar a mesma mensagem S_f, todavia com identidade sonora do falante destino.

[1] Ver detalhes em http://verbmobil.dfki.de/overview-us.html

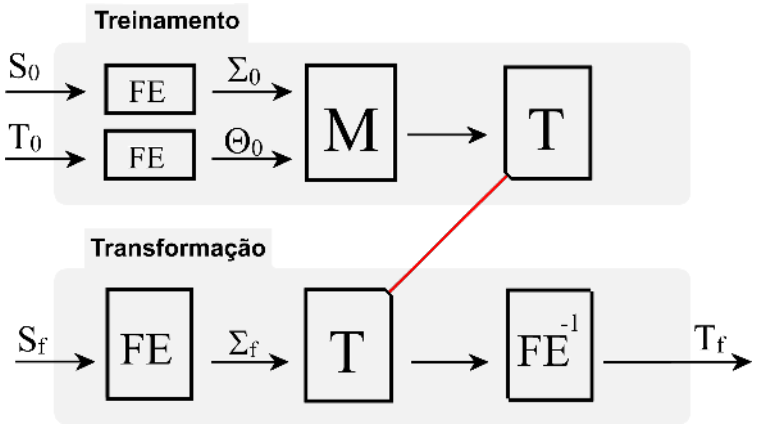

Figura 1.2: *Processo geral na conversão de voz.*

1.2.1 Treinamento

A fase de treinamento é normalmente a primeira etapa para transformações de voz em geral. Este é o estágio onde dados de ambos os falantes são coletados e processados, a fim de obter uma caracterização razoável dos parâmetros acústicas de cada falante, e assim permitir a definição da transformação que será usada no estágio subsequente. De acordo com a Figura 1.2, nesta fase o SCV:

1. Recebe Sentenças S_0 e T_0;
2. Realiza uma extração de parâmetros acústicos relevantes para cada falante, criando representações alternativas Σ_0 e Θ_0 para os falantes origem e destino, respectivamente.
3. Processa as características acústicas Σ_0 e Θ_0 a fim de obter uma base de dados de descritores acústicos tanto local (baseada em segmentos), quanto global (baseada nas sentenças), para ambos os falantes.
4. Define uma transformação dos descritores locais e globais do origem para os descritores locais e globais do destino.

Descritores acústicos são ditos locais se estes descrevem características de um único segmento, e são ditos globais se estes correspondem a características que modelam a sentença inteira que foi pronunciada por um falante. Exemplos de descritores locais são a altura musical instantânea, a energia, e a envoltória espectral, ou a classe fonética artificial à qual o particular segmento pertence. Exemplos de descritores globais são médias e desvios padrões da altura musical ou da energia de um sinal de voz.

Após obtida a transformação $\mathbf{T}$, basta utilizá-la na fase de conversão. A transformação $\mathbf{T}$ engloba todo tipo de transformação do sinal de entrada, incluindo aspectos de timbre, altura, volume e ritmo da pronúncia.

1.2.2 Transformação

Esta fase consiste em transpor os parâmetros acústicos que representam o sinal de voz, a fim de converter os trechos sonoros de uma sentença do origem em um sinal que se pareça com a mesma sentença pronunciada pelo destino.

A fase de transformação tem uma estrutura semelhante à do treinamento: as sentenças de entrada S_f são representadas como Σ_f num espaço de características acústicas, as quais são convertidas pela transformação $\mathbf{T}$ definida na fase de treinamento para uma representação Θ_f, que é finalmente transformada (por uma transformação inversa) na sentença T_f .

A representação alternativa destas sentenças num espaço de características acústicas (Σ e Θ no diagrama) supostamente preserva informações suficientes a ponto de não somente garantir a possibilidade de ressintetizar o sinal, mas também de permitir a manipulação dos aspectos de timbre e prosódia do sinal. Uma vez que muitos destes aspectos são atributos variáveis no tempo, a extração destes parâmetros acústicos do sinal de voz é realizada trecho a trecho (segmento a segmento).

A reconstrução de um sinal de voz a partir de segmentos processados deve ser bastante cuidadosa, uma vez que mudanças no conteúdo espectral podem introduzir artefatos devido a diferenças entre segmentos adjacentes, e podem se percebidos como ruído de alta frequência, "cliques" ou características não-naturais da voz (como deslocamentos de formantes).

A seção seguinte apresenta uma visão histórica geral dos artigos relacionados à conversão de voz, bem como as principais técnicas utilizadas.

1.3 Estado da Arte em Conversão de Voz

O objetivo desta seção é introduzir trabalhos anteriores, bem como relacioná-los entre si, segundo critérios de técnicas de transformação utilizadas e modelos de representação do sinal. Antes de classificarmos os trabalhos, segue uma breve descrição das técnicas mais importantes em conversão de voz, por ordem cronológica.

1.3.1 Visão Histórica das Técnicas de Conversão de Voz

Historicamente, o problema se instaura em 1985 quando Childers [33] tem a idéia de transpor parâmetros acústicos de sentenças pronunciadas de um falante para outro. No ano seguinte, Shikano [231] traz uma proposta utilizando técnicas de *vector quantization (VQ)* e sentenças *codebooks*. Em 1988, Abe [1; 2] introduz o conceito de sistemas de conversão de voz inter-linguística (SCVI) para falantes bilíngues.

Em 1991, Valbret [284] reacende a discussão, introduzindo o conceito de *Dynamic Frequency Warping (DFW)* em SCVs e sugerindo a idéia de se personalizar sistemas *Text-to-Speech*. Depois de 4 anos, Childers [32] propõe um SCV a partir da modelagem do pulso glotal, e no mesmo ano Narendranath [169] introduz conceitos de *Redes Neurais Artificiais (ANN)* em SCVs. Em 1996, Baldoin [14] apresenta um *survey* sobre SCVs, comparando métodos para transformação da envoltória espectral como ANN, VQM, LMR e GMM.

Em 1998, Arslan [6; 8] desenvolve o algoritmo *STASC (Speaker Transformation Algorithm using Segmental Codebooks)* para conversão de voz e utiliza *Line Spectral Frequencies* na representação da envoltória espectral. Ainda neste ano, Stylianou [257] introduz conceitos probabilísticos como o *Modelo de Misturas Gaussianas (GMM)* em SCVIs, passando a utilizar *MFCCs* na representação da envoltória espectral, sendo os resultados aprimorados logo em seguida por Kain [107; 108].

Em 2001, Toda et al [270] propõem uma nova técnica para conversão de voz e representação do espectro, conhecida como *STRAIGHT* (*Speech Transformation and Representation by Adaptive In-*

terpolation of weiGHTed spectrum), na qual é possível manipular parâmetros tais como a envoltória espectral, a taxa de pronúncia, a altura musical e outros parâmetros acústicos.

No ano seguinte, Türk [275] combina o algoritmo *STASC* de Arslan com a transformada Wavelet discreta (*DWT*) do sinal de voz. Ainda em 2002, Mashimo [148] propõe uma avaliação comparativa entre as abordagens dependentes de texto (para falantes bilíngues) e independentes de texto em sistemas de conversão de voz inter-linguística nos idiomas japonês/inglês. Já em 2003, Sündermann [245; 246; 247] reintroduz uma técnica proposta por Kamm et al. [110], a normalização do trato vocal denominada *VTLN*, para representação dos parâmetros espectrais do sinal de voz. A partir do levantamento bibliográfico realizado tem-se que Sündermann estabeleceu o conceito de conversão de voz independente de texto, bem como foi o primeiro a propor um SCVI independente de texto.

A partir de 2004, houve um aumento do número de trabalhos nesta área [193; 194; 195; 203; 207; 282; 283; 299; 300; 305; 313], com pesquisas principalmente em técnicas probabilísticas, em particular GMMs, sentenças *codebook* e outras técnicas como ANN e DFW. Comparações entre alguns destes trabalhos foram realizadas por Türk e Schröder [279] onde métodos de conversão e síntese de voz foram avaliados pela qualidade da conversão. No ano de 2004, Sündermann [239; 240] apresentou um arcabouço geral para um sistema de conversão de voz independente de texto.

Apresentamos a seguir alguns trabalhos que atacam este problema utilizando diversas técnicas de transformação e representação do sinal.

1.3.2 Classificação das Técnicas de Conversão de Voz

Para simplificar a apresentação das contribuições encontradas na literatura [139; 140] classificaremos os trabalhos de acordo com os parâmetros espectrais utilizados para representar os sinais de voz, e de acordo com as técnicas envolvidas na fase de transformação.

Quanto aos Modelos de Representação

Em um segmento de um sinal de voz quase estacionário a *envoltória espectral* é suficiente para caracterizar o timbre (instantâneo) da fala. Todavia, existem ainda outros parâmetros que usualmente são calculados para cada segmento, como a frequência fundamental (F_0), a energia (rms) e informações do conteúdo harmônico, fundamental para classificação e transformação da identidade da voz. Além do espectro de Fourier e seu envelope espectral, os sistemas de conversão de voz usualmente utilizam-se de muitos outros modelos para representação do sinal de voz como:

1. ***Modelos baseados em voz**: Vocal Tract Length Normalization (VTLN), Frequências Formantes, e modelos de estimação de Fluxo Glotal.*
2. ***Modelos baseados em voz/sinal**: Linear Prediction Coding (LPC), Line Spectral Frequencies (LSF), Cepstral Coefficients, e Speech Transformation and Representation using Adaptative Interpolation of weiGHTed spectrum (STRAIGHT).*
3. ***Modelos baseados em sinal**: Improved Power Spectrum Envelope (IPSE), Discrete Wavelet Transform (DWT) e Harmonic plus Noise Model (HNM).*

Modelos baseados em voz são representações baseadas nos mecanismos de produção da voz humana, usando conceitos como pulso glotal, que é o sinal produzido pelas cordas vocais, e trato

vocal, que compreende a ação dos ressoadores, tais como as cavidades oral e nasal, palatos, língua, dentes e lábios, os quais são responsáveis por caracterizar grande parte da identidade sonora de um falante.

Modelos mistos voz/sinal são na realidade modelos de sinais que fornecem representação compacta destes sinais. Uma vez que são amplamente utilizados pela comunidade de reconhecimento de fala, estes modelos incorporaram muitas interpretações relacionadas à fala. Por exemplo, regiões do cepstrum são frequentemente relacionadas ou às regiões formantes (e assim, contribuem para o trato vocal) ou à frequência fundamental, assim como os coeficientes LPC e os resíduos LPC podem ser em alguns casos associados ao trato vocal e pulso glotal, respectivamente (vistos num contexto de síntese subtrativa).

Modelos puramente baseados em sinais são representações gerais do sinal nos domínios do tempo e frequência, e usualmente não carregam nenhum conteúdo semântico no que diz respeito às especificações da fala ou informação fonética. O modelo harmônico mais ruído (Harmonic plus Noise Model – HNM) é mais específico do que os outros, porém pode ser muito útil para representar porções de sinais vozeadas (*"voiced"*) do sinal, como por exemplo as vogais.

Várias destas técnicas permitem o uso de conjuntos alternativos de frequências, tais como as escalas perceptuais *BARK* e *MEL* (ver Seção 2.1). Estas técnicas podem ser vistas detalhadamente no Capítulo 2.

Quanto às Técnicas de Transformação

A fase de transformação do sinal em sistemas de conversão de voz leva em consideração todas as características acústicas usadas na representação do sinal de voz. Esta fase inclui não somente a modulação da frequência fundamental ou compensação de energia, mas também as transformações do conteúdo harmônico nas quais os aspectos de timbre e inteligibilidade são levados em consideração. No estado-da-arte em conversão de voz destacam-se o Empenamento (ou Deformação) Dinâmico em Frequência [171] *(Dynamic Frequency Warping – DFW)* e o Modelo de Misturas Gaussianas [52; 153; 267] *(Gaussian Mixture Model – GMM)* como funções de transformação espectral. Técnicas de transformação estão estreitamente ligadas aos modelos de representação e serão detalhadas posteriormente na Seção 2.4. Algumas das técnicas comumente utilizadas [139; 140] são:

1. **Técnicas Estatísticas:** Modelos de Misturas Gaussianas (Gaussian Mixture Models – GMM), Modelos Ocultos de Markov (Hidden Markov Models – HMM), Distribuições de Probabilidade Multi-Espaciais (Multi-Space Probability Distributions – MSD), Estimadores de Máxima Verossimilhança (Maximum Likelihood Estimators – MLE), Análise de Componentes Principais (Principal Component Analysis – PCA) e Técnicas de Clusterização (Unit Selection, Frame Selection, K-means, K-histograms).

2. **Técnicas Cognitivas:** Redes Neurais Artificiais (Artificial Neural Networks – ANN), Redes Neurais com Funções de Base Radial (Radial Basis Function Neural Networks – RBFNN), Mapeamento de Características Topológicas (Topological Feature Mapping – TFM) e Mapeamento Topográfico Generativo (Generative Topographic Mapping – GTM).

3. **Técnicas de Álgebra Linear:** Modelos Bilineares, Decomposição em Valores Singulares (Singular Value Decomposition – SVD), Interpolações Lineares Ponderadas (Weighted Linear

Interpolations), Transformações Lineares Perceptualmente Ponderadas (Perceptually Weighted Linear Transformations – PWLT) e Regressão Linear (LRE, LMR, MLLR).

4. **Técnicas de Processamento de Sinais:** Quantização de Vetor (Vector Quantization – VQ), Sentenças Codebook, o algoritmo STASC (Speaker Transformation Algorithm using Segmental Codebooks) e Empenamento de Frequências (DFW, WFW).

Estas técnicas se diferenciam de acordo com o modo pelo qual elas enxergam os dados. Por exemplo, técnicas estatísticas usualmente supõem que os dados (vetores de características ou parâmetros vocais) possuem uma componente aleatória e por esta razão podem ser descritos por médias e desvios padrões (modelos Gaussianos), ou que estes estão distribuídos no tempo de acordo com regras simples baseadas no passado recente (modelos Markovianos).

Técnicas cognitivas são baseadas em processos de aprendizado usando estruturas neurais abstratas, e intrinsecamente dependem de uma fase de treinamento (onde tanto as entradas como as saídas são disponíveis). Frequentemente este aprendizado se refere a problemas de decisão (onde somente dois valores de saída são possíveis), como por exemplo em reconhecimento de fala, onde uma rede particular é treinada a partir de um fonema específico de uma palavra ou sentença que será reconhecida.

Técnicas de Álgebra Linear são baseadas em interpretações geométricas dos dados, tais como modelos simplificados usando projeções ortogonais (regressão linear), combinações convexas dos dados de entrada (interpolações ponderadas) ou decomposições em valores singulares (SVD).

Técnicas de processamento de sinais definem transformações do sinal baseadas em representações no domínio do tempo ou frequência. Estas tentam codificar o sinal usando frequentemente uma biblioteca de chaves de segmentos do sinal ou *codewords*, ou tentam converter aspectos do timbre da fala através da modificação das escalas de representações das frequências (*warping*).

As categorias acima não são disjuntas. Por exemplo, LRE e SVD são usados frequentemente como ferramentas estatísticas, assim como PCA e VQ são construídos usando ferramental proveniente da álgebra linear.

A Tabela 1.1 discrimina as técnicas e os modelos de representação do sinal utilizadas por cada autor. Nesta tabela, autores marcados com asterisco ($*$) publicaram trabalhos a respeito de SCVs independentes de texto e autores marcados com diamantes ($\diamond$) publicaram trabalhos sobre conversão de voz inter-linguística.

1.4 Limitações e Desafios

Existem muitos problemas em aberto na pesquisa de conversão de voz inter-linguística, os quais foram identificados em alguns dos artigos anteriores:

- **Problemas fonéticos**: Em conversão de voz inter-linguística, sabe-se que existem muitos fonemas da língua do origem que não existem na língua do destino. Falantes bilíngues têm sido usados [7; 148; 230] a fim de derivar transformações fonéticas que permitem aproximar (mas não tornar idênticos) os fonemas a serem convertidos entre línguas distintas. Se tais transformações poderiam ser bem sucedidas utilizando outros sujeitos (não bilíngues) é um problema em aberto.

Ano	Autor	Técnica de transformação	Modelo de representação do sinal
1986	Shikano [231]	VQ e Codebook	LPC
1988	Abe [1]	Codebook	Parâmetros Espectrais Diversos
1990	Abe$^{\diamond}$ [2]	Codebook e HMM	LPC
1991	Valbret [284]	Regressão Linear e DFW	LPC
1995	Childers [32]	Glottal origem Modeling	LF Glotal e LPC
1995	Narendranath [169]	ANN	Frequências Formantes
1996	Rinscheid [209]	Topological Features Map	LPC
1996	Verhelst [287]	VQ e Codebook	LPC
1996	Lee [130]	ANN	LPC do Cepstrum e Resíduo do LPC
1997	Kim [118]	HMM, VQ e Codebook	LPC e Cepstrum
1998	Arriola [80]	LRE	Coeficientes e Resíduo LPC
1998	Arslan [8]	HMM e STASC	LSF
1998	Kain [107; 108]	GMM	Bark e LSF
1998	Stylianou [257]	GMM	HNM, Bark e MFCC
1999	Arslan [6]	STASC	LSF
2001	Kain [109]	GMM	LPC
2001	Zhang [314]	MLLR	MFCC e LSF
2001	Lopez [135]	VQ e ANN	Coeficientes e Resíduo LPC
2001	Mashimo$^{\diamond}$ [149]	GMM	MFCC e STRAIGHT
2001	Toda [270]	GMM e DFW	STRAIGHT
2002	Türk [275]	STASC	DWT
2002	Watanabe [292]	ANN	Envelope Espectral
2003	Sündermann$^{\diamond *}$ [245; 246; 247]	DFW	VTLN
2003	Kumar* [125]	GMM	MFCC
2003	Türk [276]	Codebook	LSF
2003	Ye e Young [298]	GMM e PWLT	MEL
2003	Orphanidou [182]	Codebook e GTM	LPC
2003	Rentzos [206]	HMM	MFCC, LPC e LF Glotal
2004	Rentzos [207]	HMM	MFCC, LPC e LF Glotal
2004	Duxans [47]	GMM, HMM e árvore de decisão	LSF
2004	Orphanidou [183]	RBFNN	DWT
2004	Ye e Young [300]	GMM e Codebook	LSF
2004	Wilde [294]	PPCA	LSF
2004	Pribilova [195]	Non-linear Frequency Scale Mapping	HNM
2004	Pfitzinger [189]	Weighted Linear Interpolation	Coeficientes e Resíduo LPC
2005	Toda [268]	Maximum Likelihood	MFCC e STRAIGHT
2005	Zhang [310]	GMM	Resíduo LPC, Bark e LSF
2005	Kang [111]	GMM e Codebook	STRAIGHT e LSF
2006	Nurminen [175]	GMM	Bark e LSF
2006	Nurminen [176]	GMM e K-means	LSF
2006	Duxans$^{\diamond *}$ [48]	GMM, HMM e CART	HNM, LPC e LSF
2006	Ye e Young [301]	GMM	LPC e LSF
2006	Sündermann$^{\diamond *}$ [244]	Unit Selection	VTLN
2006	Rao [203]	Transformação linear	LPC
2006	Shuang$^{\diamond *}$ [233]	Frequency Warping	Frequências Formantes
2007	Dutoit [45]	Frame Selection	MFCC e LPC
2007	Erro$^{\diamond *}$ [49; 50]	WFW	HNM e LSF
2007	Fujii [60]	Unit Selection	LPC e MFCC
2007	Guido [79]	ANN	DWT
2007	Hanzlicek [83]	GMM	LSF
2008	Hanzlicek [84]	Warpings Functions	LSF
2008	Shuang$^{\diamond}$ [234]	Frequency Warping	Frequências Formantes
2008	Yue [305]	GMM e HMM	LPC e LSF
2008	Zhang$^{\diamond *}$ [311; 313]	Codebook	STRAIGHT e LSF
2008	Desai [41; 42]	ANN	Coeficientes Cepstrais
2008	Pozo [194]	GMM	LF Glotal e LPC
2009	Popa [193]	SVD e Asymmetric Bilinear Model	LSF
2009	Uriz* [282]	Frequency Warping e Frame Selection	LSF
2009	Uriz* [283]	K-Histogramas e Frame Selection	LSF
2009	Zhang$^{\diamond *}$ [312]	Codebook	STRAIGHT e LSF
2009	Yutani [306]	MSD, GMM e HMM	MFCC e F_0
2010	Desai [41]	ANN	Coeficientes Cepstrais
2010	Godoy* [72]	DFW e GMM	Coeficientes Cepstrais
2010	Helander [89]	Regressão Linear e GMM	Coeficientes Cepstrais
2010	Lanchantin [127]	DMS	LSF
2012	Zorilua [318]	DFW e GMM	HNM e LPC
2012	Song [249]	SVR e GMM	STRAIGHT

Tabela 1.1: *Relação de trabalhos em conversão de voz típica, independente de texto* e inter-linguística$^{\diamond}$.*

- **Problemas da prosódia**[2]: Não somente o timbre de voz, mas também a prosódia é um fator preponderante na identidade sonora, e deve ser levado em consideração [113; 260]. Existem muitos aspectos acústicos globais que são decisivos a fim de se obter boas conversões, tais como a média e desvio padrão da altura musical, energia e algumas medidas estatísticas relacionadas ao ritmo de articulação na pronúncia. Algumas destas são discutidas por Helander [90], Hanzlicek [82] e Ceyssens [27; 28].

- **Problemas da qualidade da fala**: A qualidade da fala é uma medida subjetiva que se refere ao grau de qualidade do áudio propriamente. Alguns problemas que são percebidos como falta

[2] A prosódia pode ser entendida simplificadamente como os parâmetros que controlam a altura, intensidade e velocidade de execução das sentenças pronunciadas.

de qualidade no sinal convertido são ruídos de fundo, tons inexistentes, cliques e outros aspectos no timbre que podem ser descritos como uma voz sintética ou não-natural. Não somente aspectos ruidosos interferem na qualidade do áudio, mas também outros fatores relacionados à percepção humana, como a naturalidade da pronuncia. Por exemplo, deslocamentos de altura musical muito grandes sem uma correção das frequências formantes podem degradar a qualidade (além da inteligibilidade) da voz convertida. Este problema, citado várias vezes na literatura [60; 108; 248; 257], é facilmente percebido em testes subjetivos.

- **Problemas na similaridade da fala**: Estes estão relacionados à qualidade do timbre e à identidade vocal, principalmente nos aspectos fonéticos da fala, embora sejam facilmente confundidos com problemas relacionados à qualidade e prosódia em testes experimentais. Na teoria, um sinal de voz puramente sintético poderia ser percebido como uma voz não-natural, mas similar em termos de timbre a uma voz do destino. A conversão de voz entre falantes de gêneros diferentes é particularmente suscetível a este tipo de problema [108; 109; 169].

- **Problemas na avaliação da conversão de voz**: Ainda se desconhece uma ferramenta padrão que avalie objetivamente uma conversão de voz. É comum cada trabalho realizar seus próprios testes a partir de métricas preferidas por cada autor, como a distância Itakura [295], o teste OGI [298; 299; 300], a distância P [282; 283] ou a distorção cepstral CD [149; 270], entre outras. Sabe-se que tais medidas objetivas não estão propriamente correlacionadas com a percepção humana, e em alguns casos distâncias espectrais não tão pequenas ainda podem fornecer uma boa transformação [238; 275]. Normalmente costuma-se adotar métricas subjetivas na avaliação da proposta, o que dificulta a comparação entre métodos existentes. Os testes subjetivos (ou perceptuais) mais populares são o de opinião média (*Mean Opinion Score – MOS*) e o método com duas escolhas forçadas (ABX) [107]. O teste *MOS* é bastante utilizado para análise da qualidade da fala convertida e o teste ABX é muito utilizado para análise da similaridade da fala convertida.

- **Problemas com suavização excessiva**: É um problema causado por métodos interpoladores (GMM por exemplo) na fase de transformação, que degrada o espectro do sinal ao eliminar detalhes importantes do mesmo. Tal degradação corresponde a uma alta suavização (do inglês *Excessive Smoothing*) do espectro, reduzindo muito a similaridade entre a voz convertida e a voz do destino [221; 268; 269].

- **Problemas de ajuste excessivo (*over-fitting*)**: É comum pensar em aumentar o conjunto de dados na fase de treinamento afim de se obter uma transformação mais fidedigna. Entretanto, tal estratégia tende a provocar descontinuidades (mudanças abruptas) entre segmentos consecutivos na fase de síntese do sinal [155]. Tais descontinuidades caracterizam os problemas de excesso de treinamento (*Over-fitting problems*).

Alguns cuidados e direções sobre a conversão de voz podem ser encontradas na literatura [260]. Por exemplo, transformações na escala de frequência por um fator maior que 1.2 ou menor que 0.8 influenciam significativamente na qualidade da voz convertida, por causa de efeitos decorrentes da distorção das *regiões formânticas* e de pequenos detalhes do espectro. Tais problemas podem ser solucionados por estratégias de readequação da envoltória espectral [82; 90; 263]. O processo de

análise, modificação e síntese do estilo de fala de um locutor também é um problema emergente em sistemas de conversão de voz [219; 278; 279; 280].

1.5 Objetivos

Como objetivo principal, este trabalho visa propor um sistema completo e robusto para conversão de voz inter-linguística. Para tanto, como objetivos secundários, devemos:

1. Fundamentar teoricamente o problema da conversão de voz inter-linguística, a fim de obter conhecimento a priori em termos perceptuais da conversão em questão.
2. Avaliar diferentes modelos de representação do sinal de voz.
3. Organizar as soluções existentes, classificando-as e relacionando-as entre si.
4. Propor um conjunto de técnicas e ferramentas úteis para transformação de sinais de voz em geral.
5. Apresentar um sistema completo de conversão de voz inter-linguístico.

1.6 Contribuições

As contribuições apresentadas por este trabalho compõem o Capítulo 3 desta tese. As principais contribuições deste trabalho estão discriminadas abaixo:

1. Extensão do Modelo Harmônico/Estocástico (HSM) mediante a introdução da modelagem de **coeficientes de configuração fásica**, o qual é detalhado na Seção 3.2.
2. Desenvolvimento de um novo modelo de representação espectral baseado na **decomposição espectral usando soma de bases radiais**, o qual é detalhado na Seção 3.3.
3. Composição de um **mapa fonético artificial** utilizando pressupostos linguísticos, tais como regiões formânticas, entre outras (Seção 3.4). Além disto, uma nova técnica de clusterização denominada **agrupamento k-verossímil** também é uma contribuição importante do trabalho contida nessa mesma seção.
4. Proposta de um método de **alinhamento de classes acústicas** e segmentos de voz, usando métodos de **emparelhamento** oriundos da Teoria dos Grafos, conforme apresentado na Seção 3.5.
5. Desenvolvimento de um conjunto de ferramentas de transformação de voz baseadas no conceito de **distribuição energética** espectral e **empenamento em frequências normalizadas**, os quais são discutidos na Seção 3.6.

Como resultado dessa tese, um sistema de conversão de voz completo foi desenvolvido, o qual pode ser diretamente acessado na *homepage* do autor[3].

[3] http://score.ime.usp.br/~dandy/

1.7 Organização do Trabalho

Esta tese está estruturada em três capítulos principais, um de fundamentação teórica, o Capítulo 2, um relacionado ao sistema de conversão de voz propriamente dito, o Capítulo 3, e um dedicado à parte experimental, o Capítulo 4.

O Capítulo 2 apresenta o arcabouço teórico sobre o qual é possível desenvolver um sistema de conversão de voz inter-linguístico. Neste capítulo são discutidas algumas bases fundamentais a respeito de processamento de voz em geral (Seção 2.1), e posteriormente são revisitados alguns modelos clássicos do sinal de voz (Seção 2.2), tanto da parte espectral (tais como LPC, Cepstrum e Modelos Senoidais) quanto da prosódia (PSOLA, Fujisaky, etc.). O texto passa então a apresentar conjuntos de ferramentas úteis na transformação de elementos prosódicos da fala, tais como a altura musical ou ritmo de articulação (Seção 2.3.1), e transformações espectrais (Seção 2.4).

O Capítulo 3 condensa a maior parte das contribuições do trabalho, onde o sistema de conversão de voz inter-linguístico é apresentado e implementado em pseudo-código, módulo por módulo. A Seção 3.1 apresenta o sistema de conversão de voz em uma visão estruturada e global, e entre as Seções 3.2 e 3.6 são detalhados cada um dos módulos do sistema definidos na visão estruturada.

O Capítulo 4, por sua vez, realiza dois tipos de avaliações do sistema, bem como de suas principais componentes. A primeira seção deste capítulo (Seção 4.1) introduz os aspectos a serem considerados na avaliação, bem como outros pontos comuns na experimentação, como as bases de dados utilizadas no sistema. A validação dos métodos propostos é realizada em duas categorias: avaliação subjetiva e objetiva. A avaliação objetiva apresentada na Seção 4.2 utiliza métricas quantitativas para avaliar o quão satisfatório é cada método proposto. Na parte subjetiva (Seção 4.3), por sua vez, é apresentada uma avaliação perceptual usando entrevistas de opinião a respeito dos resultados obtidos na conversão.

Finalmente, no Capítulo 5 discutiremos algumas conclusões obtidas neste trabalho, analisando as vantagens e desvantagens dos métodos propostos, bem como sugerindo algumas possibilidades de continuidade deste trabalho.

Capítulo 2

Ferramentas e Métodos

A exploração de ferramentas e métodos já conhecidos a fim de alcançar o objetivo do trabalho é de fundamental importância na fase de desenvolvimento. Entretanto, em certas ocasiões é preciso propor novos métodos adaptados ao problema a fim de solucionar detalhes técnicos que não são atendidos pelos métodos convencionais. No caso da conversão de voz inter-linguística não é diferente. A metodologia adotada por este trabalho segue o seguinte fluxo de desenvolvimento: (1) Manter o enfoque na reutilização de métodos e recursos técnicos disponíveis na literatura, procurando utilizar as ferramentas mais adequadas para a resolução do problema proposto; (2) Em casos de necessidade, propor novos métodos e ferramentas a fim de resolver o problema em questão.

Por se tratar de um problema não muito explorado em relação à conversão de voz convencional, isto é, a conversão de voz intra-linguística dependente de texto, o objetivo deste capítulo é revisar alguns conceitos e técnicas aplicáveis à **conversão inter-linguística independente de texto**. Deste modo, um levantamento geral das principais ferramentas utilizadas de modo direto e indireto na tarefa de conversão de voz é mostrado no capítulo conseguinte.

O capítulo está estruturado da seguinte forma:

- A Seção 2.1 apresentará conceitos teóricos rudimentares sobre a voz e suas propriedades, abordando aspectos básicos concernentes à fala humana e seus termos técnicos.

- Em seguida, na Seção 2.2 serão apresentados dois modelos detalhados, sobre os quais se fundamentam as técnicas de processamento de voz. Tal seção especifica detalhadamente o modelo fonte-filtro e o modelo sobre o qual se fundamenta o sistema de conversão de voz apresentado nesta tese: o Modelo Harmônico-Estocástico;

- E finalmente, a Seção 2.3 realiza uma ampla varredura exploratória do arcabouço disponível para o processamento digital do sinal de voz. Esta seção organiza o compêndio de técnicas e ferramentas em duas subseções, uma aplicada à conversão de prosódia, e outra à conversão espectral.

2.1 Rudimentos do Processamento de Voz

O Processamento Digital de Sinais é uma disciplina essencial para o desenvolvimento de qualquer tipo de sistema digital. Tal disciplina fornece um conjunto de métodos para análise de sinais do mundo real usando ferramentas matemáticas, a fim de realizar transformações, ou simplesmente

extrair informações desses sinais. Particularmente, o Processamento de Fala e Voz é uma das áreas de conhecimento que mais tem crescido na última década, principalmente mediante o avanço tecnológico em áreas como síntese [20; 164; 258; 264], reconhecimento de voz [222; 308] e de falante [254]. Além destas ferramentas, um conhecimento básico é requerido em relação aos aspectos fisiológicos da fala humana.

Fala é a faculdade humana de utilizar de uma certa maneira os recursos expressivos oferecidos por uma língua e compreende fatores linguísticos de qualquer tipo efetivamente realizados pelos falantes. A voz é o meio de comunicação empregado na fala [220]. Existem vários aspectos que discriminam a voz humana e conhecer alguns conceitos rudimentares e inerentes à voz é de grande importância para uma exploração objetiva das ferramentas de conversão de voz.

Existem algumas características da voz que devem ser invariantes dentro de um conjunto de falantes de mesma língua, com o intuito de se estabelecer uma comunicação bem-sucedida; entretanto, existem outras características que contribuem para a especificação da identidade sonora de um indivíduo. Embora a identidade sonora seja um conceito amplo e não muito bem definido pela literatura, é possível modelar e caracterizar traços da identidade sonora a partir de atributos físicos do som presentes no sinal de voz.

Atributos Físicos do Som

Sabe-se que o som é a propagação de uma frente de compressão mecânica ou onda mecânica; esta onda se propaga de forma circuncêntrica, em meios materiais, isto é, aqueles que têm massa e volume, como os sólidos, líquidos ou gasosos. Os sons vocais são, na sua maior parte, combinações de sinais mais simples, representados por uma senoide pura, que possui uma velocidade de oscilação ou **frequência** e uma **amplitude**.

A frequência (f) é uma grandeza física que indica o número de ocorrências de um evento (ciclos, voltas, oscilações, etc) em um determinado intervalo de tempo. Alternativamente, podemos medir o tempo decorrido para uma oscilação completa, que recebe o nome de **período** de oscilação (T). Desse modo, a frequência é o inverso do período:

$$f = \frac{1}{T}.$$

A unidade de medida para frequência é o Hertz (Hz), ou ciclos por segundo (s^{-1}). A frequência angular (ω) por sua vez, é a taxa de variação angular em movimentos circulares. No Sistema Internacional de Unidades, é medida em radianos por segundo, e se relaciona com a frequência em Hertz pela equação:

$$\omega = 2\pi f = \frac{2\pi}{T}.$$

O decibel (dB) é uma medida da razão entre duas grandezas, tipicamente energia, potência ou amplitude, sendo usado para uma grande variedade de medições em acústica, física e eletrônica, como por exemplo na medida da intensidade de sons. A definição do dB é obtida com o uso do logaritmo. Uma intensidade sonora I pode ser expressa em decibéis através da equação

$$I_{dB} = 10 \log_{10}(\frac{I}{I_0}),$$

onde I_0 é a intensidade de referência. No caso de se representar a pressão sonora em decibéis, o

fator multiplicativo 10 é substituído por 20, uma vez que a intensidade sonora é uma grandeza quadrática em relação à pressão sonora (dobrar a constante é equivalente a elevar ao quadrado o quociente).

Percepção Humana

A audição é um mecanismo sofisticado. Seres humanos e vários animais percebem sons com o sentido da audição, e com seus ouvidos conseguem saber a distância e posição da fonte sonora (audição estereofônica). Isso permite, por exemplo, que predadores cacem suas presas eficientemente (no caso de morcegos), bem como permite a fuga de presas ariscas diante do bramido de um leão voraz. No caso dos humanos, a audição se destaca como um dos principais sentidos receptivos da comunicação à apreciação musical.

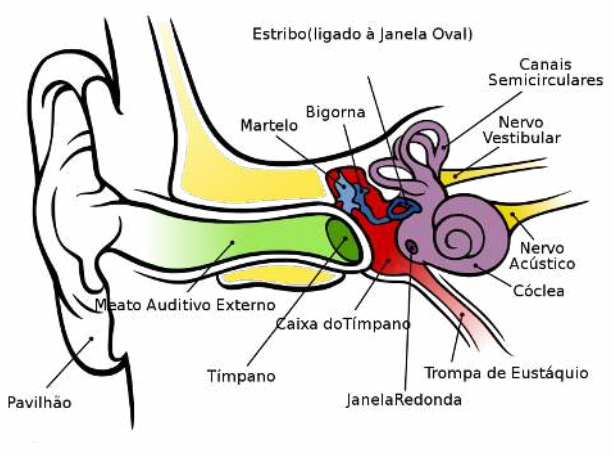

Figura 2.1: *O aparelho auditivo.*

O aparelho auditivo, o ouvido, é constituído por três partes: o ouvido externo, o ouvido médio e o ouvido interno. Observe a Figura 2.1 e acompanhe o processo de funcionamento da audição[1]:

1. O som propaga-se desde a fonte sonora até os nossos ouvidos.
2. O pavilhão auricular (ouvido externo) recebe as ondas sonoras, encaminhando-as através do canal auditivo até o ouvido médio.
3. O tímpano, a pequena membrana que separa o ouvido externo do médio, vai então vibrar de acordo com as moléculas do ar ao redor.
4. Essas vibrações vão então ser transmitidas para o interior da cóclea (no ouvido interno) através dos três ossículos: o martelo, a bigorna e o estribo, ligados em série entre o tímpano e a janela oval, e funcionando como transportadores do som recebido.
5. As vibrações são então transmitidas à cóclea, provocando a oscilação da membrana basilar.
6. Cada região da membrana basilar é especializada em responder aos estímulos provocados por oscilações de uma faixa estreita de frequências.

[1] A comunidade *Wikimedia Commons* concede ao autor a permissão para a livre utilização da imagem http://commons.wikimedia.org/wiki/File:Anatomia_do_Ouvido_Humano.svg

A membrana basilar é uma região interna na cóclea responsável pela identificação das frequências sonoras. Sabe-se que ao longo desta membrana as frequências são percebidas dentro de uma escala logarítmica de frequência. Os sons audíveis pelo ouvido humano têm frequências entre 20 Hz e 20 kHz. Acima e abaixo desta faixa estão ultra-som e infra-som, respectivamente. Daí surgem vários modelos de escalas adaptadas a este modelo perceptual, conhecido como *escalas perceptuais*. Dois destes modelos serão destacados ao longo deste trabalho: a Escala *MEL*, proposta por Stevens, Volkman e Newmann em 1937 [252]; e a Escala de *Bark*, proposta por Eberhard Zwicker em 1961 [319].

A escala MEL é uma escala perceptual musical baseada na percepção de intervalos melódicos segundo observadores especializados. Sua escala de conversão é dada pela seguinte equação:

$$m = 1127.01048 \log_e(1 + f/700),$$

e sua inversa é:

$$f = 700(e^{m/1127.01048} - 1),$$

onde f é a frequência em *Hertz* e m é a altura musical em *MELS*.

A Escala de *Bark* é uma escala psicoacústica de uso geral em acústica. O nome Bark é uma homenagem à Heinrich Barkhausen [319], que propôs a primeira medição subjetiva de intensidade sonora. Tal escala subdivide as frequências em 24 faixas de valores, denominadas *bandas críticas*. Cada banda crítica possui um frequência base. As frequências base da escala Bark de audiometria são, em Hz: 20, 100, 200, 300, 400, 510, 630, 770, 920, 1080, 1270, 1480, 1720, 2000, 2320, 2700, 3150, 3700, 4400, 5300, 6400, 7700, 9500, 12000, 15500. Sua escala de conversão é dada pela seguinte equação:

$$b = 13 \arctan(0.76f/1000) + 3.5 \arctan((f/7500)^2),$$

onde f é a frequência em *Hertz* e b é a frequência em Bark. Para encontrar a banda crítica de uma determinada frequência *Bark* b, basta tomar o piso da seguinte equação:

$$B_c = 52548/(b^2 - 52.56b + 690.39),$$

onde B_c é a banda crítica em *Hertz* e b é a frequência em *bark*.

Tais escalas servem como base para modelos de representação do sinal de voz. Além do sistema auditivo, conhecer detalhadamente o processo de produção da voz humana nos auxilia na modelagem do sinal.

Processo de Produção da Voz

O aparelho fonador compreende os órgãos encarregados por produzir a fala. Os órgãos do aparelho fonador são categorizados em três grupos:

1. Os **foles** geram uma coluna expiratória de ar dentro da região subglótica (caixa torácica, pulmões e, principalmente, o diafragma);
2. Os **vibradores** transformam a coluna de ar em vibrações sonoras (laringe e, especialmente, as cordas vocais);
3. Os **ressoadores** modificam passivamente as vibrações sonoras (nariz, boca, e faringe);

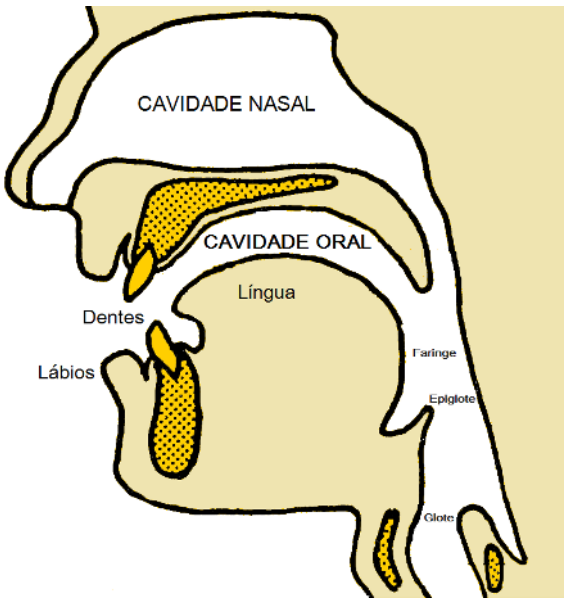

Figura 2.2: *Parte superior do aparelho fonador.*

4. Os **articuladores** se movimentam de modo a transformar as vibrações sonoras em fala (lábios, língua, dentes, palato duro e mole).

Observe na Figura 2.2 onde se localizam alguns destes órgãos do aparelho fonador. Basicamente, as vibrações sonoras são produzidas pelo afastamento e reaproximação das cordas vocais (ou pregas vocais). Tal movimentação das cordas é resultado da pressão exercida sobre elas pela coluna ar dentro da laringe [285]. O fluxo de ar resultante destas vibrações é chamado de **pulso glotal**, ou ainda, **fonte excitadora**. A Figura 2.3 ilustra o processo de geração do pulso glotal. A parte superior desta figura exibe a função da pressão do ar pelo tempo. A parte inferior desta figura mostra um diagrama simplificado do processo de controle da passagem de ar pelas pregas vocais.

É clara a contribuição do pulso glotal na composição da identidade sonora do falante. Entretanto, existem outros aspectos ligados à própria língua que devem ser observados. Então, algumas perguntas são levantadas: "*Como um ser humano pode diferenciar um fonema /a/ de um fonema /i/, ambos emitidos em uma mesma altura musical?*". Ou então: "*Como podemos associar fonemas equivalentes sendo estes emitidos em alturas musicais diferentes?*". As respostas destas perguntas estão relacionadas diretamente com as **frequências formantes** do sinal de voz.

Um formante é um pico de energia em uma região de frequências no espectro de um som, correspondente a uma concentração de energia em torno de uma determinada frequência (a frequência formante). Matematicamente, as frequências formantes são resultantes do processo de filtragem da fonte de excitação glotal pelas ressonâncias produzidas pelo *trato vocal*, formadas pela configuração dos ressoadores e articuladores da fala (ver Seção 2.2).

Na fala, as frequências formantes são responsáveis por manter o mesmo fonema independentemente da frequência fundamental emitida pelo falante. Assim, cada fonema distinto possui um conjunto de frequências formantes, que juntamente definem o contorno do **envelope** ou **envoltória espectral**.

De uma forma geral, podemos reconhecer as vogais a partir das regiões formânticas principais de cada fonema. Tais regiões se caracterizam por serem aglutinadores de frequências formantes mais importantes para especificação de um respectivo fonema. A região formântica da vogal /a/,

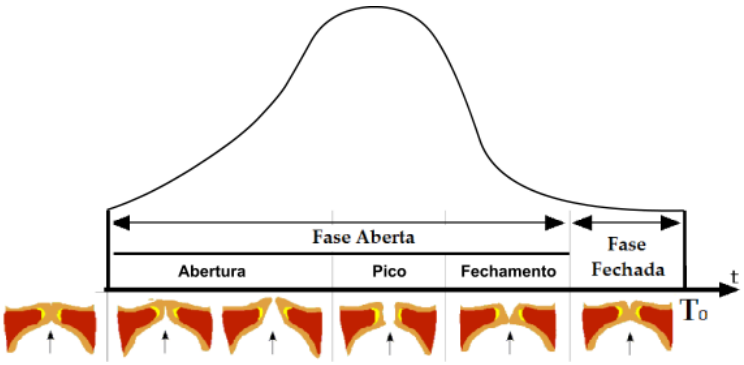

Figura 2.3: *Processo de produção do pulso glotal.*

por exemplo, compreende a faixa de frequências entre 800 e $1200Hz$. As vogais /o/ e /u/ possuem regiões formânticas na faixa de 400 a $600Hz$ e 200 a $400Hz$, respectivamente. Para as demais vogais /e/ e /i/ existem duas regiões formânticas principais: de 400 a 600 e 2200 a $2600Hz$ para a vogal /e/ e de 200 a 400 e 3000 a $3500Hz$ para a vogal /i/

Em vogais, os dois primeiros formantes são determinados principalmente pela posição da língua. O formante F_1 tem uma frequência mais alta quando a língua está mais baixa, ou seja, quanto maior for a abertura da vogal, maior é a frequência que aparece em F_1. O formante F_2 tem uma frequência mais elevada conforme o posicionamento da língua, ou seja, quanto mais "anterior" é uma vogal, maior é F_2. A Tabela 2.1 exibe a posição típica dos formantes F_1 e F_2 na pronúncia das vogais da língua portuguesa.

Vogal	**Formante** F_1	**Formante** F_2
/a/	$1000Hz$	$1400Hz$
/e/	$500Hz$	$2300Hz$
/i/	$320Hz$	$3200Hz$
/o/	$500Hz$	$1000Hz$
/u/	$320Hz$	$800Hz$

Tabela 2.1: *O primeiro e a segundo formante das vogais comuns na língua portuguesa.*

Além dos formantes, existem outras características específicas da fala como, por exemplo, os *elementos prosódicos*, que nos permitem discriminar um indivíduo dentre outros falantes a partir da altura musical, duração e energia da voz pronunciada.

Prosódia: Pitch, Duração e Energia

Em termos de linguística, a prosódia está relacionada com a correta acentuação das palavras tomando como padrão a língua considerada culta. Além disso, a prosódia relaciona todos os elementos rítmicos, melódicos e de dinâmica da fala. Tais fatores devem ser levados em consideração na tarefa de conversão de voz. Muitas vezes, até mesmo entre os seres humanos, um único erro de pronúncia torna uma simples sentença ininteligível.

A **altura** musical presente em uma voz humana, respectiva à frequência fundamental percebida por cada indivíduo, é conhecida também como **pitch**[2]. Como em vários outros aspectos da música, a percepção das alturas musicais é relativa. Poucas pessoas têm a capacidade de perceber cada frequência de modo absoluto, e quando isto ocorre dizemos que elas possuem ouvido absoluto.

[2]Adotaremos o termo *pitch* em Inglês por falta de nomenclatura padronizada em Português na área de processamento de voz.

Cada ser humano tem a capacidade emitir sons numa determinada faixa de frequência, chamada de **extensão** vocal do indivíduo. De um modo geral, se observa que a percepção do pitch da própria voz emitida por um indivíduo o torna apto a manter uma certa frequência fundamental f_0 invariável em um curto período de tempo.

A voz humana, bem como a maioria dos sons naturais, são sons compostos por uma **série harmônica** adicionados de uma componente ruidosa. Um harmônico de um sinal qualquer é uma componente cuja frequência é um múltiplo inteiro da frequência fundamental (f_0). Os múltiplos não-inteiros são chamados de **parciais**. O conjunto de frequências harmônicas $\{f_0, 2f_0, 3f_0, \ldots\}$ é denominado **série harmônica**.

Outro aspecto importante na emissão de um som é sua **duração**. No som falado existem inúmeros trechos sonoros com diferentes timbres, na maioria das vezes numa mesma sentença pronunciada. A medida de velocidade em que uma sentença é pronunciada é chamada de **taxa de pronúncia**, ou **velocidade de pronúncia**.

Como se pode perceber, a prosódia contribui com grande parcela na informação relativa à identidade sonora de um falante. Nossa percepção humana associa por exemplo vozes mais agudas a mulheres e vozes mais graves a homens. Também o volume de som irradiado, bem como o ritmo de articulação, são muito importantes na identificação de um locutor.

2.2 Modelos Clássicos do Sinal de Voz

Embora a maioria dos autores [53; 85; 128] acreditem que a modelagem do processo de produção da voz é predominantemente linear, não existe uma palavra final sobre o tema. Alguns pesquisadores pressupõem a não-linearidade da voz e se especializam em técnicas não-lineares sofisticadas [97; 102], e têm obtido êxito nos últimos anos. Não obstante, este trabalho considera que um modelo linear é mais apropriado para o desenvolvimento de sistema de conversão da voz. Seguem algumas justificativas para a escolha desta abordagem:

- A produção da voz pode ser descrita aproximadamente por comportamentos lineares, visto que as cordas vocais produzem uma fonte de excitação estável semi-periódica, cabendo aos ressoadores modelar o espectro do sinal (como um filtro).
- O arcabouço conceitual dos sistemas lineares fornece poderosas ferramentas para análise e síntese de comportamentos lineares.
- Quase por unanimidade os autores modelam o processo de produção de voz a partir de sistemas fonte-filtro, disponibilizando assim diversas ferramentas lineares para análise/síntese destes sinais.

Dado o direcionamento para a modelagem linear dos sinais de voz, passemos à especificação do modelo clássico de produção de voz, explorando a estreita relação entre as componentes de fonte e filtro do sistema.

2.2.1 Modelo Fonte-Filtro

Em geral, o **modelo fonte-filtro** decompõe os fenômenos acústicos em duas partes supostamente independentes: a **fonte de excitação** sonora, componente alimentadora do sistema, e o

filtro acústico que deforma a fonte, resultando assim no som irradiado [289]. Dado o processo de produção da voz abordado na seção anterior, podemos associar facilmente o **pulso glotal** à componente da fonte de excitação, bem como os demais elementos, isto é, os ressoadores e articuladores, ao filtro modelador do sinal de excitação, o **trato vocal**.

A Figura 2.4 ilustra o processo de composição da voz produzida pelo aparelho fonador, sob um ponto de vista de processamento de sinais. O som produzido pelas cordas vocais atravessa a região do trato vocal, e é consequentemente conformado segundo a configuração dos ressoadores e articuladores. O resultado final é o som irradiado.

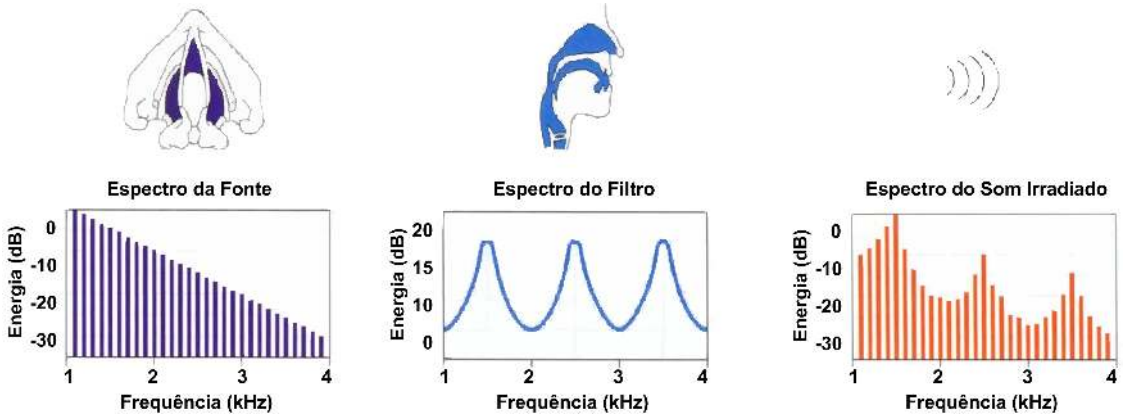

Figura 2.4: *Modelo de geração do som irradiado pelo sistema fonador.*

O trato vocal é, sem dúvidas, a componente que mais discrimina um falante de outro, dada a complexidade da configuração dos elementos ressonadores. Para se ter ideia da importância do trato vocal na composição fonética, tome como exemplo a pronúncia do fonema /e/. Perceba que enquanto este está sendo pronunciado, ao se levantar lentamente o centro da língua, gradualmente, o som passa a ser reconhecido como o fonema /i/[3]. Neste contexto, os ressoadores e os articuladores realizam um tratamento vocal no sinal de excitação, daí a terminologia dada ao trato vocal.

O pulso glotal, embora simples, é responsável por controlar elementos de prosódia de extrema importância na caracterização do falante: o pitch e a energia. Ao representar digitalmente os sons produzidos pelas pregas vocais, o sinal resultante é uma sequência quase-periódica composta por uma soma de harmônicos, cuja frequência fundamental se aproxima do pitch. Os sons produzidos por este tipo de fonte são conhecidos como sons *vozeados* (do inglês, *voiced*) [106]. Todas as vogais são exemplos de sons vozeados. Em contrapartida, sabe-se que a fonte de excitação também pode ser caracterizada pela ausência de vibrações das cordas vocais. Neste caso, uma coluna de ar passa livremente pelas pregas vocais, ganhando volume sonoro ao atravessar a faringe, e finalmente, sendo modelada pelo trato vocal. Os sons produzidos neste caso apresentam um comportamento de um sinal ruidoso, como o *ruído branco Gaussiano*. Estes sons gerados sem a excitação glotal são conhecidos como sons *não-vozeados* (do Inglês, *unvoiced*). Consoantes sem vibração de cordas acústicas, como as fricativas $[s]$ e $[x]$, bem como as plosivas $[b]$ e $[p]$ são exemplos de sons não-vozeados. Intuitivamente, costuma-se associar equivocadamente os sons vozeados somente às vogais e os sons não-vozeados às consoantes, embora haja uma certa relação entre ambas as partes. De fato, algumas consoantes como $[m]$, $[n]$, $[l]$ e $[r]$ são consideradas consoantes vozeadas. O que ocorre na maioria dos casos reais é uma mistura extremamente complexa e rica de componentes harmônicas

[3] Fonte: http://www.britannica.com/EBchecked/topic-art/457255/3598/Tongue-position-for-several-vowel-sounds

e ruído, todas ressonando ao mesmo tempo.

Conceitualmente, pode-se dizer que o som irradiado pode ser interpretado como a passagem da fonte excitadora harmônica (ou ruidosa) por uma função de transferência correspondente ao trato vocal, conforme se pode ver na Figura 2.5. Por um lado, o trato vocal funciona como um filtro que se aplica à fonte excitadora do sistema. Por outro lado, o pulso glotal é, normalmente, caracterizado como uma composição de harmônicos, que no domínio espectral corresponde a um trem de pulsos de amplitudes diversas, como foi exibido na Fig. 2.4.

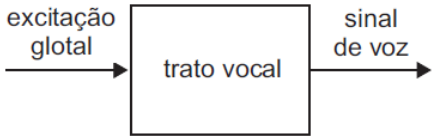

Figura 2.5: *Modelagem do processo de produção da voz por um sistema fonte-filtro.*

Em termos matemáticos, um sinal de voz $s(n)$ pode ser obtido pela convolução entre o sinal de excitação $x(n)$ correspondente ao pulso glotal e a resposta impulsiva $h(n)$ correspondente à configuração do trato vocal:

$$s(n) = x(n) * h(n). \tag{2.1}$$

No domínio da transformada $\mathcal{Z}$, observa-se a seguinte correspondência:

$$S(z) = X(z)H(z). \tag{2.2}$$

Outro aspecto importante desta modelagem é a premissa de estacionariedade em sentido amplo (*Wide-Sense Stationary – WSS*), ou seja, os primeiros e segundos momentos estatísticos de cada trecho curto tomado ao longo de um sinal de voz devem ser aproximadamente os mesmos. Sabe-se que em trechos curtos de voz (de aproximadamente 15 ms) os sinais são considerados quase-periódicos, e portanto, satisfazem esta propriedade. Tais trechos curtos de voz são denominados **segmentos** de voz. Desta forma, o sistema digital que modela a produção da voz deve considerar dois tipos de fonte de excitação: trem de pulsos e ruído branco com seus respectivos ganhos; e uma função de transferência (um filtro digital) que corresponde ao trato vocal, levando em conta a composição do sinal em pequenos segmentos consecutivos. As modelagens a seguir serão apresentadas visando a representação destes pequenos segmentos quase-periódicos de voz, em vez do sinal como um todo.

A modelagem do sistema fonte-fonte se divide em duas partes: a modelagem do trato vocal (componente do filtro) e a modelagem do pulso glotal (componente da fonte).

Modelagem do Trato Vocal

Existem vários métodos para a modelagem da componente do filtro. Normalmente, costuma-se realizar o tratamento dos segmentos vozeados com mais cuidado, por serem considerados a parte mais significativa da voz, em termos perceptuais, já que nosso ouvido é menos sensível a detalhes de alta frequência, como é o caso da maioria dos ruídos não-vozeados.

Como dito anteriormente, a voz é resultante do som produzido por uma fonte de excitação, modificado por um sistema ressonador. Qualquer mudança na forma e posição das estruturas do trato vocal resulta em modificações acústicas relevantes no som irradiado. Tal configuração do trato

vocal é percebida visualmente no espectro como o envelope que tangencia os picos correspondentes às frequências harmônicas do espectro, conforme pode ser visto na Figura 2.6.

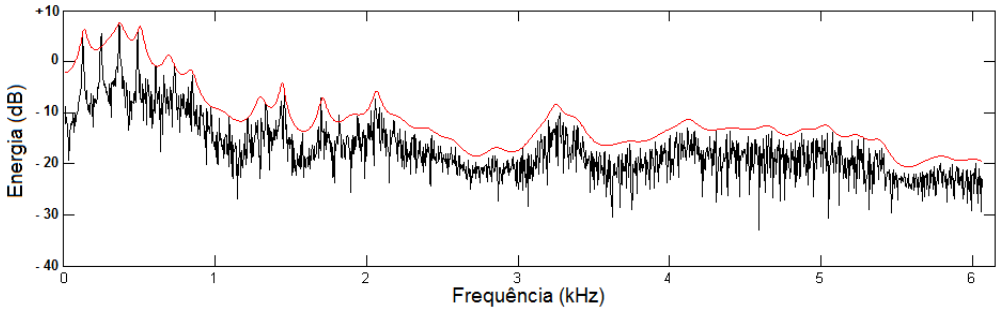

Figura 2.6: *Visualização do envelope espectral da vogal [a].*

A estimação do envelope (ou envoltória) espectral é uma das mais importantes e bem estabelecidas tarefas gerais em Processamento de Sinais, e particularmente em tecnologia de fala. Inúmeras aplicações, como reconhecimento de fala, compressão de sinal de voz, reconhecimento de locutor, síntese de fala ou mesmo conversão de voz, exigem uma representação fiel do envelope espectral usando poucos parâmetros. Além disso, a maior parte destas aplicações requerem que o envelope seja representado usando um número fixo de coeficientes. Em contrapartida, quando o sinal de fala é reconstruído a partir do envelope, como é o caso da síntese estatística de fala e da conversão de voz, é crucial que o envelope estimado represente adequadamente o trato vocal.

Existem muitas maneiras de se representar tal envelope espectral, sendo que não existe nenhuma que seja universalmente aceita como representação canônica. Por exemplo, no contexto deste trabalho, a representação do envelope espectral usando somas de funções paramétricas foi adotada, a qual tem sido bastante estudada recentemente [72]. Um caso particular é a soma de Gaussianas [137]. Alguns trabalhos anteriores também foram desenvolvidos a fim de ajustar a soma de Gaussianas a uma representação espectral. Em 1996, Zolfaghari [316] apresentou à comunidade científica um método baseado no Algoritmo EM [38] para estimação formântica usando mistura de Gaussianas (GMM). O mesmo autor [317] também propôs uma modelagem Bayesiana do envelope espectral conhecido por STRAIGHT [112] usando GMMs. Embora este último apresente melhores resultados, ambos os métodos não foram capazes de modelar todas as nuances espectrais, visto que são mais apropriados para caracterizar regiões formânticas. Goshtasby e O'Neill [75] propuseram um sistema baseado no Algoritmo de Marquardt, cuja inicialização é determinada por derivadas e taxas de cruzamento por zero (*Zero-Crossing Rates – ZCR*).

Um outro tipo de representação espectral consiste em utilizar coeficientes de um filtro digital do tipo FIR aplicado sobre a fonte de excitação glotal. No estado-da-arte em modelagem espectral, os coeficientes de predição linear (LPC) são os mais utilizados na representação do envelope espectral. LPC é o método mais extensivamente utilizado para calcular o envelope espectral de sinais de voz. Este método tem uma taxa de ajuste constante em toda a faixa de frequências, e captura bem as regiões formânticas do espectro. No entanto, o LPC também introduz distorção nas baixas frequências, assim como em algumas frequências de ressonância. Em codificadores de celulares, as limitações da modelagem LPC são compensadas pela alta qualidade na representação dos erros de predição e também pela redução dinâmica do sinal, aplicando expansão de largura de banda, entre outras técnicas. Predição Linear Perceptual (do inglês, *Perceptual Linear Prediction* [92]) é

uma extensão da estimativa LPC que leva em conta a resolução não-linear em frequência segundo o sistema auditivo humano.

A *predição* de séries temporais sempre foi um tema de grande relevância, já que exemplos de séries temporais são muito abundantes em ciências econômicas, engenharias, em inúmeros fenômenos naturais, e em particular, na voz. A predição linear [87], propriamente, consiste na previsão de uma amostra futura a partir de uma combinação linear finita de amostras anteriores. Uma série temporal $s[n]$ pode ser estimada a partir de uma combinação linear de p amostras passadas:

$$\hat{s}[n] = \sum_{k=1}^{p} a_k s[n-k],$$

onde a_k são os coeficientes da predição (pesos da combinação linear) e p é a ordem da predição.

Os coeficientes desta combinação linear são calculados através da minimização do erro quadrático médio dado por:

$$E_m = \sum_n e[n]^2 = \sum_n (s[n] - \hat{s}[n])^2, \tag{2.3}$$

onde e é o erro (ou resíduo) da predição.

Existem algoritmos eficientes tais como o algoritmo de Levinson-Durbin [179] para o cálculo destes coeficientes a partir de métodos clássicos de minimização, como o Método da Autocorrelação [143], entre outros [9; 144; 160].

Uma vez que os coeficientes de predição linear a_k são encontrados, a Equação 2.3 acima pode ser usada para calcular o erro de predição da sequência $\hat{s}(n)$. Aplicando a transformada $\mathcal{Z}$ nesta equação temos:

$$E(z) = S(z) - \hat{S}(z) = S(z) - A(z)S(z) = S(z)(1 - A(z)) = S(z)H(z)$$

Desta forma, a partir da função de transferência $H(z) = 1 - A(z)$ podemos encontrar o resíduo $e(n)$ dado um sinal de voz $s(n)$, assim como a partir da inversa desta, isto é $H(z)^{-1} = \frac{1}{1-A(z)}$, podemos encontrar o sinal de entrada $s(n)$ dado um resíduo $e(n)$.

Uma vez que os coeficientes LPC estão relacionados com a parte previsível do sinal $s(n)$, o resíduo ou erro de predição $e(n)$ tende a ter um comportamento puramente aleatório. Assim, ao usar um número suficientemente grande de coeficientes, podemos supor que o sinal $s(n)$ foi gerado pela filtragem de ruído branco por um filtro de resposta ao impulso de duração infinita (IIR), cuja função de transferência é

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

Em sinais de voz é comum utilizar uma ordem pequena de predição, p entre 10 e 30, para uma frequência de amostragem entre 8 kHz e 44 kHz, o que faz com que o filtro IIR resultante do modelo LPC possua poucos polos, tornando-o incapaz de modelar todos os detalhes do espectro do sinal de voz original $s(n)$. Neste caso, a saída do modelo tende a aproximar-se do envelope espectral de $s(n)$ em torno de seus picos dominantes.

A Figura 2.7 exemplifica este fato, ao tomar um segmento de voz com 1024 pontos de um sinal de voz - o fonema /e:/ sustentado - amostrado a uma taxa de 44100 Hz. Nesta figura, foram estimados

dois envelopes espectrais a partir do filtro IIR $H(z)^{-1}$ com p igual a 32 e 120, respectivamente. Note que o espectro superior possui um envelope modelado com poucos polos, e assim, proporciona uma estimativa mais suave, com menos detalhes do espectro. Já no segundo caso, a elevada ordem do LPC (120) fez com que o envelope espectral se ajustasse melhor ao espectro do sinal, no entanto, este envelope traz informações muito específicas do sinal, situação essa comumente referenciada pelo termo em Inglês *overfitting*, e que pode vir a ser um problema dependendo do uso que se vá fazer deste envelope.

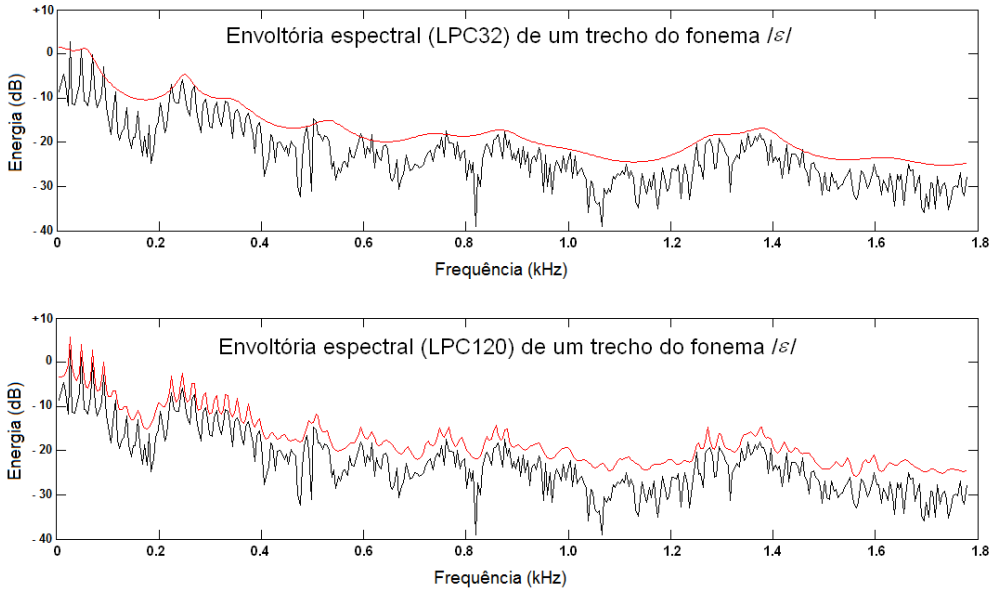

Figura 2.7: *Dois exemplos de estimativa do envelope espectral utilizando coeficientes LPC.*

Analogamente, podemos pensar que a resposta do filtro IIR $H(z)^{-1}$ ao impulso unitário também resulta numa estimativa do envelope espectral, considerando um número relativamente reduzido de coeficientes LPC. Tal resultado decorre da propriedade de linearidade (separabilidade das entradas) do filtro digital.

Uma vez que os fonemas são caracterizados pelas suas frequências formantes, e estas por sua vez, por coeficientes de representação de envelopes (como os coeficientes LPC), após a manipulação destes parâmetros acústicos deve-se garantir uma reconstrução fidedigna do espectro da fala. Resultados experimentais [163] mostram que modificações/interpolações nos coeficientes de predição linear (LPC) resultam em erros na reconstrução do espectro da fala. O problema é que este erro pode ocorrer em qualquer região do espectro, isto é, em qualquer frequência ressonante. Daí surge a necessidade de se utilizar uma nova técnica bastante popular na literatura: a LSF [132; 185; 186]. Nesta técnica, cada frequência ressonante no espectro da fala está diretamente relacionada a uma correspondente LSF. Se existe um erro numa frequência de ressonância causada pela transformação (quantização), tal erro é localizável. Assim, é comum utilizar coeficientes LSF para estimar frequências formantes de sinais de voz [36].

Os coeficientes LSF são calculados a partir das raízes dos polinômios simétrico $P(z)$ e anti-simétrico $Q(z)$, definidos a partir do polinômio $H(z)$ da Equação 2.2.1 como:

$$P(z) = H(z) + z^{-(p+1)}H(z^{-1}),$$

$$Q(z) = H(z) - z^{-(p+1)}H(z^{-1}),$$

onde $H(z) = \frac{P(z)+Q(z)}{2} = 1 - \sum_{k=1}^{p} a_k z^{-k}$, a_k são os coeficientes da predição (pesos da combinação linear) e p é a ordem da predição.

O polinômio $P(z)$ corresponde à configuração do trato vocal no instante em que a glote está fechada e $Q(z)$ quando a glote está aberta, e ambos os polinômios possuem suas raízes com módulo igual a 1, aparecendo intercaladas na circunferência de raio unitário sempre que as raízes de $H(z)$ têm módulo menor que 1. Sempre existe uma raiz de $Q(z)$ em $z = 1$ e uma raiz de $P(z)$ em $z = -1$. Os coeficientes LSF são dados pelos ângulos das raízes de $P(z)$ e $Q(z)$, e com essa informação é possível reconstruir perfeitamente $H(z)$ [250].

A Figura 2.8 ilustra, à esquerda, o lugar das raízes de $H(z)$, $P(z)$ e $Q(z)$ sobre o espaço $\Re \times \Im$, e à direita, o módulo da resposta em frequência de $H(z)^{-1}$ (filtro IIR com 6 polos) e as respectivas posições angulares dos zeros de $P(z)$ e $Q(z)$, os chamados coeficientes LSF. Note nesta figura que os pares de coeficientes LSF se aproximam dos pontos de ressonância de um trato vocal modelado por $H(z)^{-1}$.

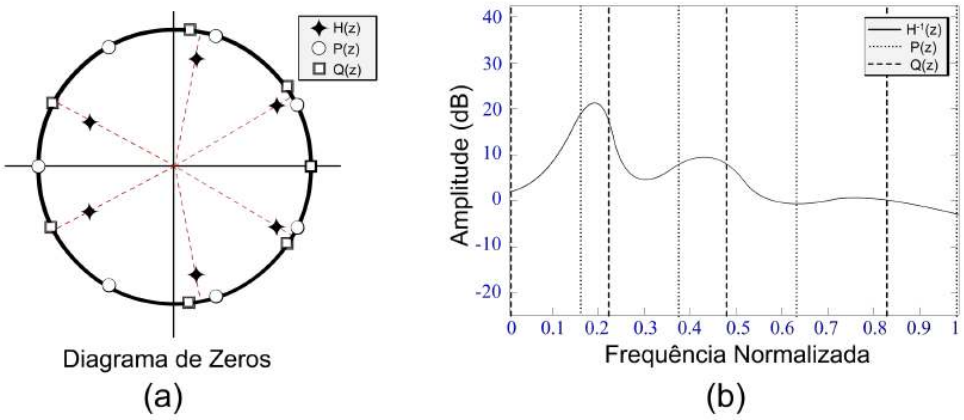

Figura 2.8: *(a) Lugar dos zeros dos polinômios $H(z)$, $P(z)$ e $Q(z)$ na circunferência de raio unitário; (b) Módulo da resposta em frequência do filtro $H(z)^{-1}$ IIR com 6 polos e as posições angulares dos zeros de $P(z)$ e $Q(z)$.*

As LSF possuem as seguintes propriedades:

1. Os zeros (raízes) de $P(z)$ e $Q(z)$ ocorrerão intercalados.
2. Os zeros de $P(z)$ e $Q(z)$ estão localizados sobre o círculo unitário.
3. O filtro é estável.
4. A interpolação dos coeficientes é possível.
5. Uma vez que as LSFs estão fortemente relacionadas entre si, elas podem ser quantizadas eficientemente.
6. Quando duas LSFs estão próximas uma da outra, um pico espectral é muito provável de ocorrer entre estas. Tal informação é muito útil para localização de regiões formantes e picos espectrais.
7. A adaptação mediante uma escala perceptual é muito mais fácil e natural, uma vez que as LSFs são valores atômicos de frequência.

A grande desvantagem de utilizar LSFs é a necessidade de se calcular as raízes de $P(z)$ e $Q(z)$, o que corresponde a resolver equações polinomiais de ordem p. O problema é ainda maior quando a taxa de amostragem e/ou a ordem de predição são altas, o que prejudica muito o desempenho do algoritmo que encontra as raízes destas equações [212]. Existem alguns trabalhos [115; 216] que tentam contornar este problema, como utilizar a transformada rápida de Fourier e coeficientes LPC [302], ou polinômios de Chebyshev [105].

Existem ainda outros métodos de modelagem do envelope espectral a partir de composição de frequências formantes. Tais formantes são representados por valores do centroide e largura de banda de cada região formante (ver Figura 2.9), obtidos de acordo com a estimativa dos polos que representam estas regiões [156].

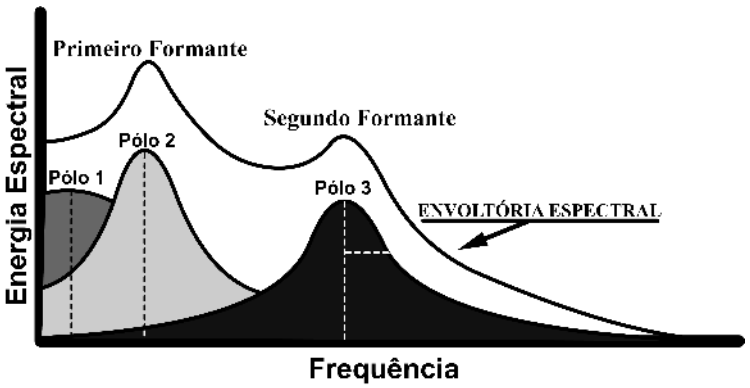

Figura 2.9: *Visualização dos formantes representados a partir de centroides e largura de banda dos polos.*

Num contexto de conversão de voz, tais centroides e larguras de banda são readaptados para realização da transformação do envelope espectral [169].

Outras técnicas mais específicas e sofisticadas também foram propostas como uma tentativa de se modelar intuitiva e consistentemente o trato vocal, baseadas na hipótese de que a representação do trato vocal seja suficiente para armazenar todas as particularidades de um falante. A normalização do trato vocal apresentada a seguir tenta compensar o efeito de dependência do trato vocal de um determinado falante a partir da aplicação de funções de deformação no eixo de frequência do espectro de potência do sinal.

Para reduzir a variabilidade entre os falantes, a normalização da magnitude do trato vocal (do Inglês, Vocal Tract Length Normalization – VTLN) é comumente utilizada para transformar características acústicas em sistemas de processamento de fala [68; 293; 308], principalmente em sistemas de reconhecimento de fala [57; 110; 187]. VTLN é uma técnica que adapta a escala do eixo de frequências dos vetores de parâmetros acústicos de modo que as observações sejam mais similares entre todos os falantes. Esta abordagem tem sido aplicada em síntese estatística de fala obtendo sentenças de fala independentes de falante, as quais podem ser readaptados a um falante específico por um conjunto pequeno de dados, e tem sido usada em conversão de voz [246; 247; 248] para transformar o espectro de uma dada classe acústica de um falante para outro. Esta é especialmente usual em sistemas gênero-independentes, uma vez que a diferença média dos ressoadores do trato vocal entre homens e mulheres é da ordem de 2 a 3 cm, o que ocasiona efeitos diretamente ligados às frequências formantes; por exemplo, em vozes femininas, as frequências formantes são cerca de 15% mais agudas do que em vozes masculinas para as mesmas vogais [57].

VTLN realiza uma transformação sobre o eixo das frequências, conhecida como função de de-

formação (*warping*), que mapeia cada frequência ω do seguinte modo:

$$\tilde{\omega}_\alpha(\omega) = f(\omega, \alpha).$$

Tal função f é uma função de deformação dependente de um *fator de empenamento* α. A maioria dos métodos para encontrar tais fatores de empenamento para VTLN utiliza métodos de otimização probabilísticos, a partir do critério de Máxima Verossimilhança (do Inglês, Maximum Likelihood – ML) [130].

O conjunto de frequências distorcidas $\tilde{\omega}$ corresponde à faixa de frequências normalizada em relação ao trato vocal, das frequências entre ambos falantes origem e destino no sistema de conversão de voz. A normalização do trato vocal corresponde a uma normalização linear do sinal no domínio *cepstral* [187; 190; 191].

Além de VTLN, outros métodos realizam a normalização do trato vocal. O método popularmente conhecido como STRAIGHT (Speech Transformation and Representation using Adaptative Interpolation of weiGHTed spectrum) realiza uma normalização do sinal de voz a partir da análise espectral adaptada ao pitch combinada a um método de reconstrução da superfície tempo-frequência da transformada de Fourier de tempo reduzido [112; 214]. A ideia central deste método é considerar que o sinal de excitação da voz é periódico, amostrado no espaço tridimensional definido pelos eixos do tempo, magnitude e frequência, que representa as características globais da fonte excitadora e aspectos de configuração do trato vocal. Em outras palavras, o STRAIGHT reconstrói o sinal em uma representação suavizada no espaço tempo/frequência que elimina a interferência da periodicidade no sinal de voz, a partir de uma análise adaptativa realizada anteriormente.

Finalmente, existe um outro tipo de representação espectral abordado nesta seção, conhecido como *coeficientes cepstrais*. O primeiro trabalho envolvendo o *cepstrum* foi realizado em 1963 por Bogert, Healy, e Tukey [22], onde foi inicialmente desenvolvido para caracterizar os ecos das ondas tectônicas provenientes de terremotos e explosões. Dentre as aplicações do *cepstrum* se destacam análise de eco em sinais sísmicos [181], detecção do pitch [174], deconvolução e processamento homomórfico de sinais [179; 253] e processamento de sinais de voz. A deconvolução homomórfica, particularmente, teve grande êxito no fim dos anos 60 na separação do sinal de excitação glotal da resposta impulsiva respectiva ao trato vocal na modelagem da fala em geral.

O termo "cepstrum" é uma combinação invertendo a ordem das primeiras quatro letras de "spectrum". Outros conceitos foram batizados segundo estas regras de recombinação de letras, tais como *quefrency* (derivado de frequency), *saphe* (de phase), *gamnitude* (de magnitude), *liftering* (filtering), *rahmonic* (de harmonic) e *repiod* (de period). O cepstrum, ou *cepstro*, é um tipo especial de espectro desenvolvido a partir de observações sobre o espectro de potência (módulo quadrado da FFT) do sinal original, e é definido como a transformada de Fourier do espectro logarítmico de um sinal. A deconvolução de um sinal de voz pressupõe que a faixa de frequências do sinal de excitação é linearmente separável (ou com pequena sobreposição no espectro) da faixa de frequências da região de ressonância do trato vocal. Se partimos do princípio de que um sinal de voz é obtido a partir da convolução da fonte de excitação com o trato vocal $s(n) = h(n) * x(n)$ (Equação 2.1), aplicando-se a transformada de Fourier obtemos $S(\omega) = H(\omega)X(\omega)$. Uma vez que a operação log tem a propriedade $\log(ab) = \log(a) + \log(b)$, temos que $\log(S(\omega)) = \log(X(\omega)) + \log(H(\omega))$. Aplicando-se

finalmente a transformada de Fourier inversa $\mathcal{F}^{-1}$, obtemos o cepstrum:

$$\mathcal{C}\{s(n)\} = \mathcal{F}^{-1}\{\log(S(\omega))\} = \mathcal{F}^{-1}\{\log(X(\omega))\} + \mathcal{F}^{-1}\{\log(H(\omega))\} = C_s(\omega). \tag{2.4}$$

A transformação cepstral inversa é definida como:

$$\mathcal{C}^{-1}\{C_s(\omega)\} = \mathcal{F}^{-1}\{\exp(\mathcal{F}\{C_s(\omega)\})\}. \tag{2.5}$$

Pelo pressuposto de separabilidade do sinal de voz quanto à fonte de excitação e o trato vocal, espera-se que no cepstrum sejam visíveis as regiões de contribuição do trato vocal, assim como a região onde se observa a fonte de excitação. Ou seja, fazendo-se uma "análise espectral do espectro" de um sinal de voz, observam-se componentes de baixas frequências respectivas à influência do trato vocal (que varia lentamente em função de ω) e componentes de altas frequências relativas ao sinal de excitação (que varia rapidamente em função de ω) [36]. A Figura 2.10 exibe o espectro do trecho de um sinal de voz vozeado (*voiced*). Acima, é possível perceber a frequência fundamental (primeiro lóbulo) e o pitch no espectro logarítmico deste sinal. Abaixo, podemos perceber no cepstrum deste sinal de voz a existência de um pico isolado na faixa de 12.5 milissegundos, que corresponde ao período de oscilação. Observa ainda no gráfico superior, em vermelho, o contorno espectral resultante da passagem de um filtro (*lifter*) passa-baixas no cepstrum deste sinal. Tal filtro elimina todas as *quefrências* acima de um certo limiar, conforme indicado no gráfico inferior.

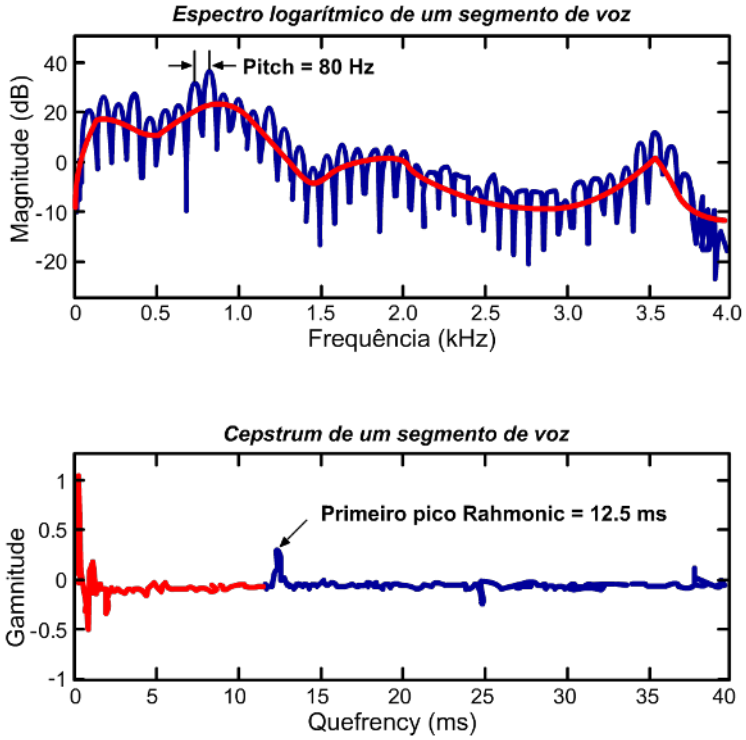

Figura 2.10: *Acima, o* log *do espectro de um sinal de voz; Abaixo, o cepstrum deste sinal de voz.*

Em casos onde a reconstrução do sinal não é necessária, o cálculo do cepstrum é simplificado de forma que somente o módulo do espectro do sinal é analisado. Em muitos casos toma-se simplesmente a potência do cepstrum, definida como:

$$\mathbf{P}[\mathcal{C}\{s(n)\}] = |\mathcal{F}^{-1}\{\log(|S(\omega)|^2)\}|^2. \tag{2.6}$$

Os coeficientes desta transformação são chamados de *cepstrum real*, são muito úteis em sistemas de detecção de pitch e têm se mostrado muito adequados quando aplicados ao reconhecimento de voz [36; 303]. Ao se transformar estes coeficientes, ou mesmo realizar uma filtragem (liftering) no cepstrum, devemos garantir que exista uma transformação inversa que reconstrói perfeitamente o sinal. Note que, dado um número complexo $A \exp j\omega$, temos que $\log(A \exp j\omega) = \log(A) + j\omega$, ou seja, podemos separar o cálculo dos coeficientes cepstrais em uma parte relativa ao módulo e outra relativa à fase da FFT [179].

Recentemente, a análise cepstral tem ganhado grande destaque na literatura [126; 188] quando utilizada juntamente com escalas perceptuais, em particular com os coeficientes mel-cepstrais (Mel-Frequency Cepstrum Coefficients - MFCC) [167; 190; 315]. Os coeficientes mel-cepstrais são tomados a partir do cepstrum de um sinal de voz adaptado à escala MEL. A Figura 2.11 apresenta a sequência de operações necessárias para obtenção do cepstrum na escala MEL.

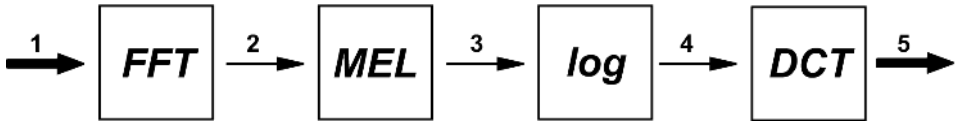

Figura 2.11: *Diagrama de blocos para obtenção do MEL-cepstrum.*

Esta figura mostra um procedimento organizado em 5 (cinco) passos:

1. Aplicar a transformada de Fourier do trecho de sinal de entrada.
2. Mapear as potências do espectro obtido sobre a escala MEL, usando *overlapping* de janelas triangulares.
3. Aplicar a operação log em cada uma das frequências empenadas na escala MEL.
4. Tomar a transformada discreta do cosseno da lista de potências log-mel.
5. Os coeficientes mel-cepstrais (MFCCs) são as amplitudes do espectro resultante.

Mais detalhes sobre este procedimento podem ser facilmente encontrados na literatura [187].

Os coeficientes Mel-Cepstrais (MFCC) são amplamente utilizados em quantização do envelope para tarefas de classificação de fala (por exemplo no reconhecimento de fala e reconhecimento de falante), e também podem ser utilizados na reconstrução da fala com bons resultados [30]. Entretanto, no uso de MFCCs o projeto do banco de filtros é fixo e não depende do sinal de entrada. Chazan et al. [30] sugerem que uma escolha apropriada da função base do banco de filtros é crítica para uma boa reconstrução.

Representações baseadas em formantes são uma alternativa interessante que permite caracterizar o envelope espectral usando poucos parâmetros. Entretanto, esta representação ainda não é suficientemente precisa e muito menos é facilmente estimada sob condições normais (não completamente controladas). Um modelo foi introduzido por Zolfaghari [316], onde o envelope espectral é representado como uma soma de algumas poucas funções Gaussianas, as quais tem uma razoável correspondência com as regiões formânticas principais. Modelos de misturas gaussianas são frequentemente usados em tecnologia de fala para representar funções densidade de probabilidade (FDP) de parâmetros espectrais [96]. Mesmo que a FDP não se ajuste estritamente ao modelo, esta fornece uma representação acurada em situações práticas, se um número suficiente de Gaussianas forem

adicionadas. Existem algoritmos bem conhecidos para estimar os parâmetros da FDP, tais como o algoritmo EM para estimação de máxima verossimilhança [301]. Muitas outras técnicas foram propostas para adaptar ou transformar a soma de Gaussianas de uma classe acústica para outra em conversão de gênero de voz [172] ou em conversão de voz propriamente dita [73].

A partir da modelagem do trato vocal é possível extrair a fonte de excitação por deconvolução. Dada esta observação, alguns sistemas optam por modelar diretamente a fonte de excitação, e por dualidade, estimar por diferenciação espectral a componente do filtro.

Modelagem da Fonte de Excitação

Por convenção, o sinal de excitação do sistema fonte-filtro também é chamado de **componente da fonte**. Tal componente que representa o pulso glotal é um sinal que além de atravessar o filtro (o trato vocal), determina características importantes da fala, tais como o pitch, energia e velocidade de pronúncia do sinal. É possível se estimar a forma de onda do pulso glotal a partir de resultados experimentais [210], sem necessariamente extrair o envelope espectral. Os pulsos glotais podem ser modelados através de um trem de impulsos quase-periódicos, representando as variações lentas da frequência fundamental. Dentre as modelagens, a modelagem residual é a mais comum e utilizada em sistemas de telecomunicações.

A modelagem residual é a modelagem mais simples do pulso glotal, que decorre a partir da modelagem do trato vocal, ou seja, dos envelopes espectrais. Pelo modelo fonte-filtro, um sinal de voz $s(n)$ é resultado da passagem do pulso glotal $x(n)$ pelo filtro do trato vocal $h(n)$, ou seja, em termos de função de transferência $S(z) = H(z)X(z)$. Assim, a componente da fonte pode ser obtida através do quociente $X(z) = \frac{S(z)}{H(z)}$. Desta forma, o envelope espectral modelado a partir de um conjunto de parâmetros (LPC, LSF, MFCC, ...) direciona a modelagem residual do pulso glotal e vice-versa, tornando-as inter-dependentes. Tal dualidade pode ser observada na situação em que dado um sinal de voz $s(n)$, conhecer $x(n)$ implica em obter $h(n)$, assim como a partir de $h(n)$ se obtém $x(n)$. Tal processo de obtenção de $x(n)$ a partir de $h(n)$ (ou vice-versa) é conhecido como *deconvolução*.

No entanto, um problema desta modelagem é a impossibilidade de se garantir que a modelagem do trato vocal não carregue consigo informações relativas ao pulso glotal, impedindo que a modelagem do pulso glotal seja fidedigna. Neste sentido, surgem as modelagens paramétricas do fluxo glotal.

O modelo Liljencrants-Fant [53; 54; 55; 56; 98] ou LF é um modelo clássico de fluxo glotal, caracterizado por quatro parâmetros independentes: frequência ω_g, amplitude E_0, uma constante α de crescimento exponencial do modelo e um parâmetro ϵ que corresponde à constante de tempo de uma recuperação exponencial, isto é, da fase de retorno desde o ponto de máxima descontinuidade até o fechamento total da glote.

O modelo LF pode ser representado a partir dos seguintes parâmetros temporais (os parâmetros T do modelo) provenientes da derivada do pulso glotal:

- T_0: instante de abertura glotal,
- T_p e U_0: instante e valor do máximo fluxo glotal,
- T_e e E_e: instante e valor absoluto do mínimo da derivada do fluxo glotal,

- T_a: fase de retorno que pode ser definida como a diferença absoluta entre o tempo T_e e a projeção da tangente da derivada do fluxo glotal em T_e,
- T_c: instante de fechamento glotal.

Observe na figura 2.12 tais parâmetros temporais em relação ao fluxo glotal (U_g) e sua respectiva derivada (dU_g) no modelo LF. No modelo LF, a derivada do fluxo glotal é composta por duas partes: a primeira representa a derivada do fluxo glotal desde a abertura glotal até o valor mais negativo; a segunda parte caracteriza o fechamento da glote.

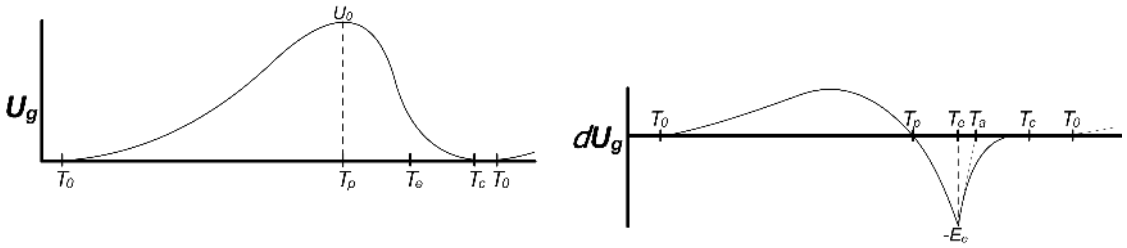

Figura 2.12: *O fluxo glotal U_g e sua respectiva derivada dU_g.*

A partir destes parâmetros temporais, podemos encontrar os quatro parâmetros independentes ω_g, E_0, α e ϵ do modelo LF:

$$\omega_g = \frac{\pi}{T_p};$$

$$E_0 = -\frac{E_e}{e^{\alpha T_e}\sin\omega_g T_e};$$

$$\alpha = \frac{-\ln(-\sin\pi\omega_g T_e)}{T_e};$$

$$\epsilon = \frac{1-e^{-\epsilon(T_c-T_e)}}{T_a};$$

E a partir destes, o fluxo glotal $g(t)$ é definido como:

$$g(t) = \begin{cases} E_o e^{\alpha t}\sin\omega_g t, & 0 \le t \le T_e \\ -\frac{E_e}{\epsilon T_a}\left[e^{-\epsilon(t-T_e)} - e^{-\epsilon(T_c-T_e)}\right], & T_e \le t \le T_c \le T_0 \end{cases}$$

onde $\int_0^{T_0} g(t)dt = 0$.

Observe que a extração do parâmetro ϵ depende da resolução de uma equação não-linear, e por esta razão, este parâmetro é um valor aproximado de forma que para um T_a pequeno, ϵT_a tenda a 1. Devido ao alto custo computacional para se estimar tal parâmetro neste modelo, modelagens alternativas de pulso glotal utilizando os mesmos parâmetros T do modelo LF são propostos [286]. Com o tempo, os modelos de representação evoluíram de modo a representar a componente da fonte usando osciladores associados às faixas de frequência espectrais.

O modelo de excitação multibanda, do Inglês *Multiband Excitation Vocoder* (MBE), que foi proposto por Griffin em 1987 e tem se tornado um modelo bastante popular para codificação e síntese em aplicações diversas [29; 71; 251], é um tipo de modelagem senoidal aplicada à componente de fonte do sinal. Este modelo é proposto a partir de observações em cada sub-banda do espectro de um sinal de voz. Tais sub-bandas são determinadas a partir da frequência fundamental, e por

esta razão estas sub-bandas são ditas sub-bandas harmônicas. O método considera que cada sub-banda harmônica pode ser vozeada ou não-vozeada. Além disso, considera-se que as regiões de alta frequência normalmente contêm componentes que não são apropriadamente modeladas por impulsos, e por esta razão, são modeladas a partir de ruídos.

O primeiro passo deste método é o estágio de análise. Primeiramente é determinado o pitch do trecho de voz a ser codificado. O autor sugere utilizar um método baseado em autocorrelação, no entanto, um detector robusto de pitch é necessário nesta fase, e por esta razão outros métodos de detecção de pitch podem ser utilizados [24; 215; 262]. O espectro de excitação é dividido em bandas harmônicas de acordo com o período do sinal. Um módulo de decisão vozeado/não-vozeado (*Voiced/Unvoiced*) é utilizado em cada sub-banda harmônica, sendo estas centradas em cada múltiplo de f_0 e cobrindo uma faixa de largura f_0 em torno deste. Em seguida, a estimativa do período de oscilação é refinada e os parâmetros são re-estimados, aumentando a robustez em casos de erro de detecção de pitch.

No segundo passo, o estágio de síntese, as amplitudes, fases e frequências centrais são interpoladas entre segmentos consecutivos. As porções vozeadas do sinal são sintetizadas usando os correspondentes parâmetros do modelo senoidal, enquanto que as porções não-vozeadas são sintetizadas a partir da multiplicação do ruído branco Gaussiano pelo envelope espectral no domínio da frequência. Então, é tomada a FFT inversa dos segmentos reconstruídos a fim de se obter a forma de onda destes no domínio do tempo. Finalmente, estas porções vozeadas e não-vozeadas são somadas a fim de obter os segmentos sintéticos. Tais segmentos são concatenados utilizando métodos de sobreposição com somas ponderadas (*Overlap Add*), similares aos métodos PSOLA [165].

A maior vantagem do modelo MBE é o fato de que o espectro da excitação é representado detalhadamente. É possível quantizar ou interpolar os parâmetros do modelo senoidal. As desvantagens estão relacionadas com os requerimentos de robustez na detecção do pitch, as decisões vozeado/não-vozeado em cada sub-banda harmônica, e principalmente, o problema da interpolação das fases. No entanto, progressos recentes em modelagem harmônica, bem como o aumento da acurácia dos detectores de pitch vêm contribuindo significativamente para a representação não só da componente de fonte, mas de ambas as componentes do sinal de voz sob uma perspectiva diferente, a qual será abordada na seção seguinte.

2.2.2 Modelo Harmônico-Estocástico

Essencialmente, o Modelo Harmônico-Estocástico [255] tem suas raízes estruturadas em modelos senoidais. Os modelos senoidais pressupõem que um trecho de fala contínua pode ser segmentado em pequenos segmentos quase-estacionários, os quais são localmente representados por uma soma de senoides, cujos parâmetros variam lentamente no tempo. A reconstrução do sinal é obtida pela soma desses osciladores, no qual cada componente é representada por sua respectiva frequência, amplitude e fase inicial. O modelo harmônico, por sua vez, é um tipo de modelo senoidal no qual as frequências em cada banco são múltiplos inteiros de uma frequência fundamental. Dada a flexibilidade na representação prosódica e espectral do sinal de voz por este modelo, tal abordagem foi adotada como modelo estrutural do sistema de conversão de voz apresentado por este trabalho, o qual será melhor explorado no Capítulo 3.

O primeiro modelo senoidal foi usado por Hedelin [88] em compressão de voz em 1981. A ideia dele foi a de usar um modelo de pitch independente para caracterizar cada faixa no espectro do

sinal. As amplitudes e fases das componentes senoidais foram estimadas usando filtros de Kalman, e as fases de cada senoide foram definidas como a integral da frequência instantânea associada.

Em meados dos anos 80, R. J. McAulay e T. F. Quatieri [152], propuseram o modelo senoidal aplicado à análise e síntese de voz, o qual conhecemos até hoje, e que tem sido utilizado em um ramo vasto de aplicações [141]. Esta ferramenta tem sido vista como um modelo completo para representar um sinal de voz por uma combinação linear de senoides. Em particular, o sinal de excitação glotal pode ser perfeitamente representado por este modelo, e além disto, este modelo proporciona uma flexibilidade maior quanto à modulação do pitch e ajustes na magnitude deste sinal.

Um modelo senoidal considera que o sinal de entrada $s^{(k)}(n)$ (em geral, o k-ésimo segmento estacionário) é composto pela soma de M senoides. Utiliza-se o modelo senoidal para aproximar uma sequência $s^{(k)}(n)$ de uma soma de osciladores

$$s^{(k)}(n) \approx \sum_{m=1}^{M} \Omega_m^k(n)$$

de modo que

$$\Omega_m^k(n) = C_m + A_m \cos\left(\frac{2\pi f_m^k n}{R} + \phi_m^k\right) + \varepsilon_m^k,$$

onde R é a taxa de amostragem, C_m é a constante DC, A_m^k é a amplitude de uma componente senoidal, f_m^k é a sua frequência, n é o tempo discreto, ϕ_m^k é a fase, e ε_m^k é o erro na componente k do modelo.

As amplitudes, frequências e fases de cada componente senoidal são obtidas a uma taxa de segmentos constante, usando um algoritmo de detecção de picos aplicado à Transformada de Fourier de Tempo Reduzido – STFT de cada segmento. A ressíntese do sinal de voz é feita de modo que as amplitudes e frequências de cada oscilador são interpoladas linearmente entre segmentos consecutivos, e as fases iniciais são interpoladas com o uso de polinômios de ordem cúbica.

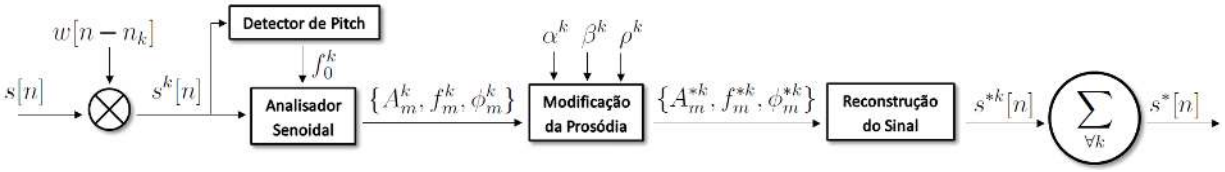

Figura 2.13: *O diagrama do Modelo Senoidal.*

A Figura 2.13 exibe um diagrama que explica o processo de análise, transformação de prosódia e síntese usando o modelo senoidal. Segundo este diagrama, um sinal de entrada s é segmentado pelo janelamento w em pequenos segmentos curtos $s^{(k)}$ de aproximadamente 15 ms. Posteriormente, cada um destes é submetido a um módulo de detecção de frequência fundamental f_0, e através do analisador senoidal, se obtém as M amplitudes A_m^k, frequências f_m^k e fases iniciais ϕ_m^k para cada oscilador. A partir daí, o sistema pode realizar as mudanças de prosódia, tais como o contorno de energia controlado pelo fator α^k, o contorno do pitch ajustado pelo fator β^k e a velocidade de pronúncia, variando o fator ρ^k. Finalmente, o sinal de saída é reconstruído com a concatenação de cada segmento s^{*k}, gerado pelos coeficientes modificados $\{A_m^{*k}, f_m^{*k}, \phi_m^{*k}\}$.

Algum tempo depois, o método *OLA - Overlap Add* foi agregado ao modelo senoidal [70]. Este

método utiliza otimização por mínimos quadrados ordinários [69] para encontrar valores ótimos de amplitude e fases iniciais de cada oscilador. O métodos dos mínimos quadrados é uma técnica de otimização matemática que procura encontrar o melhor ajustamento para um conjunto de dados tentando minimizar a soma dos quadrados das diferenças entre a curva ajustada e os dados (tais diferenças são chamadas resíduos). Maiores detalhes de como ajustar funções utilizando o método dos mínimos quadrados podem ser encontrados em [91]. Para obter um bom ajuste, o método dos mínimos quadrados não-linear precisa de bons valores de partida para as constantes C_m, as amplitudes A_m e as frequências ω_m. Um bom valor inicial para cada C_m pode ser obtido calculando-se a média da sequência. Bons valores iniciais para cada frequência ω_k e amplitude A_k podem ser obtidos a partir de frequências dominantes do sinal de entrada. Uma particularidade deste modelo é que o sinal final é reconstruído por meio de somas sobrepostas de segmentos passados por uma janela triangular. Observe que a interpolação das amplitudes, frequências e fases iniciais na fase de ressíntese não é mais necessária. Já no caso de modificação prosódica, as fases iniciais de cada componente harmônica devem ser cuidadosamente atualizadas. Sabe-se que cada fase contém dois termos distintos: o termo inerente à forma de onda do sinal, o chamado termo do trato vocal, e o termo linear em frequência, advindo da atualização da frequência instantânea ao longo do tempo. Sendo assim, para tarefas simples como o deslocamento do pitch, é necessário isolar ambos os termos, e tratar separadamente o termo linear do termo correspondente ao trato vocal. Esta série de complicações com os deslocamentos de fases fez com que Banga et al [12] propusessem uma abordagem utilizando segmentos sincronizados com o período de oscilação do sinal, os chamados **segmentos pitch-sincronizados**.

A abordagem senoidal usando segmentos pitch-sincronizados evita de certa forma distorções causadas pela alteração do pitch e velocidade de articulação do sinal de voz através do uso de marcações periódicas do sinal no tempo, conhecido como **marcas de pitch**. O resultado da análise pitch-sincronizada mostra que as fases iniciais não possuem o termo linear associado à atualização da frequência instantânea de um segmento para o outro, ou seja, a mudança da frequência fundamental não implica em uma prévia atualização das fases. Por outro lado, as estimações das marcas de pitch não muito precisas degradam significativamente o sinal ressintetizado.

Foi então que O'Brien e Monaghan [177] decidiram simplificar um pouco o modelo e descartar as marcas de pitch e restringir todas as frequências de modo que $f_m^k = mf_0^k$. Surge então o primeiro sistema harmônico para análise e síntese de voz. Novamente, os polinômios de 3^a ordem de McAulay e Quatieri foram usados para modelar as fases instantâneas, com os cálculos um pouco menos complexos de subtração dos termos lineares destas.

Como evolução deste modelo, surgem os primeiros protótipos de modelos híbridos, a saber, mesclando a modelagem estocástica à modelagem harmônica do sinal de voz. A ideia embrionária de se desenvolver um sistema híbrido surge em 1988 com Griffin e Lim [77] propondo um novo sistema chamado *Multiband Excitation Vocoder*. Neste modelo, o sinal é modelado por uma componente de fonte cujo espectro é composto por um trem de pulsos linearmente espaçados (de acordo com a frequência fundamental) adicionado de uma componente ruidosa e uma componente de filtro, tomada basicamente pela versão suavizada do espectro, o envelope espectral. Sabe-se que este foi o primeiro modelo e se preocupar com a parte ruidosa e harmônica na componente de fonte do sinal de voz.

Um ano depois, Serra [225; 226], propõe um arcabouço muito parecido com o de McAulay e

Quatieri, porém com a adição de uma componente ruidosa (ruido branco Gaussiano) modelada a partir da passagem do ruído branco Gaussiano por um filtro ressonador, a fim de obter maior naturalidade no sinal sintetizado. Esta componente residual é muito bem caracterizada apenas pelo envelope do espectro de potência, modelada com coeficientes LPC de baixa ordem. No caso deste modelo, as fases iniciais de cada oscilador são descartadas e as fases de cada segmento são geradas recursivamente tomando-se as fases finais dos segmentos anteriores. Como o termo das fases iniciais associado ao trato vocal é também descartado, este modelo se torna inaplicável em sinais de voz. No entanto, o modelo continua a ser bem aceito em outras aplicações que manipulam sinais musicais.

Finalmente, o Modelo Harmônico mais Ruído proposto por Stylianou [255; 256] (do Inglês, *Harmonic-plus-Noise Model* – HNM) foi o primeiro modelo híbrido, cuja representação do sinal de voz é dividida em duas partes: a parte harmônica do sinal corresponde às componentes quase-periódicas do sinal de voz e a parte ruidosa do modelo corresponde às componentes não periódicas do sinal, que correspondem a ruídos consonantais e variações aperiódicas da excitação glotal. A Figura 2.14 ilustra as adaptações feitas por Stylianou em cima do modelo senoidal proposto por McAulay e Quatieri. A principal diferença deste diagrama em relação ao anterior é um módulo analisador da componente estocástica que devolve um filtro digital $H^k(z)$ com o envelope espectral da componente não-harmônica do k-ésimo segmento de voz. Além disso, como todo modelo harmônico, a representação das frequências harmônicas são múltiplos inteiros da frequência fundamental f_0^k, o que compacta ainda mais a representação. Note que na fase de síntese, dois sinais são somados para compor o sinal reconstruído: a parte harmônica s_h^{*k} e estocástica s_s^{*k} do sinal. Neste esquema, o ruído usado é o ruído branco Gaussiano, e $w[.]$ é uma janela de Hamming centrada com mesmo tamanho do segmento $s^{(k)}$.

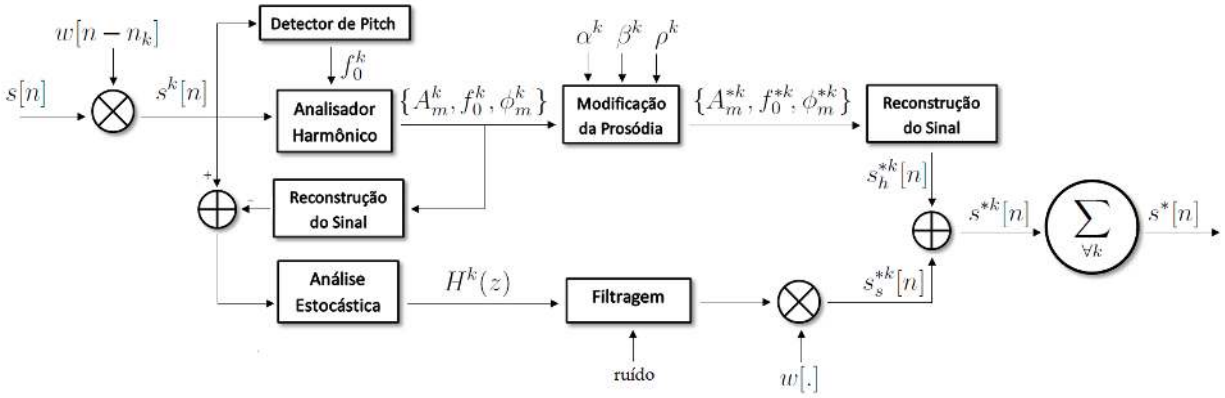

Figura 2.14: *O diagrama do Modelo Harmônico.*

Mais tarde, uma versão mais refinada do modelo HNM passou a ser a mais popular dentre os métodos propostos por Stylianou [258], adequando detalhes de janelamento e outras pequenas contribuições. Tal proposta ficou conhecida como Modelo Harmônico-Estocástico (*Harmonic plus Stochastic Model* – HSM), o qual será melhor apresentado a seguir. Por simplificação, o modelo será apresentado em dois módulos: Análise e Síntese.

Módulo de Análise do Sistema HSM

Segue uma descrição detalhada da análise do sinal de entrada, detalhando cada passo do procedimento. A decomposição do sinal de entrada é o primeiro passo dentro do módulo de análise do

modelo HSM.

Sinal de Entrada: O sistema HSM é uma ferramenta robusta na representação de **sinais de voz** e outros tipos de sinais predominantemente harmônicos. A coleta de informações fidedignas destes sinais depende diretamente da qualidade do áudio em questão. Embora a obtenção do áudio seja uma etapa razoavelmente simples, alguns cuidados importantes devem ser tomados a fim de evitar problemas como:

- Gravação de áudio com baixa qualidade: muitas vezes, as condições externas no momento da gravação (ruídos de fundo), os equipamentos inadequados e até mesmo a captação inadequada do áudio (por exemplo, gravar muito próximo do microfone) podem resultar em sinais de áudio defeituosos. Estes defeitos podem comprometer significativamente a modelagem do sinal, uma vez que a maioria das consoantes apresentam comportamento ruidoso. Um dos critérios de escolha da base de dados para teste é sem dúvida a qualidade do áudio

- Representação muito pobre do arquivo de áudio: a escolha de baixas taxas de amostragem, bem como uma quantização pobre das amostras do sinal de áudio (normalmente representado com a codificação PCM, muitas vezes é insuficiente para uma conversão de voz fidedigna. Normalmente são adotadas taxas de amostragem a partir de 8 KHz [310] ou 16 KHz [125], com amostras quantizadas acima de 16 bits por amostra, sendo que taxas de amostragem maiores implicam em uma conversão de voz de maior qualidade [275].

Todos estes aspectos devem ser observados ao princípio de uma modelagem harmônica/estocástica.

Segmentação: A segmentação do sinal de voz é uma outra etapa fundamental no processo de análise e síntese do sinal de áudio. Embora os aspectos globais sejam percebidos no contexto de sentenças, as transformações dos parâmetros modelados serão feitas em escalas menores. Sendo assim, na fase de segmentação, o segmentador deve modelar localmente as informações globais de prosódia, armazenando-as transparentemente para posterior modelagem na forma de momentos estatísticos ou outros modelos de representação [219] como no caso de conversão de voz.

Sabe-se que a escolha do tamanho do segmento pode seguir de duas abordagens distintas: a primeira corresponde à utilização de um tamanho fixo, e a segunda corresponde a escolher o tamanho do mesmo de acordo com a frequência fundamental, os chamados *segmentos pitch-synchronous* [164].

Como visto na seção anterior, os segmentos pitch-sincronizados [50; 109; 246; 247] pressupõem que o tamanho dos segmentos é ajustado de tal forma que contenha um número inteiro de períodos de um sinal de voz quase-periódico. Neste caso, as marcas de pitch são essenciais para separar um segmento do outro. Evidentemente, a utilização desta técnica considera que um sinal de voz é vozeado, uma vez que um trecho não-vozeado não possui uma série harmônica. No caso de segmentos não-vozeados, o segmento assume um tamanho fixo pré-determinado. A dificuldade em se utilizar segmentos pitch-sincronizados é a alta sensibilidade em erros de marcações de períodos associados ao pitch.

Na caso da abordagem fixa, não existe um consenso geral quanto ao tamanho ideal destes segmentos. Na teoria, segmentos menores do que 15 ms são considerados quase-periódicos para sinais de voz. Isto se deve ao fato de que tanto o processo de produção do pulso glotal pelas

Janela	Largura do lobo central a $-3dB$ (em bins)	Largura do lobo central a $-6dB$ (em bins)	Nível máximo do lóbulo secundário (em dB)
Retangular	0.88	1.21	-13
Hamming	1.30	1.81	-43
Hanning	1.44	2.00	-32
Blackman-Harris	1.62	2.27	-71
Blackman	1.64	2.30	-58
Flap Top	2.94	3.56	-44

Tabela 2.2: *Larguras de banda do lóbulo central e secundários de janelas clássicas*

cordas vocais quanto a configuração dos órgãos ressoadores responsáveis pelo trato vocal se mantêm aproximadamente estáveis em períodos com estas durações [166; 281]. No entanto, a informação harmônica contida dentro deste pequeno fragmento de voz é muitas vezes insuficiente para a detecção segura do pitch. Por este motivo, na prática, segmentos com tamanhos entre 15 ms e 25 ms são frequentemente adotados [2; 32; 80], sob a hipótese de estacionariedade do trecho em sentido amplo (WSS) [232]. No sistema HSM apresentado, adota-se a abordagem fixa com 16 ms, no qual existe sobreposição de 50% entre segmentos consecutivos, a fim de melhorar a estimação harmônica na porção central do segmento e aumentar o alcance de cada segmento. A escolha do tamanho do segmento também está relacionada com a escolha da taxa de amostragem, uma vez que estas duas determinam a acuidade espectral.

Outro aspecto importante é a escolha da janela usada na segmentação de modo que o janelamento possibilite uma estimação precisa das amplitudes e frequências harmônicas. Sabe-se que a multiplicação de um sinal no domínio do tempo implica na convolução dos respectivos espectros no domínio da frequência [36]. O janelamento ideal deveria ser avaliado em uma janela cuja resposta de magnitude fosse tão estreita quanto possível no lóbulo central para que haja maior resolução dos picos espectrais (sem interferências construtivas), bem como possuir interferência (quase-)nula dos lóbulos secundários a fim de evitar a degradação do valor real das amplitudes estimadas. No entanto, existe um problema intransponível típico da resolução tempo-frequência, em que a resolução alta do lóbulo central, implica em alto grau de interferência dos lóbulos secundários, e vice-versa. Neste ponto, o problema dual entre a largura de banda do lóbulo central e a interferência dos lóbulos secundários continua sendo um entrave na escolha da janela ideal para segmentação do sinal.

A Tabela 2.2 exibe valores do largura do lóbulo principal, responsável pelo espalhamento espectral do lóbulo central, bem como o nível máximo do lóbulo secundário cujo valor alto implica em um vazamento espectral lateral decorrente da interferência dos lóbulos secundários.

Dados os resultados da tabela anterior, é possível observar o problema dual citado anteriormente. Sendo assim, o que se costuma fazer é utilizar as janelas para propósitos específicos. Por exemplo, usar a janela retangular, cuja largura de banda do lóbulo central é a menor dentre todas, para obtenção da frequência dos osciladores, e usar outra janela com menor interferência dos lóbulos laterais, como a janela Blackman-Harris, para estimativa das amplitudes dos osciladores, por exemplo.

Grande parte dos pesquisadores trabalham com a janela de Hamming, por ser esta mais robusta em situações gerais, em especial na presença de ruído, uma vez que seu lóbulo central é relativamente

estreito, e sua taxa de decaimento lateral é aceitável [51; 258]. Supondo selecionada a janela de segmentação, a largura da janela e o deslocamento da mesma, a segmentação é imediata. Uma vez realizada a segmentação, o sistema passa a detectar a presença de uma frequência fundamental dominante.

Detecção da Frequência Fundamental: Antes de detectar uma possível frequência fundamental, é necessário saber de antemão se de fato o trecho de voz é vozeado ou não. A decisão *voiced/unvoiced*, ou decisão V/UV como é conhecida, está intimamente relacionada ao problema de estimação de uma possível frequência fundamental [146]. Ou seja, a própria existência de uma frequência fundamental elimina a hipótese de que o segmento seja não-vozeado. Embora decidir se um trecho de voz é vozeado ou não seja um problema bastante antigo em processamento de sinais digitais [180], o mesmo continua sendo bastante estudado nas últimas décadas [85; 235; 262]. Existem inúmeras técnicas que abordam este problema utilizando medidas tais como razão de bandas de frequências [173], redes neurais [34], coeficientes wavelets [101], taxas de cruzamento com zero [10], coeficientes de predição linear [199; 200] e cepstrum [3], entre outras. No entanto, existem atualmente um número grande de algoritmo detectores de pitch (do Inglês, *Pitch Detection Algorithm – PDA*) que decidem ao mesmo tempo se um segmento é vozeado ou não.

Um algoritmo detector de pitch de voz tem o propósito de localizar uma frequência fundamental de um sinal quase-periódico e que esteja dentro de uma faixa de frequência adequada em relação à extensão vocal humana. Infelizmente não existe um algoritmo ideal para detecção do pitch de voz em todo e qualquer caso realístico, mas existem bons algoritmos adequados às condições específicas do registro de voz, tais como a presença de ruído de fundo, predominância vocálica, etc. Existem algoritmos propostos nos domínios temporal e espectral, além de algoritmos que usam ambos os domínios.

Um bom algoritmo no domínio do tempo é o algoritmo YIN [31], o qual utiliza a autocorrelação clássica, combinada à função da diferença (*Average Magnitude Difference Function – AMDF*) de um sinal x de tamanho N definida como

$$d(\tau) = \sum_{n=1}^{N} (x[n] - x[n+\tau])^2,$$

para todos os *lags* $\tau = [1, N]$, para encontrar picos mais significativos, candidatos à f_0. Dentre os picos candidatos, toma-se aquele de menor frequência entre eles. Embora seja um algoritmo razoavelmente simples, este tem taxas de acerto bastante altas em sinais com pouco ruído, apresentando poucos erros de oitava (os chamados erros grosseiros).

Ainda no domínio temporal, um outro algoritmo, conhecido como Praat [21], é um algoritmo robusto na presença de ruído, especializado em detecção de pitch em sinais de voz. O mesmo utiliza como base a função de autocorrelação normalizada para pré-estimativa das prováveis frequências fundamentais (uma estimativa por segmento), e posteriormente refina a busca usando programação dinâmica, cuja função de minimização do erro penaliza tanto os erros grosseiros, quanto pondera os segmentos vozeados. Uma maneira de se aumentar o grau de confiança desta estimação inicial é alargar o tamanho da janela de análise, mantendo uma sobreposição entre os segmentos (normalmente de 50%).

A abordagem no domínio da frequência normalmente se aplica em casos de detecção de sons

polifônicos, usando ferramentas como relação harmônico-ruído (HNR) [21] e análise cepstral [181], entre outras. E dentre os algoritmos que usam a abordagem mista, se destaca o algoritmo YAAPT [307], uma recente proposta que monta uma lista de candidatos advindos da função de autocorrelação no domínio temporal, e outra lista de candidatos extraídos diretamente dos picos do espectro de magnitude. A partir destas listas de possíveis candidatos, o algoritmo também usa programação dinâmica para decidir quais são as frequências fundamentais mais adequadas, segundo um critério de minimização de erro.

No contexto deste trabalho, adotada a taxa de sobreposição entre segmentos de 50%, o algoritmo PRAAT especializado em sinais de voz foi usado para obtenção da frequência fundamental, o que não descarta a possibilidade de usar qualquer outro PDA robusto para a realização de tal tarefa. Uma vez detectada a presença de uma frequência fundamental dominante no segmento de voz, quase todos os métodos refinam a busca usando o método de interpolação parabólica. Outro fator importante é a definição da faixa de frequências fundamentais para as quais o segmento é considerado vozeado (normalmente entre 80 e 300 Hz).

Estimativa da Parte Harmônica: A partir da detecção da frequência fundamental (e consequentemente de suas componentes harmônicas), a estimativa da parte harmônica do sinal consiste em encontrar as amplitudes e fases de cada uma das componentes harmônicas. O método mais simples de estimativa, e bem explorado num contexto de sinais musicais, consiste em usar um algoritmo detector de picos harmônicos a fim de localizar componentes senoidais no domínio espectral da STFT do sinal [152]. Nesta abordagem evidentemente existe a necessidade de refinar a busca do pico espectral, dada a baixa resolução dos bins da STFT. Na maioria dos casos, a interpolação parabólica é adotada. Ao final, as amplitudes, fases e frequências de cada oscilador são encontradas a partir dos picos espectrais estimados. Em todo caso, a presença do ruído de fundo torna imprecisa a detecção das amplitudes e fases iniciais, além de ignorar a interferência entre os lóbulos laterais na configuração espectral.

Trabalhos anteriores mostram duas possibilidades de se medir com alta precisão as amplitudes e fases iniciais das componentes harmônicas de um segmento janelado com dois períodos, usando o método dos mínimos quadrados ordinários, supondo que o pitch foi anteriormente estimado com suficiente precisão. Na primeira abordagem, a otimização é realizada no domínio do tempo [255], e na segunda se utiliza o domínio da frequência [40].

Na implementação no domínio do tempo, para um dado segmento k de tamanho $N+1$ com L harmônicos, o erro a ser minimizado pode ser expresso como:

$$\varepsilon = \sum_{n=-N/2}^{N/2} (w[n]s[n] - w[n]s_h[n])^2, \qquad s_h[n] = \sum_{l=-L}^{L} c_l e^{jl\omega_0 n}, \tag{2.7}$$

onde s é o k-ésimo segmento, s_h é a parte harmônica deste segmento, c_l são valores de amplitudes complexas nos quais se verificam a condição $c_{-l} = c_l^*$ e w é a função de janelamento com valor máximo localizado no centro (tipicamente a janela de Hamming).

A quantidade de harmônicos L depende diretamente da frequência fundamental f_0. O valor de L está também associado a uma frequência máxima de corte para representação harmônica. Tal frequência de corte é um limiar hipotético, correspondendo à máxima frequência harmônica a partir da qual se acredita que as componentes harmônicas superiores estariam mascaradas pela presença

de ruído [51]. Este trabalho considera que a frequência de corte é de 5 kHz. Definida a frequência de corte F_c, o cálculo de L é imediato:

$$L = \left\lfloor \frac{F_c}{f_0} \right\rfloor$$

Reescrevendo a formulação em notação matricial, temos que

$$\mathbf{c} = [c_L^*, c_{L-1}^*, \ldots, c_0, \ldots, c_{L-1}, c_L]^t$$

e a matriz B é definida como

$$B = \begin{pmatrix} e^{j(-L)\omega_0(-N/2)} & e^{j(-L+1)\omega_0(-N/2)} & \ldots & e^{j(L)\omega_0(-N/2)} \\ e^{j(-L)\omega_0(-N/2+1)} & e^{j(-L+1)\omega_0(-N/2+1)} & \ldots & e^{j(L)\omega_0(-N/2)+1} \\ \vdots & \vdots & \ddots & \vdots \\ e^{j(-L)\omega_0(N/2)} & e^{j(-L+1)\omega_0(N/2)} & \ldots & e^{j(L)\omega_0(N/2)} \end{pmatrix} \tag{2.8}$$

e finalmente, a parte harmônica do sinal de voz é definida como $\mathbf{s_h} = \mathbf{c}B$. O objetivo da estimativa é encontrar o mínimo global do erro

$$\varepsilon = [W(\mathbf{s} - B\mathbf{c})]^t [W(\mathbf{s} - B\mathbf{c})],$$

onde D_w é o vetor (w) diagonalizado. Encontrar o ponto mínimo da função erro implica encontrar o mínimo local desta função parabólica, fazendo com que a derivada do erro $\frac{d\varepsilon}{dc} = 0$, ou seja,

$$\left[B^H D_w^H D_w B\right] \mathbf{c} = \left[B^H D_w^H D_w \mathbf{s}\right], \tag{2.9}$$

onde B^H é o hermitiano de B. Note que $D_w^H = D_w$.

Os coeficientes otimizados para $\mathbf{c}$ são obtidos tomando-se

$$\mathbf{c}^* = \left[B^H D_w D_w B\right]^{-1} \left[B^H D_w D_w \mathbf{s}\right]$$

Mediante o janelamento por w, é de se esperar que o erro de estimação seja mínimo na porção central da janela. Tal observação justifica o uso de janelamento com sobreposição. A Figura 2.15 exibe um exemplo que evidencia este fato. Note que o erro de estimação é crescente nas extremidades da janela. Tal erro será atenuado na fase de reconstrução pelo janelamento triangular.

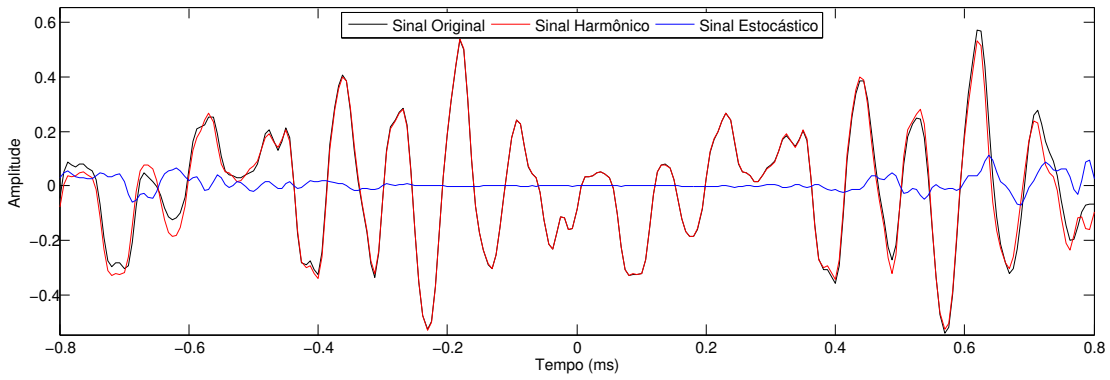

Figura 2.15: *Decomposição do segmento s em partes Harmônica (s_h) e Estocástica (s_s) no domínio temporal.*

A principal vantagem desta versão no domínio do tempo é que a matriz $\left[B^H D_w D_w B\right]^{-1}$ é Toeplitz, cuja solução pode ser encontrada eficientemente pelo clássico algoritmo de Levinson [37; 179]. Dados os coeficientes $\mathbf{c}$, as amplitudes e fases das componentes harmônicas são dadas por

$$A_l = \begin{cases} |c_l|, l = 0 \\ 2|c_l| = 2|c_{-l}|, \forall l \neq 0 \end{cases}, \quad \phi_l = \angle c_l = -\angle c_{-l}, \quad l = [-L, L]. \tag{2.10}$$

A segunda abordagem realizada no domínio da frequência considera que a Transformada de Fourier de Tempo Reduzido (STFT) S do k-ésimo segmento s pode ser aproximada de modo que

$$S(f) \approx 0.5 \sum_{l=1}^{L} A_l \left(e^{j\omega} W(f - lf_0) + e^{-j\omega} W(f + lf_0)\right),$$

onde $W(f - lf_0)$ é a transformada de Fourier da janela w centrada no l-ésimo harmônico da frequência fundamental f_0.

O espectro S pode ser expresso como $S = \mathbf{cH}$, onde $\mathbf{H}$ é uma matriz com $2L+1$ colunas, sendo que a l-ésima coluna é uma cópia do espectro H_l definido como

$$H_l = \begin{cases} \frac{1}{2}\left[W(f - lf_0) + W(f + lf_0)\right], 0 \leq l \leq L \\ \frac{j}{2}\left[W(f - lf_0) - W(f + lf_0)\right], L < l \leq 2L \end{cases}$$

O valor ótimo de

$$\mathbf{c} = [2A_0, A_1\cos(\phi_1), A_2\cos(\phi_2), \ldots, A_L\cos(\phi_L), A_1\sin(\phi_1), \ldots, A_L\sin(\phi_L)]^t$$

é dado por

$$\mathbf{c}^* = (\mathbf{H}^H\mathbf{H})^{-1}(\mathbf{H}^H S)$$

O processo de extração das amplitudes e fases é bastante simples. As amplitudes são tomadas como $A_l = \sqrt{c_l^2 + c_{L+l}^2}$ e as fases $\phi_l = \angle(c_l) = -\angle(c_{L+l})$. Matematicamente, ambos os resultados são equivalentes. No entanto, a inversão matricial realizada pelo algoritmo de Levinson na formulação do domínio temporal é mais eficiente, o que concede uma razoável vantagem a esta abordagem.

A Figura 2.16 exibe a decomposição do mesmo sinal mostrado anteriormente na Figura 2.15, porém sob um ponto de vista espectral. Note que o trem de pulsos harmônicos não excede à frequência de corte F_c. Em ambas as figuras, o erro de estimação é considerado a parte estocástica do segmento de voz.

Estimativa da Parte Estocástica: Se entende por parte estocástica do sinal toda componente cuja evolução temporal é essencialmente imprevisível, ainda que a mesma seja analisável em termos de probabilidades. Uma vez que a estimativa da componente harmônica é feita com métodos de otimização segundo o critério de mínimos quadrados, o erro de estimação entre o sinal original e o sinal reconstruído é consequentemente tratado como ruído colorido, ou seja, de evolução temporal aleatória e que pode ser espectralmente modelado. Em outras palavras, o ruído definido como

$$s_s[n] = w[n](s[n] - s_h[n])$$

corresponde à parte estocástica do sinal de voz s.

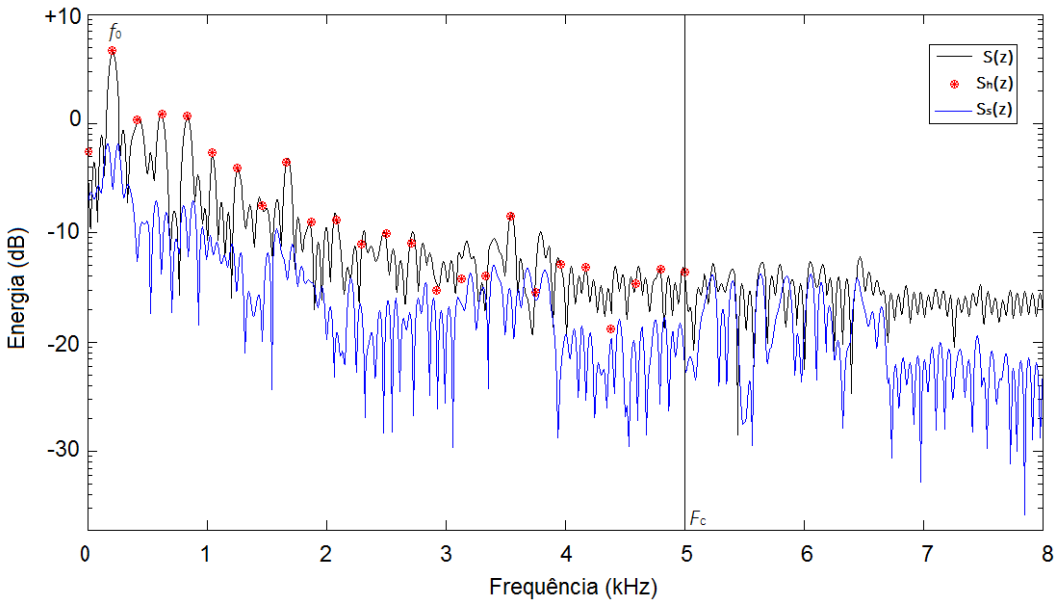

Figura 2.16: *Decomposição espectral do segmento S em trem de pulsos Harmônicos (S_h) e espectro Estocástico (S_s).*

A Figura 2.17 exibe a decomposição do sinal de voz (parte superior) nas duas componentes: a harmônica (parte central) e a estocástica (parte inferior). Evidentemente se observa uma forte correlação entre ambas as partes, devido aos erros de estimativa, principalmente em transições entre fonemas vozeados e não-vozeados ou silêncio.

Alguns trabalhos recentes descartam segmentos não-vozeados na fase do treinamento, alegando que estes não caracterizam significativamente a identidade sonora de um indivíduo [49; 305]. No entanto, alguns autores preferem tratar estes segmentos não-vozeados como se fossem a parte estocástica do sinal [32; 169].

Dado um sinal estocástico, o objetivo do sistema HSM passa a ser modelar tal sinal usando a decomposição fonte-filtro citada na seção anterior. O sinal de excitação do sistema é o ruído branco Gaussiano normalizado, e a componente do filtro é um filtro passa-faixas gerado pela versão suavizada do espectro de magnitude do sinal estocástico original (envelope estocástico). Existem várias formas de obtenção de tal envelope, sendo a mais comum entre elas baseada em coeficientes LPC. Tais coeficientes representam um filtro passa-faixas $H_s(z)$, que corresponde ao filtro gerador da parte estocástica do modelo de representação da voz. Dadas as novas propostas deste trabalho, as estimativas tanto dos envelopes das componentes estocásticas quanto dos envelopes harmônicos serão discutidas no Capítulo 3.

Módulo de Síntese do Sistema HSM

Como foi visto anteriormente, o módulo de análise toma um sinal discreto no tempo, segmenta o mesmo usando uma função de janelamento w com sobreposição de 50% em segmentos de aproximadamente 15 ms de comprimento, e devolve para o k-ésimo segmento uma frequência fundamental f_0^k, um vetor contendo L amplitudes harmônicas A_l^k e fases ϕ_l^k, além dos coeficientes de um filtro $H_s^k(z)$ que modela a parte estocástica. O objetivo deste módulo é ressintetizar o sinal usando tais informações.

Entretanto, o módulo de montagem necessita de ambas as partes harmônica e estocástica do

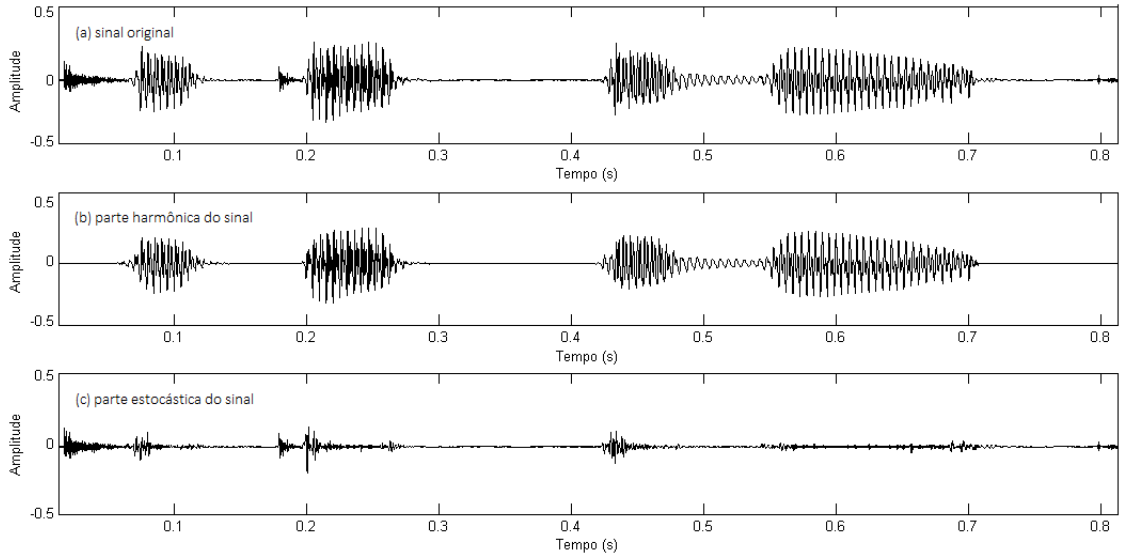

Figura 2.17: *As componentes Harmônica e Estocástica de um sinal de voz.*

segmento a ser reconstruído. Assim, segue-se o tópico de ressíntese da parte harmônica do sinal.

Ressíntese Harmônica: Na reconstrução da parte harmônica do sinal de voz, muitos autores se preocupam com a interpolação de fase e amplitude entre segmentos consecutivos [91; 152; 177]. Entretanto, com o advento do método *Overlap-Add*, a tarefa de ressíntese harmônica passou a ser trivial: basta ressintetizar independentemente cada segmento harmônico independentemente de seus adjacentes, e aplicar as sobreposições com soma do método OLA sobre estes segmentos, a fim de obter o sinal reconstruído. O tratamento das descontinuidades do sinal resultante é administrado pelo próprio módulo de montagem, o qual será apresentado num próximo tópico.

Dada uma taxa de amostragem R, bem como as fases iniciais ϕ^k de cada uma das L componentes harmônicas centradas no segmento k, a frequência fundamental $\omega_0^k = 2\pi f_0^k/R$ e as amplitudes harmônicas A^k do respectivo segmento harmônico, a reconstrução do sinal pode ser realizada sequencialmente para cada n, como

$$s_h^{(k)}[n] = \sum_{m=0}^{L} A_m^k \cos(\omega_m^k n + \phi_m^k), \tag{2.11}$$

ou pode ser reconstruída diretamente em blocos paralelos, a partir da multiplicação de matrizes

$$\mathbf{s_h}^{(k)} = \mathbf{A}^k \Omega^k,$$

onde Ω corresponde à matriz definida como

$$\Omega^k = \begin{pmatrix} 1 & \cos(\frac{-N}{2}\omega_0 + \phi_1^k) & \cos(\frac{-N}{2}2\omega_0 + \phi_2^k) & \cdots & \cos(\frac{-N}{2}L\omega_0 + \phi_L^k) \\ 1 & \cos(\frac{-N+1}{2}\omega_0 + \phi_1^k) & \cos(\frac{-N+1}{2}2\omega_0 + \phi_2^k) & \cdots & \cos(\frac{-N+1}{2}L\omega_0 + \phi_L^k) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & \cos(\frac{N-1}{2}\omega_0 + \phi_1^k) & \cos(\frac{N-1}{2}2\omega_0 + \phi_2^k) & \cdots & \cos(\frac{N-1}{2}L\omega_0 + \phi_L^k) \\ 1 & \cos(\frac{N}{2}\omega_0 + \phi_1^k) & \cos(\frac{N}{2}2\omega_0 + \phi_2^k) & \cdots & \cos(\frac{N}{2}L\omega_0 + \phi_L^k) \end{pmatrix}$$

e os coeficientes **A** definidos como

$$\mathbf{A} = [A_0^k, A_1^k, \ldots, A_L^k]^t.$$

Uma vez obtida a parte harmônica do sinal, é necessário restaurar a parte estocástica do mesmo.

Ressíntese Estocástica: A parte estocástica do sinal de voz, por se tratar de uma componente de difícil estimação, é regenerada usando o envelope espectral $H_s^k(z)$ de modo que tenha o mesmo tamanho do segmento harmônico $s_h^{(k)}$ (ou seja, $N+1$). Neste caso, o segmento estocástico é obtido sequencialmente por convolução

$$s_s^{(k)}[n] = \eta[n] * h_s^k[n]$$

, onde $\eta[n]$ é uma sequência composta por ruído branco Gaussiano e

$$h_s^k[n] = \mathcal{F}^{-1}\{H_s^k(z)\}$$

. Alternativamente, o segmento estocástico pode ser obtido (indiretamente) via filtragem espectral $\mathbf{s_s}^{(k)} = \mathcal{F}^{-1}\{H_s^k(z) \cdot \mathcal{F}\{\eta[n]\}\}$. Neste último caso, o cálculo é realizado em blocos paralelos.

Após ter em mãos a parte harmônica e estocástica do sinal, basta gerar o segmento $\hat{s}^{(k)}$ reconstruído da seguinte forma:

$$\hat{s}^{(k)}[n] = s_h^{(k)}[n] + s_s^{(k)}[n].$$

Finalmente, o módulo de síntese toma cada segmento reconstruído $\hat{s}^{(k)}$, e remonta o sinal de saída usando somas sobrepostas destes segmentos.

Montagem usando *Overlap Add*: A técnica de sobreposição com soma, mais conhecida como OLA (do Inglês, *Overlap Add*) é um eficiente modo de realizar a convolução discreta de um sinal por um filtro de resposta finita, normalmente representado por uma janela discreta: Triangular, Hann ou Hammimg [36]. O método serve para concatenar os segmentos reconstruídos pelo modelo HSM. Assim como no módulo de entrada existe o gerenciador de segmentação do sinal, na síntese existe o montador que executará as sobreposições e somas com taxa de sobreposição (*overlap*) de 50%, a fim de compor o sinal resultante.

Sob um ponto de visa prático, a reconstrução do sinal de saída pressupõe que a transição de um k-ésimo segmento $\hat{s}^{(k-1)}$ para o segmento $\hat{s}^{(k)}$, ambos de tamanho $N+1$ e centrados na origem, seja linear. Matematicamente, o sinal de saída reconstruído $\hat{s}$ é definido para todo k como

$$\hat{s}[n+kN] = \left(\frac{N-2n}{N}\right)\hat{s}^{(k-1)}[n] + \left(\frac{2n}{N}\right)\hat{s}^{(k)}[n+0.5N], \quad n = [0, 0.5N], \tag{2.12}$$

o que pode ser interpretado como OLA usando janela triangular. A Figura 2.18 ilustra este processo.

O janelamento triangular foi escolhido com base em resultados experimentais. Entretanto, outras janelas também poderiam ser usadas nesta tarefa, como por exemplo, a janela de Hann, que também tem a propriedade em OLA de soma constante igual a 1.

Uma vez definido o modelo que representa os sinais de voz, isto é, o modelo Harmônico-Estocástico, o trabalho apresentará algumas das principais técnicas de manipulação destes coeficientes, de modo a transformar tanto o conteúdo prosódico do sinal de voz (pitch, energia e,

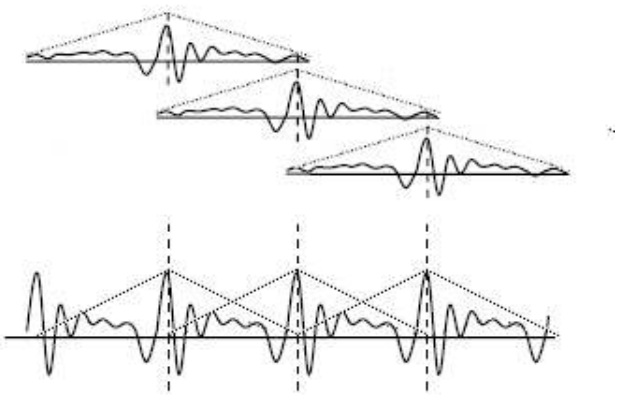

Figura 2.18: *Ilustração do método Overlap-Add.*

opcionalmente a duração), bem como transformar o conteúdo espectral do sinal de voz.

2.3 Tópicos em Conversão de Prosódia

O processo de conversão de sentenças de fala é um processo que deve levar em consideração não somente aspectos globais da sentença, como elementos prosódicos, mas também aspectos locais dentro de cada trecho estacionário do do sinal. A conversão dos parâmetros acústicos dos coeficientes de representação de cada um destes segmentos normalmente depende de dois processos básicos: o treinamento e a transformação.

Treinamento No contexto de conversão de voz inter-linguística, a fase de treinamento serve para estimar parâmetros globais e locais de controle dos sinais de voz.

Os parâmetros globais são responsáveis por controlar aspectos externos aos segmentos, tais como o pitch médio ou energia média das sentenças pronunciadas por um falante destino. Estes aspectos externos estão estreitamente relacionados com os controles de prosódia de um falante. Uma vez que a fonte de excitação está associada ao pulso glotal, é preferível utilizar estes parâmetros globais para adaptação da fonte de excitação a partir da prosódia do falante destino. De um modo geral, dizemos que os parâmetros globais da prosódia são controladores dos parâmetros acústicos da componente da fonte excitadora.

Por outro lado, os parâmetros locais estão associados às configurações do sinal de um trecho curto de voz (os segmentos). Dada a grande quantidade de fonemas associados a estes segmentos, é impossível se obter uma transformação global aplicada a um fonema qualquer.

Desta forma, para cada segmento de voz, vetores de parâmetros acústicos são obtidos dos modelos de representação do envelope espectral. Estes são organizados em um *espaço de características acústicas* e posteriormente agrupados em *classes fonéticas artificiais*. Cada uma destas classes fonéticas possui uma *chave de seleção*, que normalmente está associada ao centro de massa (centroide) da classe fonética. É a partir desta chave de seleção que um segmento do sinal de entrada localiza o segmento objetivo com o qual se realiza a conversão. O conjunto de dados deste espaço de características organizados em classe fonéticas de um falante determina o *corpus*[4] deste falante. Uma vez que o fonema de um sinal de voz está relacionado com a configuração do trato vocal, podemos

[4]Conjunto de informações acústicas capaz de caracterizar um determinado indivíduo.

associar os parâmetros locais aos parâmetros acústicos que modelam a componente do filtro.

Transformação Esta fase tem como objetivo transformar os parâmetros acústicos, tanto da fonte de excitação quanto da componente do filtro, a partir dos parâmetros globais (controles da prosódia) e locais (intra-segmento) obtidos na fase de treinamento.

A transformação da componente da fonte de excitação é muito particular do modelo de representação deste sinal. Por exemplo, se o pulso glotal esta sendo representado como frequências harmônicas de um modelo senoidal, é natural pensarmos em modificá-lo usando transformações lineares, ao passo que se o pulso glotal é representado como um sinal (no tempo), é preferível utilizar-se de técnicas como o TD-PSOLA ou FD-PSOLA [165] (vide Seção 2.3.1 a seguir). Vale lembrar que os parâmetros globais desta transformação são obtidos a partir da fase de treinamento.

A transformação da componente do filtro do sistema, também chamada de conversão espectral, exige um processo mais robusto do que a transformação da componente da fonte, e é feita em duas fases: a *fase de mapeamento* e a *fase de aproximação* destes parâmetros. Dado um vetor de parâmetros acústicos do falante origem que queremos converter, a fase de mapeamento seleciona um vetor de parâmetros do corpus do destino, que posteriormente será utilizado para conversão na fase de aproximação.

Detalharemos a seguir o primeiro tipo de transformação (componente da fonte), ao passo que a seção 2.4 tratará da transformação da componente do filtro.

2.3.1 Modelagem da Prosódia

A fonte de excitação sonora está intimamente ligada aos controles de prosódia, uma vez que estes controles parametrizam as transformações de pitch, energia e ritmo de articulação no processo de transformação do pulso glotal. De um modo geral, a prosódia é um conjunto de parâmetros acústicos responsáveis pelo volume do som irradiado (energia do sinal), frequência fundamental percebida (pitch) e o ritmo de articulação, que variam em larga escala no tempo. Em conversão de voz é muito importante que estes aspectos sejam levados em consideração, e desta forma, é imprescindível a existência de um conjunto de parâmetros que controlem tais aspectos.

No caso de conversão de voz inter-linguística, nem todos os aspectos são transpostos, visto que para determinados pares de línguas não existem correspondências prosódicas entre elas. Por exemplo, em Inglês, as perguntas possuem linhas melódicas diferentes das perguntas em Português, ou ainda num outro exemplo, o ritmo de articulação do Chinês é completamente diferente daquele do Inglês ou do Português. Deste modo, estamos interessado em converter de alguma forma o timbre dos falantes e os aspectos globais da prosódia, mantendo os aspectos locais da prosódia compatíveis com a língua original. Assim, apenas a conversão de momentos estatísticos associados à energia e altura musical serão consideradas neste trabalho.

Modelagem por Contornos

Chamamos de *contorno* a curva correspondente à variação dos parâmetros acústicos prosódicos ao longo do tempo. Sendo assim, temos um contorno para cada um dos dois aspectos de prosódia de uma sentença pronunciada. A Figura 2.19 mostra um exemplo de contorno de pitch e energia de um sinal de voz dado. Os valores de pitch e energia são tomados a partir de segmentos curtos do sinal de voz, recortados e janelados apropriadamente, de acordo com cada contexto. Por exemplo,

ao se estimar o pitch do trecho selecionado é apropriado utilizar uma janela de Hamming, ao passo que para se estimar a energia do sinal a janela retangular é mais apropriada. Consideraremos que as curvas de pitch e energia são obtidas por interpolação dos valores estimados a cada segmento, conforme ilustrado na Fig. 2.19.

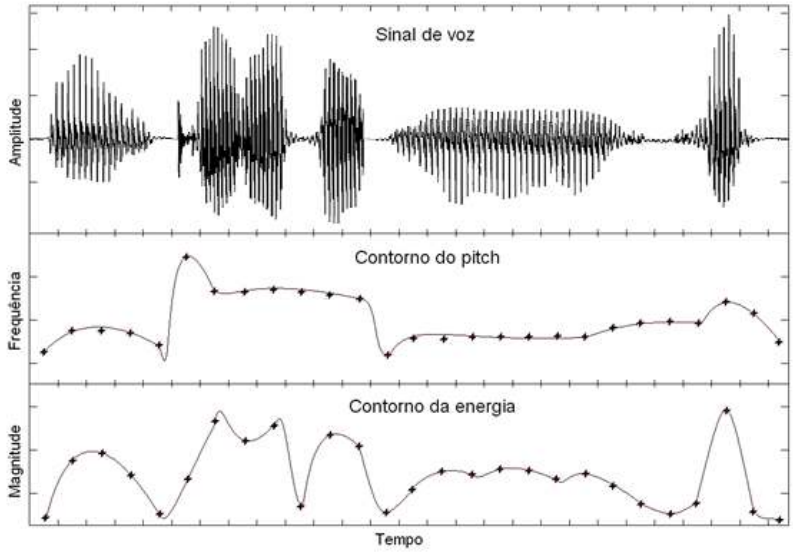

Figura 2.19: *Contorno de pitch e energia do sinal de voz.*

No contexto deste trabalho, modelar contornos não implica necessariamente em armazená-los, mas sim coletar momentos estatísticos de cada um deles. A conversão da prosódia neste caso significa apenas a conversão destes momentos estatísticos entre os falantes envolvidos. A partir destes momentos é possível se adaptar o sinal de voz modificado, a fim de torná-lo mais próximo estatisticamente do sinal do falante destino. Normalmente, somente a *média* e a *variância*, ou seja, os dois primeiros momentos, são tomados para modelar o contorno do *pitch* e da *energia* do sinal de voz. Dependendo da aplicação, pode-se optar por armazenar a média e a variância, somente a média, ou ainda partes dos contornos.

Modelo de Fujisaki

Fujisaki propôs no início dos anos 70 um modelo analítico que descreve variações da frequência fundamental F_0 (ou pitch) da fala [61; 62; 211]. Seu trabalho envolve os mecanismos de produção da fala que são responsáveis, em particular, pelo modelo de entonação das estruturas da prosódia. Em seu modelo, os contornos do pitch são compostos por duas diferentes componentes: a componente de *frase* (y_f) e a componente de *acento* (y_a). Cada componente possui vários eventos temporais de frase e acento. A componente de frase modela o contorno global do pitch, enquanto que a componente de acento modela as variações de pitch de curta escala. Tais componentes y_f e y_a são obtidas a partir da filtragem de um trem de impulsos de Dirac x_f pelo filtro de frase e de um trem de impulsos retangulares x_a pelo filtro de acento, respectivamente.

A resposta impulsiva h_f do mecanismo de controle de frase é definida como

$$h_f(t) = \alpha^2 t e^{-\alpha t} u(t),$$

onde $\alpha \in [2, 4]$ é sua frequência angular natural, assim como a resposta impulsiva h_a do mecanismo

de controle de acento é definida como

$$h_a(t) = [1 - (1 + \beta t)e^{-\beta t}]u(t),$$

onde $\beta \in [19, 21]$ é sua frequência angular natural.

O contorno do pitch é representado pela seguinte equação:

$$y(t) = \ln F_0(t) - \ln F_{\min} = y_f(t) + y_a(t)$$

onde $F_{\min}$ é o valor mínimo de F_0 do falante; N_f e N_a são os números de eventos de frase e acento, respectivamente; $A_{f,k}$ e $t_{p,k}$ são a magnitude e o tempo de início do k-ésimo evento de frase; $A_{a,k}$, $t'_{a,k}$ e $t''_{a,k}$ são a magnitude, o início e o fim do k-ésimo evento de acento, e $y_f(t)$ é definido como

$$y_f(t) = \sum_{k=1}^{N_f} A_{f,k} h_f(t - t_{f,k}),$$

assim como para $y_f(t)$ temos que

$$y_f(t) = \sum_{k=1}^{N_a} A_{a,k}[h_a(t - t'_{a,k}) - h_a(t - t''_{a,k})].$$

O método de Fujisaki é muito particular para cada língua. Sendo assim, é possível elaborar métodos para estimativa dos parâmetros deste modelo automaticamente para cada idioma [63; 78; 170].

É bem verdade que o conteúdo harmônico da componente da fonte (pulso glotal) não varia muito entre pessoas de um mesmo gênero. Tal premissa pode ser aceita mediante a observação de pessoas que imitam vozes de outras pessoas, em alguns casos com muita naturalidade e similaridade. No entanto, a distribuição de energia nas componentes da série harmônica em relação à frequência fundamental (o pitch) e a concentração de energia média no espectro de cada segmento são aspectos que devem ser levados em consideração.

2.3.2 Transformação da Prosódia

Dentre as técnicas existentes para conversão de prosódia, este trabalho destaca duas das mais populares. A primeira delas é a clássica transformação linear, que é aplicada sobre os momentos estatísticos que modelam os contornos de pitch e energia, e detalhada na seção abaixo. A segunda técnica abrange todos os métodos da família "*PSOLA*", a qual será abordada na sequência.

Transformação Linear

De um modo geral, a transformação linear de um conjunto de parâmetros acústicos $\mathbf{x} = [x_1, x_2, \ldots, x_N]$ com média μ_x e variância σ_x^2 para um conjunto objetivo $\mathbf{y} = [y_1, y_2, \ldots, y_M]$ com média μ_y e variância σ_y^2 é realizada a partir do ajuste linear

$$\hat{x}_n = \mu_y \frac{\sigma_y}{\sigma_x}(x_n - \mu_x), n = [1, N]. \tag{2.13}$$

A transformação da energia de um falante origem para um destino é realizada tomando-se o log do contorno de energia, por causa da percepção auditiva humana (ver Seção 2.1). Neste caso, a média e as variâncias também são extraídas deste conjunto de valores logarítmicos estimados durante a fase de treinamento. No caso da transformação do pitch de um sinal de voz [82] é comum ajustar o contorno de pitch usando as frequências fundamentais re-amostradas na escala MEL.

Tais transformações também são aplicáveis sobre os parâmetros de representações senoidais (SM) ou harmônico-estocásticas (HSM). No entanto, é mais comum representar o sinal de excitação em sua forma de onda natural, amplitude × tempo, devido à simplicidade de manipulação direta do sinal no domínio do tempo. Deste modo, é comum se utilizar técnicas de manipulação de duração e frequência do sinal, cuja segmentação é pitch-sincronizada . Tais técnicas, muito populares em processamento de sinais, são conhecidas como PSOLA, em suas versões para o domínio temporal (*Time Domain Pitch Synchronous Overlap-Add* (TD-PSOLA)) e espectral (*Frequency Domain Pitch Synchronous Overlap-Add* (FD-PSOLA) [165]).

PSOLA

Este é um método simples para modificações de pitch e velocidade de pronúncia em determinados trechos de voz, bastante utilizado em sistemas de tempo real. O algoritmo é comumente aplicado sobre o sinal de voz, mas pode ser aplicado normalmente sobre a componente da fonte. A ideia geral é obter segmentos sobrepostos sincronizados com o pitch e concatená-los apropriadamente, de acordo com a modificação desejada.

O processo geral pode ser descrito como:

1. Os instantes de início e fim de cada período de oscilação nas regiões vozeadas são determinados por um algoritmo marcador de pitch [74]. Métodos de detecção de pitch não são apropriados para esta tarefa, uma vez que são necessários os instantes exatos de início e fim do período de oscilação. Sons não-vozeados não são modificados por este método.

2. Segmentos de voz pitch-sincronizados são extraídos de modo a cobrir 2 a 5 períodos do sinal por segmento a partir de um janelamento (por exemplo, usando a janela triangular).

3. As modificações de pitch e escala são realizadas e posteriormente a saída é reconstruída usando síntese por sobreposição e adição por janelamento (*Overlap-Add* – OLA).

As modificações de escala estão diretamente relacionadas às modificações no ritmo de articulação de um sinal de voz. Sabe-se que dilatar um sinal de voz no tempo resulta em reduzir a frequência, ao passo que uma compressão temporal corresponde a um aumento de frequência. O objetivo deste método é obter uma saída escalada no eixo do tempo, sem afetar o timbre ou o pitch do sinal original. O algoritmo TD-PSOLA modifica o conteúdo temporal a partir da repetição ou remoção de um número inteiro de segmentos de voz. A repetição de segmentos produz um sinal dilatado no domínio do tempo, enquanto que a remoção produz um sinal comprimido no domínio do tempo (ver Figura 2.20). Note que a manipulação de um número inteiro de segmentos não modifica significativamente nem o conteúdo espectral nem o pitch do sinal de saída.

As modificações de pitch visam manipular o conteúdo espectral dos segmentos sem alterar suas características temporais. Além disso, tais modificações espectrais não devem degradar o envelope espectral do sinal, apenas modificar as localizações dos picos harmônicos relacionados com o pitch.

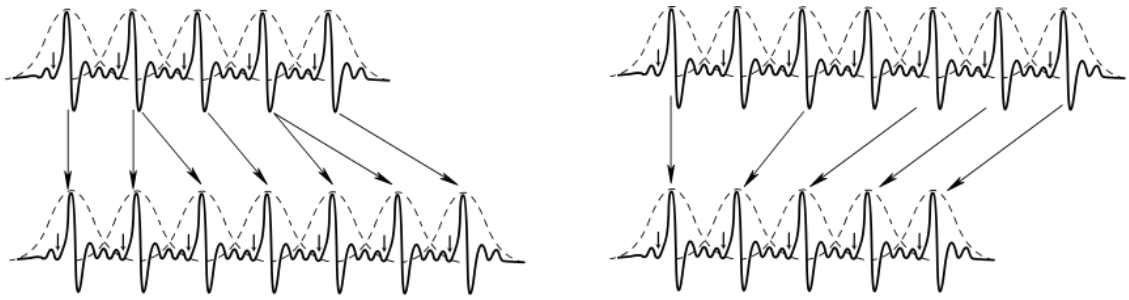

Figura 2.20: *Dilatação (esquerda) e compressão (direita) na escala temporal de um sinal de voz.*

Neste caso, o algoritmo TD-PSOLA modifica a porção de segmentos sobrepostos e sincronizados pelo pitch conforme mostrado na Figura 2.21. A elevação do pitch implica em uma aproximação dos segmentos consecutivos no tempo, enquanto que a diminuição do pitch implica no afastamento destes segmentos no tempo. Evidentemente, a modificação do pitch ocasiona uma modificação na escala temporal, que deve ser convenientemente adaptada mediante a repetição ou remoção de segmentos.

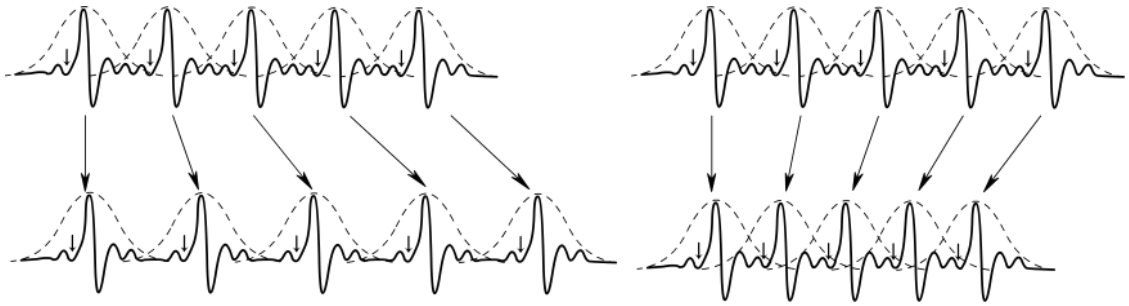

Figura 2.21: *Diminuição (esquerda) e elevação (direita) do pitch de um sinal de voz.*

Síntese por Overlap-Add: Reconstrói o sinal de saída a partir da soma de segmentos janelados (por janela triangular ou Hamming) sobrepostos, a fim de prevenir descontinuidades no sinal. Alguns cuidados devem ser tomados a fim de evitar problemas na síntese, tais como:

- Adaptação da escala no eixo do tempo do sinal de voz devido à modificação do pitch,
- As fases de segmentos consecutivos devem ser compatíveis no momento da concatenação,
- Na dilatação temporal, a repetição de segmentos pode introduzir efeitos não-humanizados aos sinais de voz. Este efeito pode ser compensado embaralhando-se segmentos consecutivos, procurando repetir o menor número de vezes possível cada segmento.

No caso do algoritmo FD-PSOLA, o mesmo procedimento do TD-PSOLA é aplicado sobre o espectro da fonte de excitação em cada segmento. Diferentemente do TD-PSOLA, este método só pode ser aplicado sobre o espectro da fonte, uma vez que não é desejável degradar o envelope espectral do sinal todo. A operação mais comum neste caso é a dilatação e compressão do espectro com prováveis repetições de picos harmônicos introduzidas nas altas frequências do espectro.

O cuidado básico que deve ser tomado é a que o sinal resultante seja real. Neste caso, deve-se garantir que o espectro seja simétrico em magnitude e anti-simétrico em fase. É comum tratar este problema realizando transformações em meio espectro, e ao final, refletir o resultado em relação à origem e conjugar a parte refletida.

2.4 Técnicas de Transformação Espectral

A transformação é realizada em quatro estágios: (1) clusterização dos dados em classes acústicas, (2) mapeamento de classes de um falante para o outro, (3) seleção da classe origem que mais se aproxima dos parâmetros respectivos no sinal de entrada e (4) conversão dos parâmetros acústicos da entrada, de modo que os mesmos pertençam à sua correspondente classe acústica no corpus destino após a conversão.

A Figura 2.22 mostra um diagrama que ilustra este processo. Primeiramente, o sistema recebe um conjunto de vetores acústicos X e Y que representam os dados de treinamento dos falantes origem e destino, respectivamente. Tais vetores, que representam os envelopes espectrais de cada segmento do sinal, são então organizados em classes acústicas de cada falante (origem O e destino D). Então, ambos os conjuntos de classes são submetidos a um estágio de mapeamento, no qual se alinham as classes de ambos os falantes de acordo com um critério de similaridade específico. Em seguida, o processo de transformação seleciona a classe artificial O_i do falante origem na qual s é pertinente, usando de estruturas de decisão, funções de custo mínimo ou chaves de seleção. Dada a classe O_i, a função de aproximação $\mathcal{T}_i$ associada é devolvida. Finalmente, a partir desta função de aproximação o sistema realiza a conversão do respectivo vetor acústico por meio de estruturas de dados de mapeamento de características espectrais, de aproximadores de funções universais, de utilização de bancos de parâmetros acústicos indexados, de funções de empenamento de frequências ou por meio de transformações lineares obtidas a partir dos momentos estatísticos de classes artificiais do falante origem e destino. Finalmente, após o processo de transformação o sistema devolve o vetor de parâmetros acústicos $\mathcal{T}_i(x) = y'$ que representa o envelope espectral do sinal convertido.

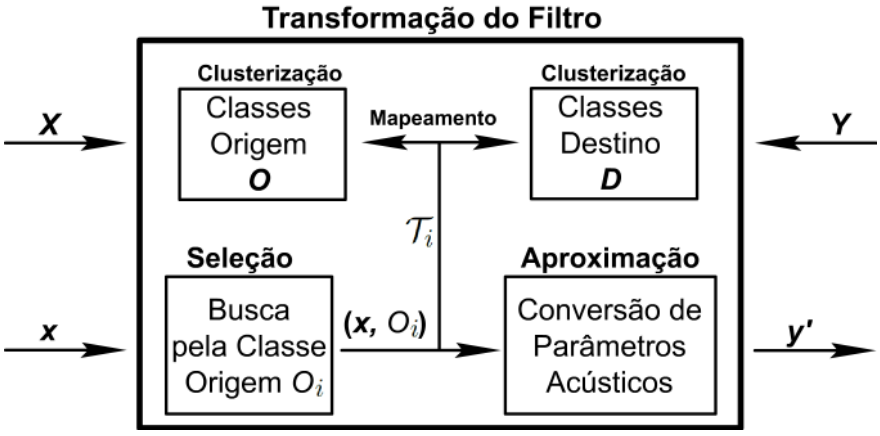

Figura 2.22: *As duas fases do processo de conversão de parâmetros acústicos.*

Na fase de treinamento, cada função de transferência $\mathcal{T}_i$ é determinada a partir de um alinhamento entre as classes acústicas do falante origem O_i e destino D_i. Então, para cada dupla (O_i, D_i), é possível estimar a transformação $\mathcal{T}_i$ que melhor aproxima $x \in O_i$ a um vetor $y' \in D_i$. Para realizar tal alinhamento, duas técnicas são atualmente bastante exploradas:

1. Alinhamento classe a classe: É intuitivo pensar no alinhamento de classes acústicas após a fase de clusterização. Entretanto, alguns cuidados devem ser tomados a fim de garantir a consistência do alinhamento: garantia de uma clusterização estável, ou seja, para um conjunto de amostras, o algoritmo de clusterização deve devolver o mesmo conjunto de classes; garantia de

equivalência de classes acústicas entre corpus de falantes distintos; e correspondência fonética entre as classes alinhadas. Dados os fatores listados, e mediante a dificuldade em solucioná-los, esta não é a mais explorada das duas técnicas.

2. Alinhamento segmento a segmento: A maioria dos trabalhos [49; 243] adota esta técnica por ser mais simples e menos sensível ao erro de alinhamento. Note que um erro de alinhamento de classes apresenta resultados mais catastróficos do que erros de alinhamento entre segmentos isolados. Após o alinhamento dos segmentos, as amostras são agrupadas em pares, ou seja, compõem um vetor artificial de dados acústicos $\mathbf{x} = [s_1, s_2, \ldots, s_L, t_1, t_2, \ldots, t_L]$, de tal forma que $\mathbf{s}$ e $\mathbf{t}$ são as amostras acústicas do falante origem e destino, respectivamente. O alinhamento é evidente, já que a clusterização em classes acústicas é feita de modo paralelo. Não obstante, tanto a complexidade computacional do alinhamento quanto o número de classes necessárias para a modelagem do corpus aumentam significativamente.

Para muitas destas tarefas existem técnicas bem conhecidas na literatura, tanto para os estágios de clusterização, alinhamento e seleção de classes quanto para a fase de aproximação de parâmetros acústicos. Deste modo, é conveniente categorizar as técnicas para conversão da componente do filtro em duas categorias: as técnicas de classificação, que envolvem tarefas relacionadas às classes, isto é, clusterização, mapeamento entre classes e seleção de classes acústicas dado um vetor de características; e as técnicas de aproximação, que são especializadas propriamente em converter vetores acústicos.

2.4.1 Técnicas de Classificação

Esta seção tem como meta fazer um levantamento detalhado das principais técnicas de classificação, dentre as citadas anteriormente na introdução deste trabalho (Seção 1.3.2). Um bom método classificador deve ser capaz de tomar um conjunto de vetores acústicos e sistematicamente organizá-los em **classes acústicas** de modo que os dados estejam aglutinados em torno de um elemento central. Além disso, se espera que estes elementos centrais estejam tão afastados quanto seja possível.

Clusterização por Distâncias

Sabe-se que, em treinamento independente de texto, as sentenças, em geral, são fragmentadas e aglutinadas em classes acústicas num espaço de características. As técnicas de agrupamento (**clustering**) [312] exploram semelhanças entre padrões e agrupam os padrões parecidos em categorias ou grupos. Segundo Jain [100], a clusterização é um método que utiliza o aprendizado não-supervisionado ou auto-organizável que busca extrair informação relevante de dados não-rotulados. Tais informações, de um modo geral, são obtidas a partir de medidas de similaridade entre dois clusters assim como um critério global como a soma do erro quadrático. Os algoritmos de agrupamento são classificados como hierárquicos ou sequenciais (iterativos). Dentre os algoritmos de clusterização estão o K-médias [176] e K-histogramas [282]. Em ambos os casos, organiza-se os dados a partir de ajustes estatísticos.

Existem alguns padrões de seleção de clusters tais como a seleção a partir de um módulo de seleção (*Unit Selection*) que leva em conta outros aspectos entre segmentos. No Unit Selection [48; 244], o objetivo é realizar a transformação de modo a encontrar o pareamento que minimize a distância

das características do falante origem e destino, e ao mesmo tempo preservar a máxima continuidade. Em geral, a unidade de seleção pondera duas funções de custo: o custo objetivo $C_t(u_m, t_m)$ é uma estimativa da diferença entre a unidade abstrata u_m e um destino t_m, e o custo de concatenação $C_c(u_{m-1}, u_m)$ é uma estimativa da qualidade na junção entre duas unidades consecutivas u_{m-1} e u_m. Com o tempo, o Unit Selection se restringiu aos segmentos na conversão de voz, passando a ser uma técnica denominada Frame Selection [282]. Pensando em segmentos, é razoável balancear o custo entre concatenar segmentos consecutivos e selecionar segmentos mais similares aos do falante destino. O principal objetivo proposto por esta técnica é minimizar a descontinuidade entre segmentos adjacentes, e favorecer a seleção de segmentos consecutivos, evitando não somente efeitos de descontinuidade entre segmentos consecutivos, mas também a suavização excessiva [283].

Modelos Ocultos de Markov

Uma abordagem estatística muito comum para o mapeamento de segmentos é a utilização de cadeias Markovianas. Além dos modelos ocultos de Markov (HMM) [201] serem muito utilizados em segmentação de fonemas [8], estes modelos também são bastante utilizados para estimação da transformação $\mathcal{T}$, e têm sido aplicados em muitos trabalhos de conversão de voz e modificação de prosódia [118; 305]. HMMs são amplamente utilizadas em alinhamento de sequências em treinamento independente de texto, por agruparem classes de segmentos em mapas de estados cujas transições são estatisticamente ajustadas.

Um modelo oculto de Markov é um modelo estatístico em que o sistema a ser modelado é considerado um processo Markoviano com estados inobserváveis (com parâmetros desconhecidos), e o desafio é determinar os parâmetros ocultos a partir dos parâmetros observáveis. Uma sequência de valores s_t de uma variável aleatória discreta S_t caracteriza uma cadeia de Markov se:

$$P(S_{t+1} = s_{t+1}|S_t = s_t) = P(S_{t+1} = s_{t+1}|S_t = s_t, S_{t-1} = s_{t-1}, \ldots, S_1 = s_1)$$

A grande popularidade das HMMs se encontra em problemas de reconhecimento de voz [65; 129; 205]. A utilização de HMMs em reconhecimento de voz parte do pressuposto que um sinal de voz segmentado pode ser visto como processos estacionários de tempo curto, sendo representados por um modelo Markoviano. Outra razão pela qual o uso de cadeia de Markov é popular é porque elas podem ser treinadas automaticamente e são simples e computacionalmente eficientes na utilização. A HMM tende a ter em cada estado uma distribuição estatística que é uma mistura Gaussiana de covariância diagonal, dada pela verossimilhança de cada vetor observado (ver técnicas de aproximação da seção seguinte). Para cada fonema, espera-se ter uma distribuição diferente na saída; uma HMM para uma sequência de fonemas é criada a partir da concatenação de HMMs treinadas para cada fonema.

Em conversão de voz, uma HMM é constituída por estados destinados a agrupamentos (ou classes acústicas), e Kim [118] propõe modificar a estrutura das HMMs para realização de conversão espectral. O modelo possui dois conjuntos de estados-dependentes que incluem relações de mapeamento espectral entre os falantes. Um deles representa o espaço particionado do falante origem e o outro contém parâmetros sintéticos do falante destino. Existem trabalhos que combinam HMM com diversas outras técnicas, como por exemplo distribuições de probabilidade multi-espaciais (MSD) [271; 272; 306].

As técnicas cognitivas também são alternativas apropriadas em tarefas de clusterização e mape-

amento de classes acústicas, dada sua natureza de ajuste não-linear. Dentre as estruturas de dados cognitivas mais utilizadas, se destacam as redes neurais e as redes topológicas. No entanto, outros tipos de estruturas de decisão podem ser utilizadas tais como SVM [66; 67] ou redes fuzzy [122; 162].

Redes Neurais Artificiais - Estrutura de Decisão

As redes neurais artificiais (ANN) são muito utilizadas em reconhecimento de padrões [17] e particularmente em sistemas de reconhecimento de fala [133], e vêm sendo muito utilizadas em sistemas de conversão de voz [42; 79]. Os modelos ANN consistem em uma estrutura de dados com nós interconectados, onde cada nó representa o modelo de um neurônio artificial, e as interconexões possuem pesos sinápticos associados. Um neurônio é ativado a partir de uma função denominada função de ativação. Existem vários tipos destas funções, sendo as mais conhecidas a *sigmoide* e as funções de base radial (Radial Basis Function Neural Networks - RBFNN [236]).

Existem duas utilizações clássicas para o uso de redes neurais: a ANN como estrutura de decisão, isto é, a rede responde '*sim*' ou '*não*' para um estimulo dado; ou a ANN como estrutura aproximadora de funções universais (usada na fase de aproximação).

No primeiro caso, cada rede neural $\mathcal{R}_i(s)$ recebe um conjunto de valores de entrada (parâmetros acústicos) e está associada a uma determinada classe fonética $\mathcal{C}_i$ do corpus do destino. Cada classe fonética armazena um conjunto de vetores de parâmetros acústicos do destino. Supondo já realizada a fase de treinamento, onde se atualizam e ajustam os pesos das redes, o sistema recebe um vetor s de parâmetros do falante origem, e deve encontrar a correspondente classe fonética $\mathcal{C}_i$ na qual s melhor se encaixa. Neste caso, devemos encontrar o índice k tal que

$$k = arg\min_i(\mathcal{R}_i(s)).$$

Ou seja, as redes neurais funcionam como chaves seletoras de classes acústicas.

Redes Topológicas

Assim como em uma rede neural, um mapeamento acústico pode ser realizado a partir de informações provenientes de redes de dados previamente estabelecidas na fase de treinamento. Tais redes definem operações denominadas mapeamentos topológicos. Dentre os tipos de mapeamentos topológicos, destacam-se as redes de mapeamento topológico de características (Topological Feature Mapping - TFM), as redes de mapeamento topográfico generativos (Generative Topographic Mapping - GTM) e as árvores de decisão.

Topological Feature Mapping: Neste caso, a seleção é feita pela escolha do vencedor dentro de um mapa topológico de características. O mapa de características realiza uma quantização vetorial (VQ) que subdivide o espaço de características em um número fixo de subespaços representados cada um por um neurônio. O mapa de características é auto-organizável durante a fase de treinamento, utilizando vetores com parâmetros acústicos de ambos os falantes (falante origem e destino). Esta auto-organização cria um tipo de memória associativa do mapa de características com a qual é possível se selecionar o mapeamento vencedor [209].

Generative Topographic Mapping: Conhecido como GTM, é um método de aprendizado computacional auto-organizável, cujos dados de entrada são considerados como pontos de um espaço de baixa dimensionalidade a serem mapeados para um novo ponto observado num espaço de latência de alta dimensionalidade, constituído por uma função de mapeamento suave adicionada de ruído [18; 182]. Os parâmetros da distribuição de probabilidade de baixa dimensão, o mapeamento suave e o ruído são todos aprendidos na fase de treinamento usando o algoritmo de máxima expectativa (EM). A ideia central de GTM é que variáveis de alta dimensionalidade dos dados observados são geradas a partir de um número pequeno de variáveis latentes. É comum pensar que GTMs são como uma versão não-linear da análise de componentes principais (PCA). Este modelo permite o relacionamento não-linear entre os dados latentes e o espaço dos dados, a partir de um mapeamento paramétrico do espaço de latência para o correspondente ponto no espaço dos dados. As características acústicas do sinal de voz representado pela GTM são posteriormente convertidas em chaves seletoras, com as quais se realiza um mapeamento entre os vetores de entrada (falante origem) e saída (falante destino).

Árvores de decisão: Uma árvore de decisão é um bom exemplo de estrutura topológica de decisão, cujo critério para construção é baseado em conceitos de entropia. Uma árvore de decisão é uma estrutura de dados em que cada nó interno avalia um atributo (vetor de parâmetros acústicos, por exemplo), cada aresta corresponde a um valor de atributo e cada folha representa uma classe. Uma árvore de decisão muito interessante, e adaptada à conversão de voz, é a estrutura de dados CART (*Classification And Regression Tree*) [47; 48], que permite tanto trabalhar com dados numéricos (tais como características espectrais) como com dados categorizados (tais como características fonéticas) no momento em que é construído um modelo acústico. A ideia geral é que os espaços acústicos de ambos os falantes são organizados em classes acústicas, e uma função de conversão (uma GMM, por exemplo) pode ser estimada para cada classe. A partir do vetor de parâmetros acústicos de entrada (do falante origem), pode-se percorrer a árvore de decisão a fim de selecionar a classe acústica mais apropriada do destino, isto é, aquela que melhor se ajusta na fase de treinamento.

Self Organizing Maps (SOM): Propostos por Teuvo Kohonen [121], um mapa auto-organizável (também chamado de mapa de Kohonen) é um tipo de rede neural artificial, com treinamento não-supervisionado que produz uma representação discreta de baixa dimensionalidade (comumente bidimensional) do espaço de amostras de treinamento de entrada. As SOMs tipicamente apresentam duas fases: O treinamento constrói o mapa usando exemplos de entrada de vetores quantizados, e na fase de treinamento, o algoritmo classifica automaticamente os novos vetores da entrada. Cada componente (neurônio) do mapa se associa a um nó de peso e dimensão relacionados com o dado de entrada. Assim, a SOM representa um mapeamento de um espaço de entrada de alta dimensionalidade em uma rede de baixa dimensionalidade. O critério de conexão entre o espaço de entrada e o mapa de Kohonen é definido a partir de uma função de distância, que agrupa nós mais próximos entre si, preservando as propriedades topológicas do espaço de entrada [120].

Dynamic Time Warping

Além das estruturas topológicas, bem como outras estruturas cognitivas, podemos realizar o mapeamento de classes acústicas a partir de técnicas clássicas de processamento de sinais. Estas técnicas compreendem todo tipo de transformação do sinal, baseadas em representações no domínio do tempo ou frequência, que por sua vez tentam codificar o sinal usando frequentemente uma biblioteca de chaves de segmentos do sinal ou *codewords*. O que fundamentalmente diferencia esta das demais técnicas é o fato de que as ferramentas de transformação foram desenvolvidas com o propósito específico de manipular sinais de voz a partir de ferramentas peculiares em sistemas digitais. Em geral, além destas técnicas serem de fácil intuição, isto é, permitirem uma compreensão clara e sucinta da operação a ser realizada, elas são facilmente implementáveis.

O Dynamic Time Warping [15] é um algoritmo bastante conhecido em processamento de sinais, usado para medição de similaridade entre duas sequências variantes no tempo. A partir destas medidas, podemos inferir distâncias entre pares de vetores de características acústicas. Dado um vetor de características v_1, o DTW busca, dentre um conjunto de vetores, o vetor v'_1 que minimiza a diferença $|v_1 - v'_1|$. O algoritmo faz o uso de programação dinâmica [64], bastante comum em sistemas de reconhecimento de voz.

Vector Quantization/Codebook

Shikano [231], o primeiro a propor o problema de conversão de voz, utilizou esta técnica, que consiste em quantizar os vetores de parâmetros acústicos dos sinais de entrada, organizando esses dados em espaços de características de ambos os falantes, e finalmente associando estes vetores entre si a partir de um critério de minimização de erro.

Uma quantização de vetor (*Vector Quantization*) é uma técnica que permite a modelagem de funções de densidade de probabilidade a partir da distribuição de vetores protótipos, originalmente usada para compressão de dados. O método divide um conjunto grande de pontos (vetores) em pequenos grupos com aproximadamente o mesmo número de pontos próximos entre si, onde cada um destes grupos é representado pelo seu *centroide*, analogamente aos algoritmos de clusterização. Os critérios de agrupamento dependem muito da aplicação. Por exemplo, Abe [1] propõe uma estratégia que agrupa os vetores a partir de um histograma montado segundo distâncias ponderadas entre o falante origem e o destino. Como pode-se observar, a VQ organiza o espaço de características de forma que os clusters são acessíveis por uma chave de seleção (*codeword*) correspondente ao centroide do agrupamento. O conjunto destas *codewords* é conhecido como *codebook*.

A ideia central dos codebooks é definir um mapeamento entre classes acústicas representadas pelos codewords correspondentes. Dado o mapeamento, é possível realizar qualquer tipo de transformação que converta um parâmetro acústico do falante origem para o destino. Lopez [135], por exemplo, utiliza redes neurais para mapear estes parâmetros acústicos.

A partir da técnica VQ/codebook, Arslan [6] propõe uma técnica bastante utilizada em conversão de voz, conhecida como *Speaker Transformation Algorithm using Segmental Codebooks (STASC)*. Esta técnica é uma fusão de VQ/codebook com outras técnicas de mapeamento espectral, do pulso glotal e de parâmetros da prosódia. A ideia geral é gerar um codebook para cada falante a partir de parâmetros acústicos que representam o trato vocal (neste caso, LSF).

O algoritmo STASC propõe dois métodos de associação entre os codewords dos falantes. A

primeira abordagem é dependente de texto, o que torna a tarefa de associação entre os centroides dos clusters de classes fonéticas correspondentes trivial, uma vez que a informação de associação entre segmentos é dada. A segunda abordagem é independente de texto e supõe que existe uma quantidade suficiente de fonemas comuns a ambos os falantes, a ponto de permitir uma associação biunívoca entre as classes fonéticas. Neste caso, é utilizada uma HMM para alinhar estes fonemas, e assim, definir a associação entre as classes fonéticas. Após tais associações, o algoritmo propõe uma transformação linear dos parâmetros acústicos, assim como um mapeamento ponderado do espectro do pulso glotal, ambos do falante origem para o destino. O algoritmo STASC ainda realiza um ajuste dos parâmetros da prosódia do sinal de saída, levando em consideração a energia, a média e o desvio padrão do pitch, e o ritmo de articulação deste sinal em relação aos parâmetros estimados do destino.

2.4.2 Técnicas de Aproximação

Técnicas de aproximação são destinadas a realizar a conversão de um vetor de características acústicas do falante origem para um vetor mapeado na fase anterior.

Modelo de Misturas Gaussianas - GMM

A partir de duas sequências de parâmetros espectrais $s = (s_1, s_2, \ldots, s_N)$ e $t = (t_1, t_2, \ldots, t_N)$, correspondentes aos coeficientes que representam o trato vocal do falante origem e do destino, respectivamente, o objetivo desta técnica é encontrar uma função de conversão F que minimiza o erro médio quadrático

$$E_m = E\{|t - F(s)|^2\},$$

onde $E\{\cdot\}$ corresponde à esperança estatística.

O Modelo de Misturas Gaussianas (GMM) é um modelo paramétrico clássico utilizado em muitas técnicas de reconhecimento de padrões [208; 296], sendo de longe o mais utilizado em processamento de voz [178; 273]. A utilização de GMM foi proposta a fim de estimar parâmetros para utilizar uma função de conversão linear obtida a partir de um arcabouço probabilístico [107; 257; 261].

Um GMM permite que a distribuição de probabilidade de s seja escrita como uma soma de Q funções Gaussianas multivariadas [5; 103]

$$P(s) = \sum_{i=1}^{Q} \alpha_i N(s; \mu_i, \Sigma_i), \tag{2.14}$$

onde $\sum_{i=1}^{Q} \alpha_i = 1,\ \alpha_i \geq 0$ e $N(s; \mu, \Sigma)$ denota a distribuição normal d-dimensional com vetor de média μ e matriz de covariância Σ, definida como

$$N(s; \mu, \Sigma) = \frac{1}{(2\pi)^{d/2}} \Sigma^{-0.5} \exp\left[-\frac{1}{2}(s-\mu)^T \Sigma^{-1}(s-\mu)\right]. \tag{2.15}$$

Os parâmetros deste modelo podem ser estimados a partir do algoritmo de máxima expectativa [154].

Como se pode observar, o modelo de misturas Gaussianas $(\alpha_i, \mu_i, \Sigma_i, i \in [1..Q])$ se ajusta ao vetor de parâmetros espectrais do falante origem s. O mapeamento entre o vetor do falante origem s

e cada classe fonética $\mathcal{C}_i$ mais adequada pode ser encontrado em termos probabilísticos pelo cálculo da probabilidade condicional dado por:

$$P(\mathcal{C}_i|s) = \frac{\alpha_i N(s;\mu_i,\Sigma_i)}{\sum_{j=1}^{Q}\alpha_j N(s;\mu_j,\Sigma_j)}.$$

Segundo Kang et al. [111], a função de conversão F que transforma cada vetor do falante origem s em seu correspondente vetor t' respectivo ao falante destino é definida como

$$F(s) = \sum_{i=1}^{Q} P(\mathcal{C}_i|s)[\nu_i^{<2>} + \Sigma_i^{<1,2>}(\Sigma_i^{<1,1>})^{-1}(s - \mu_i^{<1>})],$$

onde o vetor de média μ_i e a matriz de covariância Σ_i são definidos como

$$\Sigma_i = \begin{bmatrix} \Sigma_i^{<1,1>} & \Sigma_i^{<1,2>} \\ \Sigma_i^{<2,1>} & \Sigma_i^{<2,2>} \end{bmatrix} \quad e \quad \mu_i = \begin{bmatrix} \mu_i^{<1>} \\ \mu_i^{<2>} \end{bmatrix}.$$

Espera-se que $E\{|t - F(s)|^2\}$ seja o menor possível.

Devido ao grande número de técnicas estatísticas disponíveis na literatura, mencionamos apenas duas técnicas que foram utilizadas em trabalhos anteriores de conversão de voz:

Maximum Likelihood Estimators (MLE): Estimação com máxima verossimilhança [301] é um método usado para ajustar um modelo estatístico a um conjunto de dados, provendo estimativas para os parâmetros do modelo.

Principal Component Analysis (PCA): Análise de componentes principais [294] compreende um procedimento matemático que transforma um número de variáveis provavelmente correlacionadas em um número menor de variáveis descorrelacionadas conhecidas como componentes principais. A primeira componente principal representa o máximo da variabilidade dos dados, e cada componente sucessora representa o máximo da variabilidade dos dados restantes.

Mesbahi [155] publicou um artigo em que se comparam métodos baseados em GMM para transformações de vozes. Ele compara funções lineares para conversão de voz, sugerindo soluções relativas aos defeitos causados por suavização excessiva do espectro e *overfitting*.

Redes Neurais Artificiais - Funções Aproximadoras Universais

Uma rede neural recebe um vetor de parâmetros de entrada s pertencente ao corpus do falante origem e o aproxima dos vetores de parâmetros de t pertencentes ao corpus do falante destino, de modo que o resultado desta aproximação é um sinal convertido s' pertencente à classe fonética mais apropriada do destino. No contexto de conversão de formantes, Narendranath [169] e Desai [42] utilizam redes neurais como modelo universal de aproximação de funções [161; 265]. Neste contexto, existe uma rede neural $\mathcal{R}_{(S,T)}(s)$ para cada par de classes acústicas $\mathcal{C}_S$ e $\mathcal{C}_T$ dos falantes origem e destino, respectivamente, especializada em aproximar vetores $s \in \mathcal{C}_S$ em vetores $t \in \mathcal{C}_T$. Deste modo, tais redes $\mathcal{R}_{(S,T)}(s)$ são treinadas e posteriormente selecionadas na fase de transformação, de acordo com o vetor de entrada s. Note que podemos utilizar as duas aplicações para realizar uma conversão de voz. No entanto, é comum alguns autores utilizarem técnicas estatísticas para o mapeamento de classes e ANN para transformar os parâmetros acústicos.

Transformações Lineares

Uma outra alternativa quanto às técnicas de mapeamento são as técnicas de álgebra linear, que se baseiam em interpretações geométricas dos dados, e são muito utilizadas no contexto de aproximação de parâmetros acústicos, particularmente na conversão de parâmetros espectrais a partir de transformações lineares.

Uma transformação linear é um tipo particular de função entre dois espaços vetoriais que preserva as operações de adição vetorial e multiplicação por escalar. No contexto deste trabalho [203], os espaços vetoriais são definidos a partir da dimensão dos vetores de parâmetros acústicos utilizados no sistema. As transformações lineares mais comuns na literatura são aplicadas a partir de *Regressões Lineares*.

As regressões lineares [76] são modelos que relacionam uma variável escalar y a uma ou mais variáveis X, de modo que o modelo depende linearmente de um conjunto desconhecido de parâmetros a serem estimados a partir dos dados de treinamento. De um modo geral, dizemos que uma regressão linear encontra uma função linear que melhor se ajusta a um conjunto de valores distribuídos num espaço vetorial, conforme pode ser observado na Figura 2.23. Tal função linear pode ser vista como um mapeamento de parâmetros acústicos distribuídos num espaço Euclideano de características.

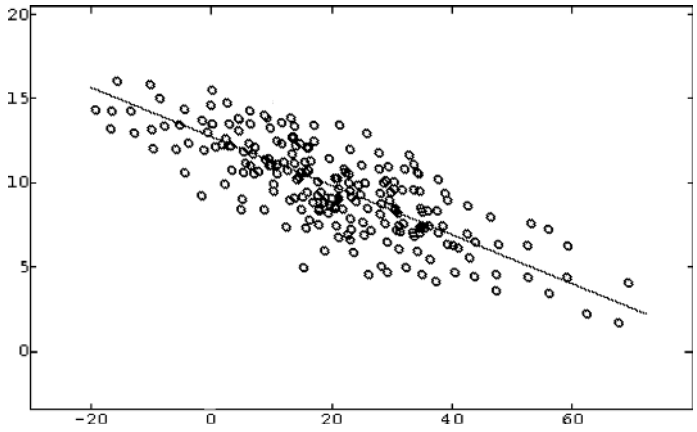

Figura 2.23: *Ajuste linear a partir de um conjunto de pontos no espaço Euclideano.*

Dado um valor de saída y_i e um conjunto de dados de entrada $x_i = [x_{i1}, x_{i2}, \ldots, x_{ip}]$, para $i \in [1..n]$, um modelo de regressão linear pressupõe que a relação entre a variável y_i e o vetor de regressores p-dimensional x_i é aproximadamente linear. Tal aproximação é modelada a partir da adição de ruído ao modelo. Assim, o modelo é definido como:

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \epsilon_i = x_i'\beta + \epsilon_i, \quad i = 1, 2, \ldots, n,$$

onde $x_i'\beta$ é o produto interno entre os vetores x_i e β.

É comum reescrever estas equações em forma matricial, como

$$y = X\beta + \epsilon,$$

onde

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \; X = \begin{bmatrix} x'_1 \\ x'_2 \\ \vdots \\ x'_n \end{bmatrix} = \begin{bmatrix} x_{11} & \dots & x_{1p} \\ x_{21} & \dots & x_{2p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \dots & x_{np} \end{bmatrix}, \; \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{bmatrix}, \; \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}.$$

Existem vários métodos de estimação dos coeficientes de regressão linear β. Tais métodos são dirigidos por critérios de minimização de erro, levando em consideração estratégias como a máxima verossimilhança (*Maximum Likelihood Linear Regression – MLLR* [254; 314]) ou a minimização do erro médio quadrático. Neste último caso, temos que

$$\beta = (X^t X)^{-1}(X^t y).$$

Alguns autores preferem utilizar a Regressão Linear Multivariada [104] (*Linear Multivariate Regression – LMR*). Como qualquer modelo de regressão linear clássico, a transformação é ajustada a partir de um conjunto de treinamento de parâmetros entre as classes acústicas do falante origem e destino. No caso de modelos multivariados, a variável de resposta é multivariada, isto é, da forma $T = (t_1, t_2, \dots, t_p)$, sendo que as variáveis $t_1, t_2, \dots, t_p$ são medidas na mesma unidade amostral, e estão associadas a p conjuntos de variáveis preditoras não-aleatórias. Para cada variável de saída t_k existe um conjunto de variáveis de entrada $s_k = [s_{k1}, s_{k2}, \dots, s_{kN_k}]$, para $k \in [1 \dots p]$.

Desta forma, cada uma das p regressões lineares é definida como

$$t_k = s_i \alpha_i + \epsilon_i, \quad i = 1, 2, \dots, p$$

onde $\alpha_i = [\alpha_{1i}, \alpha_{2i}, \dots, \alpha_{N_i i}]'$ são os coeficientes da regressão e ϵ_i é o ruído associado à i-ésima predição. É comum se utilizar um estimador de mínimos quadrados para calcular os coeficientes α; Valbret [284] propõe este modelo para realizar o mapeamento entre as classes acústicas do falante origem para o destino, minimizando o erro médio quadrático entre as duas amostras a serem convertidas.

Interpolações Ponderadas

Em conversão de voz, uma transformação linear é útil quando queremos encontrar uma regra que mapeia um vetor de parâmetros acústicos s para um correspondente t. De um modo geral, tal transformação é definida como:

$$s = f(t) = \alpha t + \epsilon.$$

Uma interpolação linear ponderada (*Weighted Linear Interpolation*) é um método de transformação espectral que surge como uma alternativa em conversão de voz [189].

Suponha que tenhamos M transformações lineares de classes acústicas do falante origem para o destino. Um modelo de transformação mais robusto é obtido se todas as M transformações contribuem para a conversão de cada vetor de entrada. O peso de cada matriz de transformação depende, então, de uma medida da probabilidade com que o vetor pertence à correspondente classe fonética. Assim, a função de conversão que se aplica sobre o vetor do falante origem x é definido

pela seguinte interpolação:

$$F(x) = \left(\sum_{m=1}^{M} \lambda_m(x) W_m \right) \bar{x}, \tag{2.16}$$

onde $\bar{x} = [x', 1]'$ é o vetor estendido de x e λ_m é a interpolação ponderada da matriz W_m, cujo valor é dado pela probabilidade do vetor x pertencer à m-ésima classe fonética, ou seja, $\lambda_m(x) = P(\mathcal{C}_m|x)$. A matriz W_m é obtida segundo um critério de minimização de erro, tal como o clássico critério de mínimos quadrados.

Ye & Young [298] propõem uma técnica para conversão de voz inspirada em modelos de interpolação ponderada e uso de escalas psicoacústicas conhecida como PWLT (*Perceptually Weighted Linear Transformation*). No contexto desta aplicação, parâmetros acústicos formados por coeficientes mel-cepstrais normalizados são interpolados de acordo com uma matriz de pesos W.

Modelos Bilineares

Muitas vezes, os modelos lineares utilizados para conversão de parâmetros acústicos são insuficientes para uma transformação de alta qualidade, uma vez que estas apresentam um certo grau de não-linearidade. Uma boa alternativa para transformação destes parâmetros é a utilização de modelos bilineares [39; 59; 193; 266], que apresentam a vantagem de serem mais simples que os não-lineares e, em geral, mais expressivos que o modelo linear.

Em geral, uma transformação bilinear [25] de s para t é definida como:

$$t = \rho \frac{1 - s}{1 + s}, \quad \rho > 0.$$

A ideia básica consiste em realizar primeiramente a interpolação numa direção, e partir desta, realizar na outra direção. O resultado da interpolação bilinear não depende da ordem de interpolação, ou seja, interpolar sobre o eixo x após o eixo y é equivalente a interpolar sobre o eixo y após o eixo x. Embora cada passo seja linear sobre os valores e posições amostradas, a interpolação como um todo não é linear, mas sim quadrática.

Os modelos bilineares são também muito utilizados em transformações e adaptações de escalas psicoacústicas [237]. Neste contexto, a transformação bilinear é um mapeamento definido como

$$\hat{z}^{-1} = \frac{z^{-1} - \rho}{1 - z^{-1}\rho},$$

cuja inversa é

$$z^{-1} = \frac{\hat{z}^{-1} + \rho}{1 + \hat{z}^{-1}\rho},$$

que leva o círculo unitário no plano z para o círculo unitário do plano ρ de modo que, para $0 < \rho < 1$, as baixas frequências são espaçadas e as altas frequências são comprimidas, analogamente à transformação da frequência em Hertz para a escala Bark.

Decomposição de Valor Singular

A decomposição de valor singular (SVD) [145] é um método que realiza uma transformação que torna um conjunto de variáveis correlacionadas em um conjunto descorrelacionado que melhor representa os relacionamentos entre os dados do sinal original. O método identifica e ordena as

dimensões espaciais de modo que os pontos (neste caso os dados) estejam mais concentrados nas dimensões com maior variação. A partir das dimensões identificadas como de maior variação, podemos encontrar a melhor aproximação dos dados originais usando poucas dimensões. Assim, SVD pode ser visto como um método para redução de dados.

A ideia básica por trás de SVD é tomar um conjunto extenso de dados de alta dimensionalidade e reduzi-los a um espaço de menor dimensionalidade que expõe a subestrutura dos dados originais explicitamente e os ordena decrescentemente de acordo com sua variação no espaço. A partir de um limiar conveniente sobre o conjunto de dados resultante é possível recompor as componentes mais significativas do sinal. Tal recomposição pode ser vista como um sinal passado por um filtro de supressão de ruído.

A partir destas componentes mais representativas do sinal de voz podem ser realizadas transformações espectrais quaisquer que aproximem o sinal do falante origem para o destino. Por exemplo, Popa [193] adota uma abordagem que utiliza SVD para obtenção dos sinais reduzidos, juntamente com a transformação bilinear para a conversão dos parâmetros.

Deformação em Frequência (DFW)

Uma vez que os parâmetros acústicos representam perfeitamente o envelope espectral do segmento, a manipulação direta na escala das frequências é uma forma intuitiva de aproximar o espectro dos sinais selecionados do falante origem e do destino. A técnica de deformação em frequências, ou *Frequency Warping* como é tradicionalmente conhecida, tem sido bastante utilizada em normalização de falantes [130; 131; 309] e empenamentos para escalas psicoacústicas [157]. Um empenamento de frequências pode ser realizado a partir de transformações bilineares.

O empenamento de frequências corresponde a uma função de mapeamento que modifica a escala no eixo das frequências, levando frequências ω de um espectro para frequências $\tilde{\omega}$ do correspondente espectro empenado, conforme mostra a Figura 2.24. Tal função de empenamento é comumente utilizada para aproximar parâmetros acústicos do falante origem para o destino, representados no domínio das frequências, assim como LSF [283] ou MFCC [187].

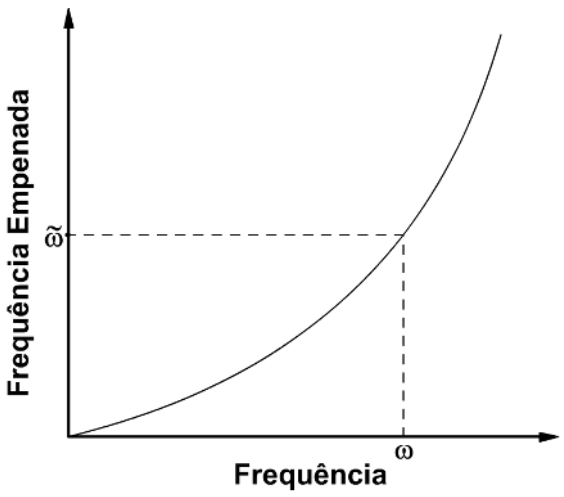

Figura 2.24: *As duas fases do processo de conversão de parâmetros acústicos.*

Nos casos em que as funções de empenamento são estimadas por otimização na fase de treinamento, dizemos que tais funções são de empenamento de frequência dinâmico [270] (*Dynamic*

Frequency Warping – DFW). Tal otimização utiliza uma estratégia de seleção de segmentos com máxima similaridade, assim como a DTW, e por esta razão, estas são ditas duais [171].

Daniel Erro [50] propôs recentemente um método de empenamento de frequências ponderado (*Weighted Frequency Warping – WFW*) que combina as abordagens DFW e GMM. Para cada segmento é calculada uma combinação linear de funções base. Os pesos destas combinações e formatos das funções base são obtidos de um treinamento GMM, que também é utilizado para aumentar a similaridade entre os sinais do falante origem e do destino. Tais pesos e funções base definem uma função de empenamento entre um par de espectros de segmentos associados às duas classes acústicas a serem aproximadas.

O modelo adotado para representar estes segmentos é o modelo harmônico-estocástico, o qual será utilizado como a base do sistema proposto por este trabalho, e detalhado no capítulo que segue.

Capítulo 3

Conversor de Voz Inter-Linguístico

O sucesso no desenvolvimento de um sistema de conversão inter-linguístico com alto índice de qualidade depende intrinsecamente da escolha de boas ferramentas especializadas em tarefas específicas dentro do processo. Embora o sistema exija uma computação de alto desempenho, fortemente estruturada em algoritmos fundados sobre bases complexas e sólidas da disciplina de Processamento de Sinal Digital, o trabalho visa manter o foco em conceitos estabelecidos pela Ciência da Computação, tais como análise de complexidade algorítmica, paradigmas de programação imperativa estruturada e orientada a objetos, entre outras ferramentas de apoio, como a teoria dos grafos.

Esta seção está categorizada em módulos funcionais, a fim de facilitar a exposição das novas propostas técnicas da tese. Antes de introduzir cada módulo funcional, o sistema é apresentado na Seção 3.1 sobre uma ótica algorítmica, relacionando cada módulo funcional do programa a um método proposto na seção seguinte. Se trata de uma visão sistemática do sistema de conversão de voz inter-linguístico, bem como sua estrutura básica:

- A Seção 3.2 mostra uma adaptação do modelo HSM para Análise e Síntese de sinais de voz usando envelopes interpolados das partes harmônica de amplitude e estocástica.

- Na Seção 3.3 é apresentada uma técnica de decomposição paramétrica destes envelopes, a qual é posteriormente submetido ao módulo de Clusterização. Esta seção aborda também a reconfiguração dos parâmetros para a fase de síntese, mediante a parametrização inversa.

- A Seção 3.4 apresenta uma proposta de clusterização na qual se utilizam conceitos de morfologia matemática para o agrupamento de vetores de características em classes acústicas. Esta clusterização monta os corpora contendo as classes acústicas de cada um dos falantes origem e destino.

- Para cada par de corpora distintos, estes corpora serão alinhados usando um algoritmo clássico na Teoria dos Grafos para emparelhamento de custo mínimo na Seção 3.5.

- Finalmente, a Seção 3.6 apresentará uma nova técnica de transformação de parâmetros acústicos, inspirada em técnicas de Transformação Linear combinadas à deformação em frequência (*Normalized Frequency Warping* – NFW).

3.1 Visão Estrutural do Sistema

Um sistema robusto precisa definir de maneira clara e precisa as relações entre os módulos do sistema. Antes de passar à parte de implementação, segue-se a definição de cada módulo funcional do sistema.

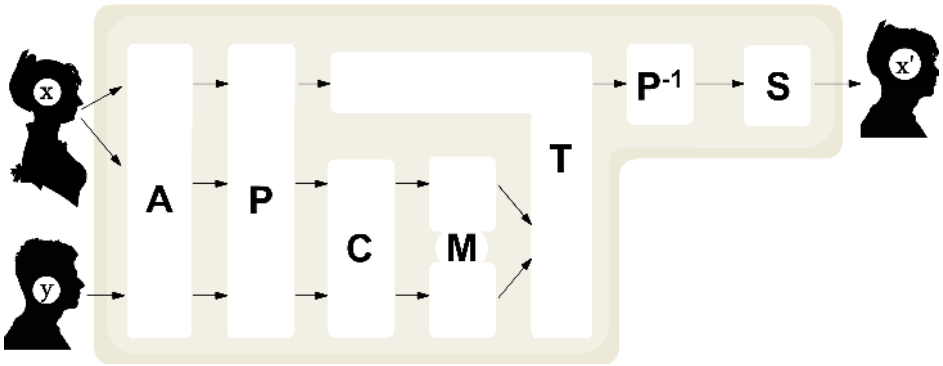

Figura 3.1: *Visão do sistema sob o ponto de vista de fases de Treinamento e Conversão.*

A Figura 3.1 exibe um diagrama geral, no qual estão identificados os fluxos de dados de ambos os falantes. Denotam-se por **x** as sentenças relativas ao falante origem, do qual se espera modificar a voz, e por **y** as sentenças do falante destino, para o qual se deseja converter a voz, obtendo as sentenças convertidas $\mathbf{x}'$. Os processos estão discriminados como letras maiúsculas, de acordo com suas funcionalidades, as quais serão descritas em ordem:

A– O módulo de **A**nálise toma uma sentença de fala pronunciada, segmenta-a em pequenos trechos fonéticos e fornece uma representação intuitiva dos mesmos para posterior quantização.

P– O módulo de **P**arametrização é responsável pela modelagem e quantização dos parâmetros acústicos devolvidos pelo módulo anterior.

C– O módulo de **C**lusterização é a parte do sistema responsável por organizar o *corpus* de cada falante em classes acústicas, o que implica em modelar o espaço de características acústicas respectivas ao falante em questão.

M– O módulo de **M**apeamento corresponde à parte do sistema que associa as classes acústicas entre ambos os falantes, usando técnicas de alinhamento de custo mínimo.

T– O módulo de **T**ransformação diz respeito às funções de conversão de parâmetros acústicos, estimadas a partir do alinhamento das classes do módulo anterior.

S– O módulo de **S**íntese, por sua vez, monta o sinal de saída $\mathbf{x}'$ tomando os parâmetros transformados. Estes parâmetros finais são dados pela parametrização inversa $\mathbf{P}^{-1}$, na qual os parâmetros quantizados são readaptados ao padrão do modelo de análise.

Estes **módulos funcionais** que conjuntamente compõem o sistema de conversão de voz estão mais profundamente explanados na seção subsequente. Por questões de implementação, nesta seção os módulos estão categorizados em duas fases distintas do processo, conforme é observado na Figura 3.2.

A fase de treinamento toma dois conjuntos de sentenças pronunciadas por cada um dos falantes (origem X e destino Y), modela (A) tais sentenças e quantiza (P) todo o conjunto de informação.

Estes vetores quantizados são clusterizados (C) em classes acústicas e organizadas em corpora, e os pares de classes acústicas de ambos os falantes são alinhados (M). Como resultado final de treinamento é gerada uma função de transformação $@\mathcal{T}$, a qual é usada na fase de conversão.

Na fase de conversão, a função $@\mathcal{T}$ toma um vetor quantizado de características acústicas Ψ_x, pertencente ao falante origem, e o converte em um vetor $\hat{\Psi_x}$, de modo que este passe a possuir propriedades acústicas que o caracterizem como pertencente ao corpus do falante destino.

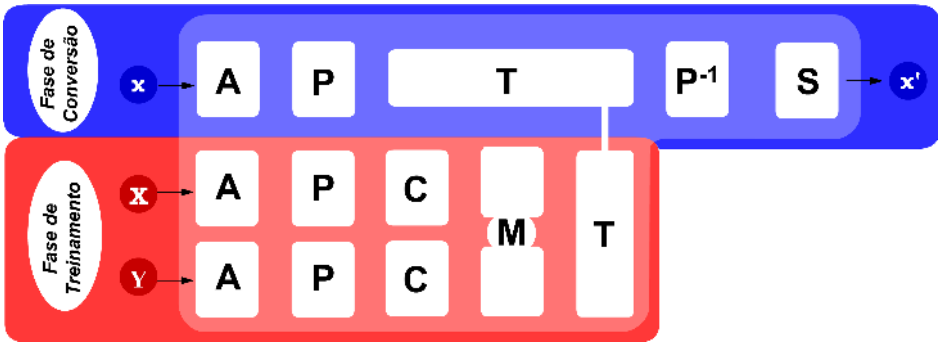

Figura 3.2: *Visão do sistema sob o ponto de vista de fases de treinamento e transformação.*

A fim de apresentar o modelo mais claramente, é conveniente apresentar separadamente ambas as fases da conversão: o treinamento (Seção 3.1.1) e a conversão propriamente dita (Seção 3.1.2).

3.1.1 Fase de Treinamento

Como dito anteriormente, o principal objetivo da fase de treinamento é definir uma função de conversão $@\mathcal{T}$ que será utilizada na fase subsequente. A explicação do processo será feito com base nos algoritmos construídos para cada módulo da conversão.

O Algoritmo `TREINAMENTO` (Alg. 3.1) corresponde ao processo geral de treinamento, que toma como entrada os conjuntos de sentenças de treinamento dos falantes origem (X) e destino (Y), e devolve a função de mapeamento de parâmetros acústicos $@\mathcal{T}$. Todos os algoritmos são descritos em pseudocódigo baseado em linguagens de alto nível para computação científica, tais como *Matlab/Octave*. Embora seja possível compor todas as funções em cascata, como no caso do padrão *Command* [1] em Programação Orientada a Objetos [26], tais métodos são compostos sequencialmente para tornar a apresentação mais didática. O padrão adotado na composição dos pseudocódigos que seguem foi:

1. Os nomes dos módulos que correspondem às fases do processo de conversão, isto é, às fases de treinamento e de conversão, sempre são escritos usando CAIXA ALTA;

2. Os nomes dos módulos funcionais descritos anteriormente (Análise, Parametrização, Clusterização, etc) sempre são iniciados por uma letra maiúscula e continuados com letras minúsculas;

3. Nomes de funções utilitárias usadas dentro dos grandes módulos são sempre escritas com todas as letras minúsculas;

4. As variáveis globais são sempre maiúsculas e itálicas.

[1] *Command* é um dos 23 padrões clássicos de projeto de software em Programação Orientada a Objetos.

Algoritmo 3.1 $@\mathcal{T} \leftarrow$ `TREINAMENTO(`X`,`Y`)`

▷ *Modelagem do conjunto treinamento de ambos os falantes*
1 $[F_0^X, S^X, A^X, \varphi^X] \leftarrow$ `Análise(`X`)`;
2 $[F_0^Y, S^Y, A^Y, \varphi^Y] \leftarrow$ `Análise(`Y`)`;

▷ *Quantização dos valores de entrada*
3 $\Psi^X \leftarrow$ `Parametrização(`F_0^X, S^X, A^X`)`
4 $\Psi^Y \leftarrow$ `Parametrização(`F_0^Y, S^Y, A^Y`)`

▷ *Agrupamento dos vetores acústicos em classes acústicas*
5 $[C^X, G^X] \leftarrow$ `Clusterização(`Ψ^X`)`
6 $[C^Y, G^Y] \leftarrow$ `Clusterização(`Ψ^Y`)`

▷ *Alinhamento de classes e geração da função de mapeamento M*
7 $@M \leftarrow$ `Mapeamento(`C^X, C^Y`)`

▷ *Estimação da função de transformação $\mathcal{T}$*
8 $@\mathcal{T} \leftarrow$ `Transformação(`$@M, C^X, C^Y, G^X, G^Y$`)`
9 **return**$(@\mathcal{T})$

O algoritmo acima possui cinco módulos funcionais, cada um deles descrito detalhadamente em cada seção posteriormente. No entanto, a fim de esclarecer alguns aspectos de implementação da fase de treinamento, segue uma descrição detalhada de cada linha de código:

1–2: O módulo de análise, baseado no modelo Harmônico/Estocástico, é aplicado sobre o conjunto de treinamento de ambos os falantes, identificados pelos índices superiores. Este módulo toma cada um dos sinais compostos pelas concatenações de todas as sentenças pronunciadas por um mesmo falante, e devolve o conjunto de vetores F_0 (frequências fundamentais de cada segmento), o conjunto S de segmentos estocásticos do modelo, bem como o conjunto A de vetores representando o envelope harmônico de cada segmento. As fases iniciais não são modeladas, ou seja, são descartadas na fase de treinamento. Na fase de conversão, no entanto, estas são transferidas diretamente do módulo de Análise para o módulo de Síntese, uma vez que a manipulação das mesmas pode prejudicar a continuidade temporal (vide Seção 3.2). Caso um segmento k seja avaliado como não-vozeado, então $F_0[k] = 0$ e as respectivas entradas de A são nulas.

3–4: O módulo de parametrização toma as estruturas de dados devolvidas pelo módulo de análise e as converte em um vetor de características acústicas multidimensional. Tal processo, conhecido como quantização vetorial, foi bem explorado em outros sistemas de conversão de voz [135; 231]. No entanto, o processo de quantização apresentado utiliza um módulo de decomposição espectral, sendo esta uma contribuição original deste trabalho e apresentada na Seção 3.3.

5–6: O módulo de clusterização, apresentado na Seção 3.4, também é uma proposta original do trabalho, e utiliza ferramentas de morfologia matemática para obter uma clusterização estável dos vetores Ψ já quantizados pelo módulo anterior. Tal clusterização devolve um conjunto de classes acústicas C, o qual juntamente com as informações globais das sentenças de treinamento P compõem o corpus do falante respectivo (X ou Y).

7: Uma vez definidos os corpora de X e de Y, o módulo de Mapeamento é responsável por alinhar as classes acústicas usando funções de minimização de distância global do mapeamento. Esta

distância global é definida de acordo com uma métrica de diferença acumulada entre centroides de classes emparelhadas. O objetivo deste módulo, detalhado na Seção 3.5, é encontrar o emparelhamento de custo mínimo, que associa cada classe do corpus origem a uma classe do corpus destino.

8: Finalmente, o módulo de transformação coleta as informações de ambos os corpora, além da função de mapeamento $@M$, que associa classes acústicas C^X à C^Y. A partir das médias e variâncias destas classes, o módulo gera um conjunto de funções de transformação $@\mathcal{T}_i$ associadas a cada dupla de classes $[C_i^X, C_j^Y]$, de tal forma que $M(C_i^X) \rightarrow C_j^Y$. Ademais, a função de transformação é também responsável pela conversão da prosódia e outros parâmetros acústicos do sinal de entrada ao nível da sentença (ver Seção 3.6).

Encerrada a fase de treinamento, o sistema de conversão de voz está apto a realizar as transformações prosódicas e espectrais do sinal de origem.

3.1.2 Fase de Conversão

Resumidamente, a fase de conversão codifica o sinal de entrada, quantiza-o em vetores acústicos de modo análogo à fase de treinamento, e em seguida realiza a transformação destes vetores para posterior reconstrução. A transformação destes vetores utiliza a função de conversão $@\mathcal{T}$ construída na fase de treinamento.

O Algoritmo CONVERSÃO (Alg. 3.2) é o próprio conversor de voz inter-linguístico, o qual recebe uma sentença de voz x pronunciada por um falante origem, juntamente com a função $@\mathcal{T}$, e devolve a mesma sentença com o timbre de voz do falante destino.

Algoritmo 3.2 $\hat{x} \leftarrow$ `CONVERSÃO(`x`,`$@\mathcal{T}$`)`

▷ *Modelagem do sinal de entrada*
1 $[F_0, S, A, \varphi] \leftarrow$ `Análise`(x);

▷ *Quantização dos valores de entrada*
2 $\Psi \leftarrow$ `Parametrização`(F_0, S, A)

▷ *Conversão de Ψ usando a função $\mathcal{T}$*
3 $\hat{\Psi} \leftarrow \mathcal{T}(\Psi)$

▷ *Aplicação da parametrização inversa de $\hat{\Psi}$*
4 $[\hat{F}_0, \hat{S}, \hat{A}] \leftarrow$ `Parametrização_Inversa`$(\hat{\Psi})$;

▷ *Síntese do sinal de saída*
5 $\hat{x} \leftarrow$ `Síntese`$(\hat{F}_0, \hat{S}, \hat{A}, \varphi)$
6 **return**$(\hat{x})$

A estrutura do algoritmo de conversão é bastante similar à do algoritmo de treinamento. De fato, observe que a Linha 1 executa exatamente a mesma instrução das linhas 1 e 2 do Algoritmo 3.1, assim como a Linha 2 corresponde às linhas 3 e 4 daquele. No entanto, as particularidades deste algoritmo estão nas linhas seguintes. A Linha 3 converte os vetores de origem quantizados usando a função de transformação $@\mathcal{T}$ obtida na fase de treinamento. Esta função realiza não somente a conversão espectral das componentes harmônicas e estocásticas do modelo, como também adéqua os controles globais (como da prosódia por exemplo) de modo a conformá-los ao falante destino.

Maiores detalhes podem ser encontrados na Seção 3.6. Na Linha 4, o módulo de parametrização inversa é utilizado a fim de recompor os parâmetros do modelo Harmônico-Estocástico, de tal modo que estes sejam utilizados no módulo de Síntese do sinal de voz (Linha 5).

Antes de dar prosseguimento à implementação de cada um destes módulos funcionais, um conjunto de valores iniciais devem ser definidos de modo a serem utilizados globalmente no sistema.

Configurações Iniciais do Sistema

As configurações iniciais do sistema assumem valores iniciais associados à codificação e segmentação do sinal de voz. Evidentemente, o conjunto de sinais de entrada devem aderir estritamente a estes parâmetros.

O sinal de voz tem a particularidade de concentrar a maior parte da energia sonora em baixas frequências. Sabe-se que, a partir de uma taxa de amostragem de 16 kHz, as diferenças na representação de sentenças de fala são praticamente indistinguíveis ao ouvido humano [281] e, por esta razão, o sistema apresentado adota a taxa de amostragem $R = 16$ kHz.

Outro aspecto importante na configuração inicial é a escolha do tamanho das janelas na segmentação. Como discutido na Seção 2.2.2, os segmentos de voz se mantêm estáveis em períodos entre 15 ms e 25 ms [80; 166; 281]. Sendo assim, o sistema usa uma segmentação em janelas de 16 ms, visando preservar todas as nuances do trecho longo de voz. Com uma taxa de amostragem de 16 kHz, o número de amostras de cada segmento de voz é $N_w = 256$. Além disso, foi adotada uma abordagem sobreposta na segmentação do sinal, com taxa de 50% de sobreposição entre segmentos consecutivos, a fim de melhorar a estimação harmônica na porção central do segmento. Tal fator implica que o segmentador deve realizar um "salto" de $N_{\text{adv}} = 128$ amostras entre um segmento e outro.

A escolha do tipo de janela na fase de análise tem grande importância na medida em que impacta no espalhamento espectral da respectiva resposta em frequência. Naturalmente, a voz possui uma componente ruidosa, cuja energia é distribuída ao longo do espectro, concentrada principalmente nas altas frequências. Uma vez que a segmentação é o produto da janela com o segmento de voz, o espectro do trecho segmentado é convoluído com o espectro da janela, e por isso o uso de janelas com lóbulos principais no espectro demasiadamente largos são descartados (ver Figura 3.3). A janela de *Hamming* foi escolhida como janela padrão para segmentação. No entanto, qualquer outra janela poderia ser utilizada para este proposito.

Um outro parâmetro global requerido pelo modelo é a Frequência de corte F_c, a partir da qual as componentes harmônicas superiores são consideradas componentes ruidosas. Este trabalho adota o padrão $F_c = 5$ kHz proposto por Erro [51], visando comparar os resultados com tal sistema. Não obstante, valores maiores podem ser utilizados, acarretando em uma melhora perceptual da estimação harmônica.

Sabe-se que o módulo de Parametrização toma um conjunto de vetores de tamanho variável devolvidos pelo modulo de Análise, e sendo assim, uma quantização é requerida a fim de normalizar os dados. Considere que L^* é o fator de normalização usado nesta fase do sistema, o qual corresponde à ordem de grandeza de representação dos vetores acústicos. O valor L^* usado pelo trabalho foi $L^* = 48$, associado ao número de bandas perceptuais distribuídas no espectro pelo algoritmo de quantização (ver Seção 3.3).

O valor de ϵ é um valor extremamente pequeno (*default* $\epsilon = 10^{-10}$), que é usado em módu-

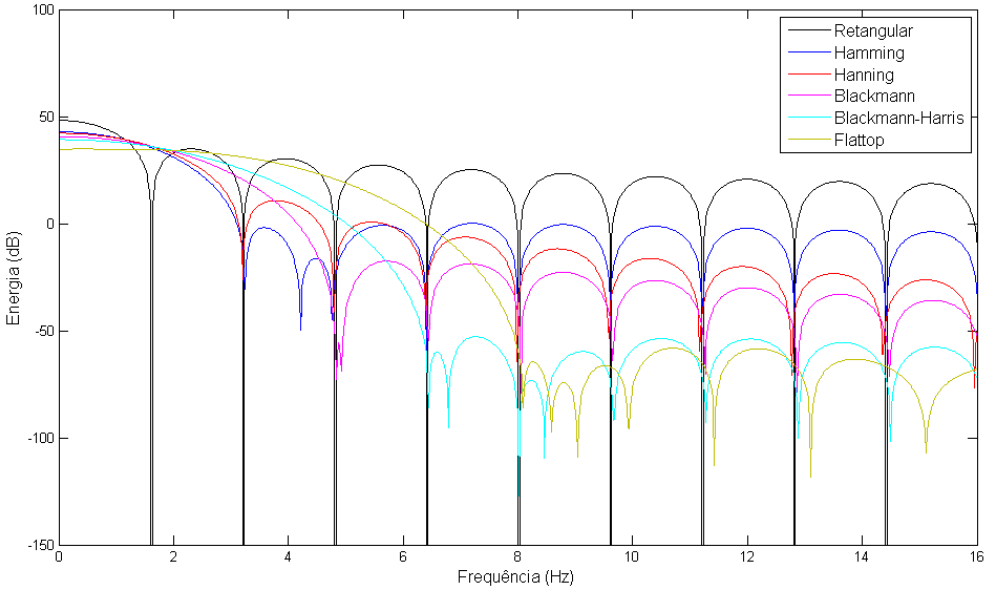

Figura 3.3: *Espalhamento espectral de algumas janelas de segmentação.*

los cujas transformações necessitam de um valor mínimo diferente de zero, a fim de manter uma representação numérica estável, como é o caso do log, por exemplo.

Quanto ao módulo de Clusterização, é necessário que seja definida a taxa de amostragem do mapa fonético. Neste caso, assume-se que $M_f = M_{f1} = M_{f2} = 35$. Já a largura de cada filtro triangular usado para a classificação fonética é estabelecida como sendo $B_f = 120$ mels.

O módulo de Transformação por sua vez utiliza uma variável que define a quantidade de classes $M_c = 15$ a serem utilizadas na mistura para compor a transformação linear ponderada.

Segue a tabela 3.1 que lista todas as variáveis globais do sistema, com seus respectivos valores default.

Identificador	Valor	Descrição
R	16 kHz	Taxa de Amostragem
N_w	256	Tamanho dos Frames
N_{adv}	128	Avanço da Janela de Segmentação
L^*	48	Ordem dos Vetores Quantizados
W_{hamming}	Hamming	Janela de Segmentação
F_c	5 kHz	Frequência de Corte
ϵ	10^{-10}	Menor Valor Considerável
B_f	120 mels	Largura de Banda dos filtros
M_f	35	Taxa de Amostragem do Mapa Fonético
M_c	15	Número de Classes usadas na Transformação

Tabela 3.1: *Variáveis globais do sistema e valores default*

Definidas as variáveis globais do sistema, segue-se a fase de implementação. No entanto, dada a redundância nas chamadas dos módulos funcionais, optou-se por organizar os módulos funcionais em etapas de processamento. Observe na Figura 3.4, os módulos agrupados por cores distintas. Cada agrupamento é uma etapa distinta do sistema de conversão de voz, que será aprofundada na seção seguinte.

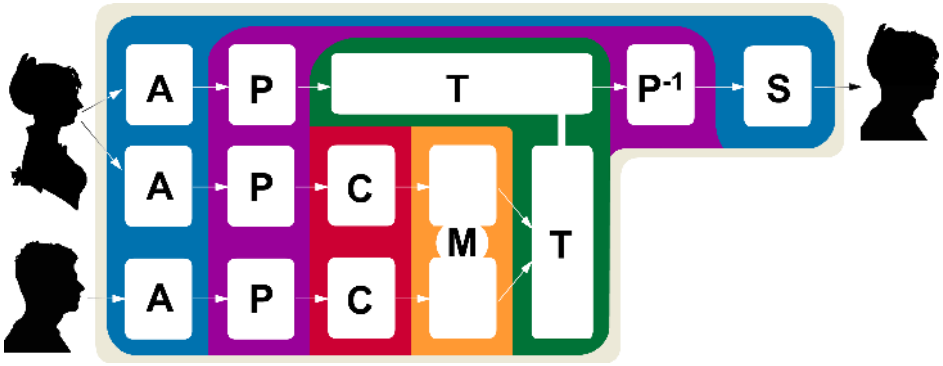

Figura 3.4: *Visão do sistema dividido em módulos funcionais.*

3.2 Estágio I: Modelagem Harmônico-Estocástica

Dentro do esquema estrutural do sistema, o módulo de análise e síntese (ver Figura 3.5) é a primeira fase do sistema, a qual deve propiciar uma representação intuitiva dos segmentos de voz (análise), a fim de se obter uma reconstrução fidedigna do sinal de voz (síntese). A premissa de reconstrução perfeita é descartada, uma vez que a reconstrução perfeita da componente estocástica do sinal é impossível a partir de uma modelagem paramétrica.

Dado o direcionamento desta pesquisa, a escolha de um bom modelo para Análise e Síntese leva em consideração os seguintes requisitos:

- O modelo deve reconstruir o sinal de voz com qualidade e naturalidade da sentença pronunciada, de modo que o resultado da análise-síntese sem modificação seja perceptualmente indistinguível do sinal original, ou o mais próximo disto o possível;

- A representação compacta e transparente do sinal de voz deve ser feita com poucos parâmetros acústicos significativos.

- A modelagem deve se adaptar aos contornos prosódicos precisamente;

- Deseja-se que o modelo represente adequadamente as componentes de filtro e fonte de cada segmento de voz;

- O modelo deve favorecer uma transformação flexível e intuitiva tanto dos elementos prosódicos quanto dos envelopes espectrais, tanto em escala global quanto local, sem a introdução de artefatos audíveis.

Os modelos gerais de decomposição do sinal a partir da envoltória espectral, e posteriormente, de estimação da fonte de excitação glotal, tem o inconveniente de representar o pulso glotal através de um sinal ruidoso, com informações revelantes e de difícil interpretação, onde se mistura a parte ruidosa e harmônica. Neste sentido, a conversão desta componente da fonte se torna um grande problema. Em contrapartida, as modelagens do fluxo glotal ainda não apresentam resultados satisfatórios em termos perceptuais, dada a dificuldade de encontrar precisamente os parâmetros do modelo [54; 98].

Recentemente o Modelo Harmônico-Estocástico [51] (*Harmonic plus Stochastic Model* ou HSM), proposto pelo pesquisador grego Yannis Stylianou [255; 258], se destaca como uma das mais apropriadas ferramentas para modelagem de voz, além de ser o que melhor se ajusta aos requisitos listados

acima. Diferentemente da modelagem fonte-filtro, a estimação é feita em ambas as componentes conjuntamente, representando a parte harmônica do sinal como uma soma de osciladores senoidais munidos de uma certa envoltória espectral. A partir de então, presume-se que a parte ruidosa do sinal é composta por um tipo de ruído colorido, posteriormente modelado usando o ferramental disponível em modelagem fonte-filtro (neste caso, o ruído branco gaussiano é a componente da fonte). Note que o modelo HSM é também uma poderosa ferramenta para filtragem de ruído em sinais de voz.

Este modelo oferece um arcabouço para reconstrução de voz com alta fidelidade, com eficiente decomposição do sinal em termos de componentes harmônicas e estocásticas, transparência na representação das amplitudes e fases das componentes harmônicas do sinal e flexibilidade na manipulação destes parâmetros acústicos. Além disso possui boas propriedades concatenativas aplicadas à síntese do sinal e fornece uma representação compacta do sinal de voz.

Existem ainda muitas facilidades com respeito à transformação de prosódia que provém imediatamente do modelo. Uma delas é a flexibilidade e robustez na modificação das componentes harmônicas, independentemente da componente estocástica. No caso do modelo HSM, um simples fator de escala multiplicado às frequências fundamentais de cada segmento é suficiente para transpor o pitch (*pitch-shifting*) de todo sinal de voz sem adição de ruídos indesejados. Esta modificação prosódica de alta qualidade é resultante da preservação da coerência entre as fases de cada componente harmônica em segmentos consecutivos, o que propicia uma reconstrução contínua da forma de onda do som reconstruído.

Esta facilidade na manipulação do pitch do sinal de voz confere uma vantagem ao modelo HSM em relação a outras alternativas. A transformação prosódica baseada em PSOLA [165], por exemplo, exige ferramentas auxiliares para marcação de pitch, as quais são bastante imprecisas na presença de ruído. Essa característica, somada ao fato de que o PSOLA é bastante sensível aos erros destas marcações, faz com que tal técnica se aplique principalmente a registros com alta relação sinal-ruído. Por outro lado, técnicas de manipulação espectral direta, como por exemplo o Phase-Vocoder [58], por não preservar a envoltória espectral acabam por distorcer a estrutura formântica do sinal de voz.

Por estas razões, neste trabalho foi adotado o HSM como o núcleo central para o processo Análise/Síntese do sinal de voz, o qual está em conformidade com a estrutura adotada pelo sistema (ver Figura 3.5).

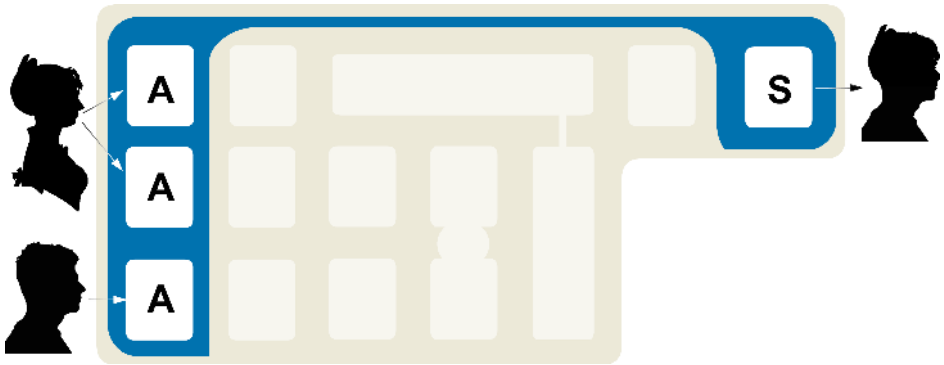

Figura 3.5: *Módulo de Análise (**A**) e Síntese (**S**)*

Antes de apresentar a implementação, será discutido a seguir um tema que traz dificuldades

para o modelo, que é o problemas das fases iniciais dos segmentos consecutivos no sinal de voz.

O problema das fases iniciais

O problema da modelagem de fases é um tópico antigo, mas ainda bem estudada atualmente dada a complexidade envolvida no problema. A modelagem Harmônico-Estocástica devolve um conjunto de fases instantâneas relativas à configuração central da respectiva componente harmônica. No entanto, tais informações de fase são tradicionalmente consideradas sem importância [116; 117; 151] sob o ponto de vista perceptual, e dada sua dificuldade intrínseca, sua modelagem é comumente negligenciada, sendo inclusive descartado o uso da fase em alguns sistemas de realce na reconstrução da fala [291].

Obviamente, em trechos curtos de sinais periódicos, a mudança abrupta de fases instantâneas é de fundamental importância para o reconhecimento de sons plosivos e outras descontinuidades do sinal. No entanto, considera-se frequentemente que a insensibilidade do ouvido quanto às fases se estenda também a trechos razoavelmente longos de um sinal periódico. A intuição de que o ouvido seria insensível às fases em trechos (quase-)periódicos se apoia no fato de que a percepção humana é insensível a diferenças de fase na chegada simultânea de dois sinais senoidais de diferentes frequências ao ouvido [150].

Entretanto, a importância perceptual das fases em sons musicais e de voz tem sido recentemente estudada em áreas como música eletroacústica e fisiologia auditiva. Por um lado, em trabalhos musicais as fases iniciais são frequentemente tratadas como componente descartáveis e redundantes na representação do sinal [196; 225; 226]. Por outro lado, em tecnologia da fala alguns autores tentam investigar os fenômenos obscuros causados pela contribuição das fases iniciais em sinais de voz [4; 134; 192].

Este problema foi detectado pelo autor desta tese a partir de resultados obtidos no deslocamento de frequência fundamental em sinais modelados somente a partir de espectro de magnitude [137]. Estes resultados indicavam que as vozes masculinas eram mais prejudicadas perceptualmente pela transformação, em contraposição às vozes femininas. A conjectura de que nosso sistema auditivo é sensível às fases iniciais, pelo menos em sinais sonoros harmônicos foi atestada no Apêndice A.2. Tal resultado sugere que a modelagem perceptual do sistema auditivo deva considerar tais relações fásicas entre componentes harmônicas, tema este que fica como sugestão para pesquisas futuras. No escopo do trabalho, a modelagem desta configuração será realizada visando diminuir o impacto da degradação do sinal, principalmente na presença de um número grande de componentes harmônicas.

Para modelar apropriadamente as componentes de fase de uma série de segmentos harmônicos, o termo linear em fase (o qual depende unicamente do instante de análise) deve ser removido da fase instantânea. Inúmeras técnicas têm sido propostas para contornar este problema, por exemplo, utilizando a abordagem pitch-sincronizada, ou cálculos de pontos invariantes em cada período (tais como deslocamentos temporais [197], centros de gravidade [259], seções de Poincaré [81] ou instantes de fechamento glotal [53]).

3.2.1 Modelagem da Configuração Fásica

O conjunto de fases iniciais do modelo HSM corresponde a um conjunto de sequências $\phi^k = [\phi_1^k, \phi_2^k, \ldots, \phi_L^k]$ associados ao k-ésimo segmento vozeado de voz dentro da representação do sinal.

Tais fases são estimadas pelo método de análise do modelo HSM e posteriormente são usados na fase de síntese, conforme é mostrado. na Equação 3.1 abaixo:

$$s_h^{(k)}[n] = \sum_{m=0}^{L} A_m^k \cos\left(\frac{2\pi m f_0^k n}{R} + \phi_m^k\right), \quad n = \left[-\frac{N}{2}, \frac{N}{2}\right]. \tag{3.1}$$

A fase instantânea de cada componente harmônica $s_h^{(k)}$ depende de ϕ_m^k que por sua vez depende essencialmente do instante de análise, assim como da frequência fundamental associado ao termo linear em fase

$$\theta_m^k(n) = \frac{2\pi m f_0^k n}{R}.$$

Entretanto, cada uma das fases ϕ_m^k, tais como são acima descritas, carregam consigo os termos lineares, propagados pelo segmento anterior, isto é,

$$\theta_m^{(k-1)\to(k)} = \theta_m^{k-1}\left(\frac{N}{2}\right) = \frac{\pi m f_0^{k-1} N}{R},$$

de tal maneira que a simples manipulação de frequência fundamental, ou mesmo da duração do sinal, exige uma série de cálculos complicados, dada esta "contaminação temporal". Desta maneira, a fase inicial da k-ésima componente harmônica pode ser decomposta em três componentes

$$\phi_m^k = \varphi_m^k + \theta_m^{(k-1)\to(k)} + \eta_m^k$$

correspondendo respectivamente ao termo associado ao trato vocal φ_m^k, ao termo linear propagado $\theta_m^{(k-1)\to(k)}$ e a um ruído de fase η_m^k proveniente do erro de estimação harmônica.

Embora alguns autores [11] tratem o termo do trato vocal agregado ao ruído de fase, é conveniente pensar que as fases estão associadas às configurações dos polos que modelam o trato-vocal, sendo assim considerados como uma componente determinística, pelo menos em sentido amplo. Sendo assim, tal informação é considerada importante na modelagem fásica proposta neste trabalho. Em contrapartida, o descarte da componente $\theta_m^{(k-1)\to(k)}$ na fase de análise pode ser realizado, uma vez que tal componente está associada à propagação linear de um segmento para outro, que pode ser facilmente re-estimada na fase de síntese. A eliminação do termo linear possibilita, por exemplo, que sejam realizadas as operações de *pitch-shifting* e mudança de velocidade do áudio sem nenhuma complicação adicional. Segue-se então a modelagem da configuração fásica do termo associado à configuração dos polos do sinal.

Define-se como *Configuração Fásica* (ou configuração em fase), a relação existente entre os termos φ^k correspondentes ao trato vocal em relação a φ_1^k (fase da harmônica fundamental) ao longo do tempo. Neste caso, cada fase φ_m^k, com $m \geq 2$, armazena o deslocamento relativo à fase da harmônica fundamental. Obviamente, tanto a fase da componente DC ($m = 0$), quanto a da harmônica fundamental ($m = 1$) são descartadas.

Intuitivamente, pode-se pensar em estimar as componentes φ, tomando-se a diferença de fases iniciais entre dois segmentos consecutivos $k-1$ e k, ou seja,

$$\varphi_m^k = \phi_m^k - \left(\phi_m^{k-1} + \frac{2\pi m f_0^{k-1}}{R}\right), \quad m = [0, L].$$

No entanto, a derivada discreta aplicada sobre as fases distorce completamente a configuração em fase original, dado o não descarte do ruído inerente η^k. Assim, por menor que seja a influência deste ruído em φ_m^k, o resultado da propagação acumulada do mesmo em segmentos seguintes é catastrófico. Alguns autores inclusive descartam tais termos e usam técnicas sofisticadas para restaurar tal configuração a partir dos centroides de cada fase ao longo do tempo [259]. No entanto, existe uma abordagem muito mais simples, que estima diretamente as componentes φ^k unicamente a partir de ϕ^k, para cada segmento.

Para fins de notação, considere sem perda de generalidade que ϕ é a sequência de fases de um segmento arbitrário, computada pelo modelo HSM.

O deslocamento temporal (atraso) em fase (*Path Difference*) corresponde à variação sofrida pela m-ésima componente harmônica, desde o instante em que se verifica que sua respectiva fase cruzou a origem, ou seja, $\phi_m = 0$. Matematicamente, o deslocamento temporal em fase de um harmônico m de um segmento de voz qualquer é dado por

$$\delta_m^t = \frac{R\phi_m}{2\pi m f_0}, \tag{3.2}$$

onde f_0 é a frequência fundamental do sinal harmônico e R é a taxa de amostragem.

A determinação da configuração de fase é então obtida tomando-se as frequências instantâneas do segmento no instante no qual o deslocamento temporal em fase de frequência fundamental for nulo, i.e. $\delta_1^t = 0$. Neste caso, todas as componentes estarão sincronizadas com o harmônico fundamental e o ruído de fase η é eliminado por este deslocamento. Então, a Configuração Fásica (CF) do segmento é obtida ao eliminar este lapso de tempo de todas as componentes de fase. Assim, cada fase ϕ_m é definida como

$$\varphi_m = \frac{2\pi m f_0 \left(\delta_m^t - \delta_1^t\right)}{R}.$$

Expandindo a Equação 3.2 dentro desta última, é derivada a equação

$$\varphi_m = \phi_m - m\phi_1 \tag{3.3}$$

correspondendo à CF do segmento em questão.

A Figura 3.6 exibe um exemplo de alinhamento das fases usando a abordagem de CF. O sinal de entrada é composto por três sinais naturais, da vogal [a] [2] sustentada por 4 falantes arbitrários. Cada linha exibe três segmentos consecutivos no tempo reconstruídos sem a propagação linear das fases de um segmento para outro, ou seja, cada segmento do sinal é reconstruído como se fosse o primeiro segmento da sequência. Em vermelho claro ao fundo está o mesmo segmento original com a componente de fase linear presente.

Esta transformação remove a contribuição linear da fase associada à frequência de cada harmônico e permite que esta seja modelada facilmente para posteriores manipulações. Note que tal alinhamento preserva a forma de onda reconstruída ao longo do tempo, dadas as propriedades de estacionariedade e periodicidade do sinal de voz. Para a reconstrução no sinal usando a CF é necessário ter uma variável de propagação do termo linear, atualizada ao final da reconstrução de cada segmento. Ou seja, supondo que no segmento $k = 0$ tem-se que $\theta = 0$, para todo segmento $k > 0$

[2] A vogal [a] foi adotada como padrão vocálico em todos os experimentos deste capítulo, dada sua riqueza espectral na representação harmônica.

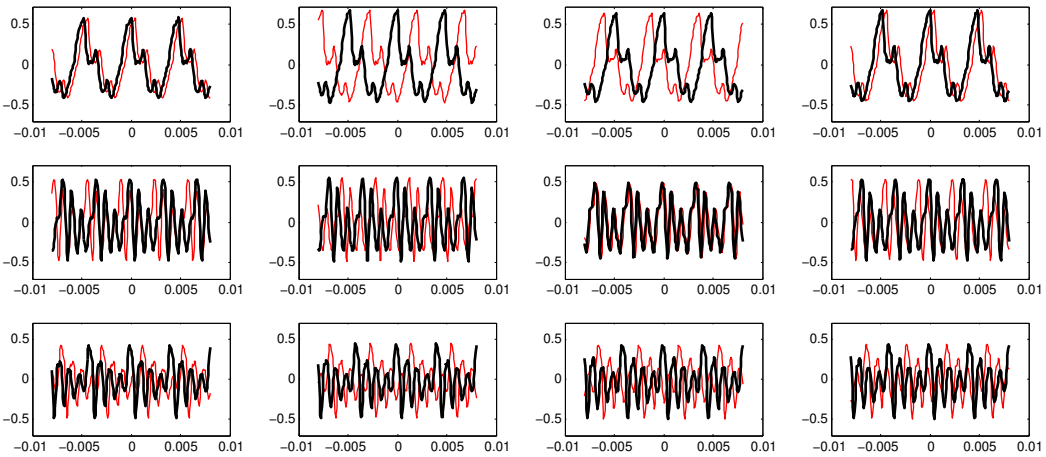

Figura 3.6: *Evolução temporal dos segmentos sem o termo de fase linear.*

as fases iniciais do modelo HSM são restauradas como

$$\phi_m^k = \varphi_m^k + m\theta, \tag{3.4}$$

e ao final do segmento, o valor de θ é atualizado de modo que

$$\theta' = \theta + \frac{2\pi f_0}{R}. \tag{3.5}$$

Existem ainda outros modelos para síntese de fala, baseados em LPC e MFCC [167], que descartam as componentes de fase e tentam reconstruí-las a partir do espectro de magnitude somente, usando a modelagem espectral sob hipótese de fase mínima [11; 36]. Recentemente, Saratxaga *et al* [217] concluíram, com experimentos comparativos usando reconstrução de fases a partir dos modelos de fase linear, fase mínima, fase zero e fase randômica, que a abordagem de fase mínima apresenta melhores resultados perceptuais dentre os modelos comparados [218], no entanto ainda com baixas taxas de aceitação perceptual quanto à qualidade da reconstrução.

A Figura 3.7 exibe um espectro de fase suavizado ao longo do tempo e normalizado entre $-\pi$ e π, de um sinal obtido pela concatenação de 3 vogais [a] sustentadas durante 2 segundos, pronunciadas por três locutores distintos. Observe que a transição suave é mantida entre um segmento e outro. Outro aspecto interessante é que os três sinais possuem representações em fase próximas uma das outras, uma vez que a vogal sustentada é a mesma nos três casos.

A seguir serão apresentadas as implementações dos módulos de Análise e Síntese, respectivamente.

3.2.2 Implementação

Esta seção visa desenvolver os módulos funcionais de Análise e Síntese descritos nas Seções 3.1.1 e 3.1.2. A implementação do modelo HSM utiliza basicamente as formulações detalhadas que podem ser encontradas na Seção 2.2.2.

O Algoritmo 3.3 corresponde ao Módulo `Análise` do sistema. Tal módulo é responsável por

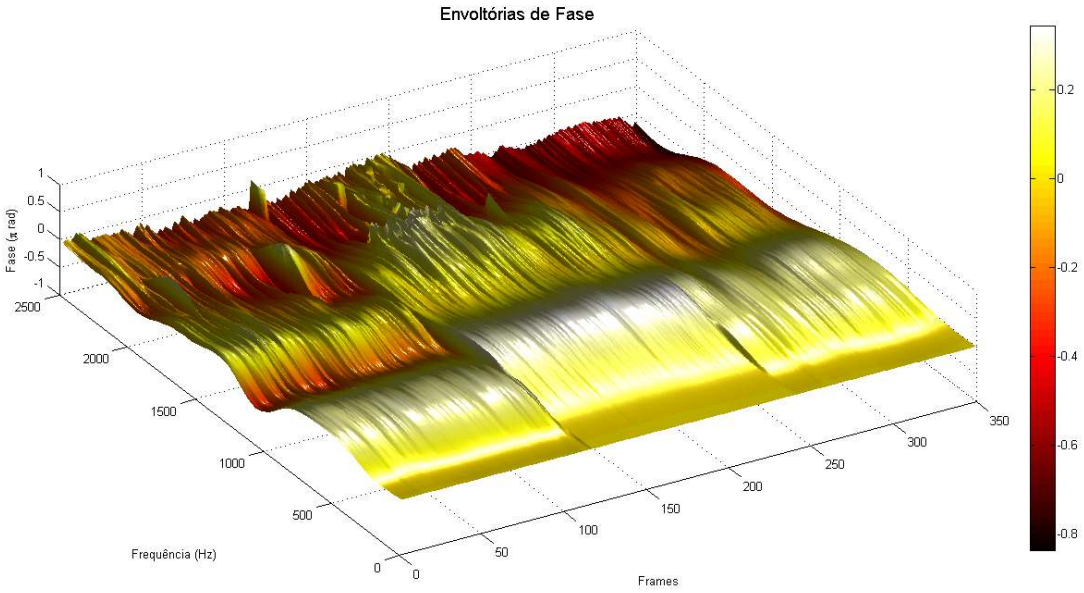

Figura 3.7: *O espectro de fases interpolado.*

tomar o sinal de voz, segmentá-lo em pequenos trechos estocásticos de acordo com os parâmetros globais do algoritmo, e extrair informações suficientes de cada segmento, para parametrização no seguinte módulo.

O problema de detectar a frequência fundamental é um problema muito explorado na literatura, de modo que este trabalho não se concentrou e propor novas alternativas. Por esta razão, a função `detecta_f0` na Linha 1 utiliza o Algoritmo Praat proposto por Paul Boersma [21], que atualmente é considerado um dos mais robustos na presença de ruído, e será textualmente explicado abaixo.

O Algoritmo Praat é um algoritmo detector de pitch (*do inglês, Pitch Detector Algorithm* – PDA) paramétrico, sendo atualmente um padrão na literatura para detecção de pitch de sinais de voz [51; 258]. Um breve resumo dos passos do algoritmo é apresentado abaixo:

(1) – O primeiro passo do algoritmo consiste em remover os lóbulos laterais da janela de Hann usada na segmentação do sinal em componentes próximas à frequência de Nyquist. Uma super-amostragem é realizada e um filtro passa-baixas elimina as componentes de alta frequência do sinal.

(2) – Encontrar o maior valor absoluto (pico) do sinal no tempo.

(3) – Uma vez que o método é baseado na análise de trechos curtos do sinal de voz, a análise pode ser feita usando um número pequeno de parâmetros. Sendo assim, o algoritmo toma deslocamentos do sinal original em passos, definidos pelo parâmetro *TimeStep* (o padrão é 0.01 segundos). Então, para cada um destes segmentos, o algoritmo busca por *lags* da função de autocorrelação, cujos valores o classificam como pertencente ou não ao grupo de *lags* candidatos à periodicidade do segmento. O número de candidatos é definido pela variável *MaximumNumberOfCandidatesPerFrame* (o padrão é 4). Todos os passos a seguir são realizados para cada segmento.

(3.1) – Tomar um segmento do sinal; o tamanho deste segmento, assim como da função de janelamento é determinado pelo parâmetro *MinimumPitch*, que se adapta à menor frequência

fundamental que se deseja detectar. A janela deve ser ajustada de modo a conter pelo menos três (para detecção de pitch) ou seis períodos (para medições de razões harmônico-ruído) do pitch mínimo (MinimumPitch).

(3.2) – Eliminar a componente DC do segmento.

(3.3) – Ajustar os parâmetros de limiar *VoicingThreshold* e *SilenceThreshold*. Ambos definem um limiar sobre os picos da autocorrelação normalizada (ACN) e do sinal no tempo, respectivamente. Adotando $VoicingThreshold = 0.4$ e $SilenceThreshold = 0.05$, um segmento é considerado não-vozeado se não tiver pelo menos um pico acima de 0.4 na ACN, e silêncio, caso o maior pico do sinal no tempo for inferior a 0.05.

(3.4) – Janelar o sinal usando a função de Hann.

(3.5) – Anexar ao início do segmento uma sequência de zeros com a metade do tamanho da janela.

(3.6) – Preenche com zeros até que se complete uma potência de dois (*Zero Padding*).

(3.7) – Tomar a FFT do sinal.

(3.8) – Construir espectro de energia (quadrado da FFT anterior).

(3.9) – Tomar a FFT do espectro de energia.

(3.10) – Dividir o sinal pela auto-correlação $r(\tau)$ da janela calculada no passo 3.5.

(3.11) – Encontrar posições e alturas máximas na versão contínua reconstruída de $r(\tau)$, definida como

$$r_c(\tau) = \sum_{n=1}^{N} \{s_n^N(\alpha_l) + s_n^N(\alpha_r)\},$$

onde

$$s_n^N(\alpha) = \mathrm{sinc}(\alpha_l + n - 1)\left(\frac{1}{2} + \frac{1}{2}\cos\frac{\pi(\alpha_l + n - 1)}{\alpha_l + N}\right),$$

$n_l = \lfloor \frac{\tau}{\Delta\tau} \rfloor$, $n_r + n_l = 1$, $\alpha_l = \frac{\tau}{\Delta\tau} - n_l$ e $\alpha_r + \alpha_l = 1$. Somente picos máximos localizados entre *MinimumPitch* e *MaximumPitch* são considerados. Além disso, um fator de força relativa que toma medidas como máximo local do segmento, máximo global do sinal, bem como os valores pré-definidos *VoicingThreshold* e *SilenceThreshold* são usados para definir se um segmento é considerado vozeado ou não.

Ao final da detecção do pitch em cada janela, o algoritmo ainda refina a busca pelas frequências fundamentais usando programação dinâmica a fim de eliminar erros de oitava. Este detalhes técnicos do algoritmo levam em conta elementos particulares do processo de produção da voz, como relações energéticas entre harmônicos vizinhos, o que ajuda a corrigir os erros de oitava. Embora o método *Praat* tenha sido usado no Algoritmo `Análise`, qualquer outro algoritmo robusto para estimação de pitch pode ser usado.

Algoritmo 3.3 $(F_0, S, A, \varphi) \leftarrow$ `Análise`(x)

▷ *Módulo Detector de Pitch*
1 $F_0 \leftarrow \text{detecta\_f}_0(x)$

▷ *Define Parâmetros de Janelamento*
2 $w \leftarrow \text{janela}(W_{\text{hamming}}, N_w + 1)$
3 $W \leftarrow \text{diag}(w)$
4 $N_x \leftarrow \text{dimensão}(x)$

▷ *Módulo Segmentador*
5 **for** $k = 1 : N_{\text{adv}} : N_x$
6 $\quad s \leftarrow x(k : k + N_w + 1)$

▷ *Decisão V/UV*
7 $\quad$ **if** $F_0(k) > 0$ **then**

▷ *Cálculo do Número de Harmônicos*
8 $\quad\quad L \leftarrow \left\lfloor \frac{F_c}{F_0(k)} \right\rfloor$

▷ *Estimação Harmônica por Mínimos Quadrados*
9 $\quad\quad$ Calcule B pela Eq. 2.8 usando $(N_w, L, F_0(k))$
10 $\quad\quad c \leftarrow (B^{-H}WWB)(B^{-H}WWs)$

▷ *Extração das Amplitudes e Fases da Parte Harmônica do Sinal*
11 $\quad\quad A^{(k)} \leftarrow |c(L + 1 : 2L + 1)|$
12 $\quad\quad \phi \leftarrow \angle c(L + 1 : 2L + 1)$

▷ *Elimina o Termo Linear das Fases*
13 $\quad\quad \varphi^{(k)} \leftarrow \phi - \phi(2)(0 : L)$

▷ *Obtenção da Parte Estocástica do Sinal*
14 $\quad\quad \hat{s} \leftarrow cB$
15 $\quad\quad s_s \leftarrow s - \hat{s}$
16 $\quad$ **else**

▷ *Frame Não-Vozeado é Considerado Parte Estocástica do Sinal*

17 $\quad\quad s_s \leftarrow s$
18 $\quad$ **end if**
19 $\quad S^{(k)} \leftarrow s_s$
20 **end for**
21 **return**(F_0, S, A, φ)

Comentando cada linha seguinte do Algoritmo `Análise` (Alg. 3.3), temos que a Linha 2 representa o carregamento da janela de Hamming. Note que seu tamanho é ímpar, a fim de facilitar posteriormente o *Overlap-Add* no módulo de síntese. A função `diag` da Linha 3 toma um vetor v de tamanho N e cria uma matriz diagonal com as entradas de v.

A decisão V/UV contida na Linha 7 está apoiada na detecção de pitch anterior, e considera um segmento como vozeado se existe uma frequência fundamental maior que zero (para os não-vozeados, $f_0 = 0$). A Linha 8 define o número de harmônicos (L) de acordo com a razão entre a frequência de corte e a frequência fundamental. As linhas 9 e 10 são fundamentadas nas Equações 2.8 e 2.9, respectivamente, assim como as linhas 11 e 12 são fundamentadas na Equação 2.10. Todas estas

equações estão descritas no Capítulo 2, Seção 2.2.2.

A Linha 13, por sua vez, aplica os conceitos de configuração fásica abordados anteriormente. Sua formulação pode ser encontrada na Equação 3.3. A Linha 14 reconstrói o segmento harmônico a fim de obter a parte estocástica da diferença apresentada na Linha 15.

No caso da Linha 17, por não haver parte harmônica, s é a própria parte estocástica do sinal. A Linha 19 armazenam o segmento estocástico do trecho analisado.

Algoritmo 3.4 $x \leftarrow$ `Síntese` (F_0, S, A, φ)

▷ *Define Parâmetros de Janelamento*
1 $w \leftarrow \text{janela}(W_{\text{hamming}}, N_w + 1)$
2 $W \leftarrow \text{diag}(w)$
3 $N_\Psi \leftarrow \text{dimensão}(F_0)$
4 $N_x \leftarrow 0$
5 $\theta \leftarrow 0$

▷ *Módulo Montador*
6 **for** $k = 1 : N_\Psi$

▷ *Reconstrução Harmônica*
7 **if** $F0(k) > 0$ **then**

▷ *Cálculo do Número de Harmônicos*
8 $L \leftarrow \left\lfloor \frac{F_c}{F0(k)} \right\rfloor$

▷ *Formula de Reconstrução Harmônica*
9 $\phi \leftarrow \varphi^{(k)} + \theta(0 : L)$
10 $s_h \leftarrow \sum_{m=0}^{L} A^{(k)}(m) \cos\left(\frac{2\pi m F0(k)}{R}(-N_{\text{adv}} : N_{\text{adv}}) + \phi_m\right)$

▷ *Propagação do Termo Linear*
11 $\theta \leftarrow \text{empacotar\_fase}\left(\theta + \frac{2\pi N_{\text{adv}} F0(k)}{R}\right)$
12 **end if**

▷ *Reconstrução Estocástica*
13 $s_s \leftarrow S_s^{(k)}$

▷ *Composição do segmento*
14 $s^{(k)} \leftarrow (s_h + s_s)$
15 **for** $n = 0 : N_{\text{adv}}$
16 $x(n + kN_W) \leftarrow \left(\frac{N_w - 2n}{N_w}\right) s(n-1)^{(k-1)} + \left(\frac{2n}{N_w}\right) s(n)^{(k)}$
17 **end for**
18 **end for**
19 **return**(x)

A parametrização em filtros passa-bandas, como sugere a modelagem original HSM, será descrita na seção de parametrização (próxima seção). Naquele contexto, será tomado o espectro de magnitude da parte estocástica, que será envelopado e parametrizado; estes procedimentos serão detalhados na próxima seção. O módulo de `Síntese`, por sua vez, não apresenta grandes complicações diante da explanação do módulo de Análise. Seu objetivo é tomar um conjunto de coeficientes harmônicos e estocásticos (uma sequência por segmento) e devolver o sinal reconstruído no tempo devidamente montado. O Alg. 3.4 corresponde ao respectivo módulo do sistema.

Como proposta do trabalho para incrementar o modelo HSM, foram inseridas as linhas 5 e 9 no

módulo de Síntese apresentado, onde se realiza a propagação do termo linear em fase θ, associado à frequência fundamental (ver Eq. 3.5).

Particularmente, a Linha 10 do Alg. 3.4 usa a função `empacotar_fase`, cujo objetivo é transferir a fase para o intervalo $[-\pi, \pi]$, a partir de deslocamentos de $\pm\pi$, a fim de evitar estouro de memória na implementação algorítmica.

A Linha 9 corresponde a restauração da fase inicial da componente harmônicas, definida pela soma da contribuição do termo linear e do deslocamento de fase φ (ver Eq. 3.4).

A Linha 10 reconstrói o segmento harmônico, somando as contribuições de cada componente harmônica, de acordo com a Equação 2.11 (síntese sequencial). A Linha 13, analogamente, retoma a parte estocástica do sinal de voz, a fim de compor o segmento reconstruído (Linha 14).

Finalmente, as Linhas 15, 16 e 17 realizam o *Overlap-Add* usando janelamento triangular, como discutido na Seção 2.2.2, Equação 2.12.

Uma vez definidos os módulos de Análise e Síntese do sistema, segue o próximo módulo, para modelagem destes parâmetros acústicos.

3.3 Estágio II: Decomposição Paramétrica

O modelo de parametrização propõe-se a normalizar os parâmetros acústicos de cada segmento, a fim de caracterizá-los e fornecer um espaço no qual eles possam ser comparáveis entre si. O objetivo desta seção é apresentar implementações do módulo de parametrização e sua respectiva inversa (Fig. 3.8).

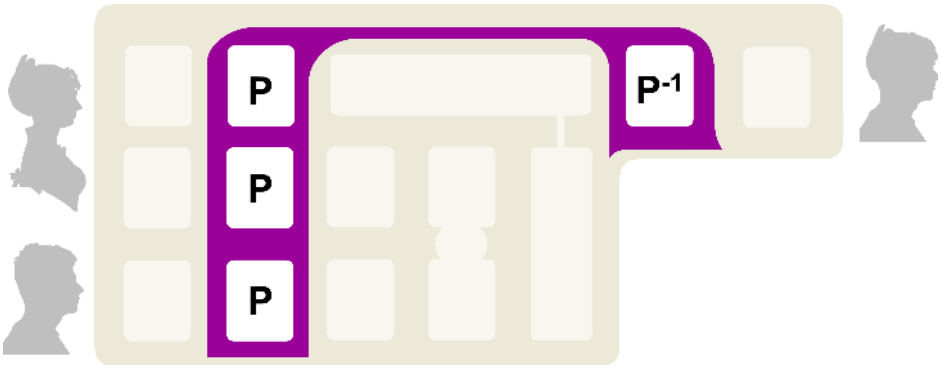

Figura 3.8: *Fase de Parametrização (**P**) e sua respectiva Inversa (**P**$^{-1}$)*

A quantização dos parâmetros acústicos devolvidos pelo modelo HSM é de grande relevância em tarefas de conversão espectral e prosódica. Uma vez que as frequências fundamentais de cada segmento de voz já estão bem definidas para cada segmento, é necessário representar os parâmetros A e S do modelo usando um número fixo de coeficientes. No entanto, não basta selecionar as k primeiras componentes harmônicas, pois não se deseja perder informações relevantes do espectro.

O módulo parametrizador segue implementado pelo Alg. 3.5.

Algoritmo 3.5 $\Psi \leftarrow$ `Parametrização`(F_0, S, A)

▷ *Estimação das envoltórias espectrais dos espectros de entrada*

1 $[@\upsilon_S, @\upsilon_A] \leftarrow \text{envelopador}(S, A, F_0)$

▷ *Quantização usando Decomposição Espectral*

2 $\Psi_{\{S,A,E_0\}} \leftarrow \text{quantizador}(@\upsilon_S, @\upsilon_A)$

▷ *Modelagem das frequências fundamentais (escala Mel) e das energias*

3 $\Psi_{F_0} \leftarrow 1127.01048 \log\left(1 + \frac{F_0}{700}\right)$

4 $\Psi_{E_0} \leftarrow E_0$

5 **return**(Ψ)

A ideia central do módulo parametrizador é providenciar dois envelopes espectrais de amplitude, o envelope harmônico (1) e o envelope estocástico (2), e os decompor em somas de funções bases (filtros) com largura de banda e formas flexíveis. Vale salientar que a modelagem das fases iniciais por meio de envelopes espectrais de cada segmento serão consideradas no trabalho, de modo que as mesmas passarão diretamente do módulo de análise para síntese na etapa de conversão. A partir dos envelopes de magnitude, o método de representação espectral destes envelopes é acionado, a fim de os decompor em somas de funções paramétricas distribuídas em bandas de frequência (banco de filtros).

Analogamente, é apresentado a seguir o módulo de parametrização inversa, cuja função é recuperar cada vetor (F_0, S, A) a partir de Ψ.

Algoritmo 3.6 $(F_0, S, A) \leftarrow$ `Parametrização_Inversa`(Ψ)

▷ *Restaura as frequências fundamentais e as energias*

1 $F_0 \leftarrow 700\left(e^{\frac{\Psi_{F_0}}{1127.01048}} - 1\right)$

2 $\Psi_A \leftarrow \Psi_A \cdot E_0$

▷ *Sub-amostragem das envoltórias espectrais dos espectros de entrada*

3 $[S, A] \leftarrow \text{sub\_amostrador}(\Psi_{\{A,S\}}, F_0)$

4 **return**(F_0, S, A)

Vale salientar que na parametrização os coeficientes de envoltória harmônicos estão normalizados na escala logarítmica, fazendo com que E_0 também esteja. O próprio módulo `sub_amostrador` se encarrega de transpor os valores de amplitude novamente à escala linear.

A proposta de representação espectral deste trabalho visa analisar diferentes funções base (filtros com decaimento radial) para representação de sinais de voz. A partir da imposição de diferentes restrições no algoritmo de estimação é possível oferecer o controle entre eficiência e acurácia do método. Note que a função a ser representada, o envelope espectral, depende somente de uma única dimensão (a frequência), e esta é estritamente positiva. Além disso, deseja-se que todos os parâmetros do modelo (localização, amplitude e largura de banda) de cada base mantenham a continuidade temporal entre segmentos adjacentes. Algumas propostas de envelopamento espectral, bem como métodos de estimação para obter estes parâmetros serão derivados. Por exemplo, um destes métodos considera que existe uma função base (ou filtro) por banda fixa sobre a escala MEL, sendo similar ao banco de filtros sobre a escala MEL representados pelos coeficientes MFCC, porém o resultado do ajuste não é simplesmente um conjunto de parâmetros isolados, mas sim uma aproximação contínua do envelope espectral [93].

Envelopamento Espectral

O primeiro passo do Módulo de Parametrização é caracterizar as componentes harmônica (amplitudes) e estocástica através de seus respectivos envelopes espectrais, para posterior modelagem paramétrica. Define-se por *Envelopamento Espectral* o processo de obtenção de uma versão suavizada super-amostrada do espectro original, que corresponde a conformação da fonte de excitação sonora pelos ressonadores do trato vocal. Deseja-se, em primeiro lugar, que esta modelagem seja representada por funções contínuas, de tal modo que seja possível, por exemplo, comparar dois espectros harmônicos de magnitude ainda que estes estejam amostrados em frequências fundamentais distintas. Além disso, deseja-se ter uma representação quantizada que caracterize acusticamente cada segmento.

A versão suavizada de tais espectros pode ser obtida de várias formas, e dentre elas se destacam:

1. LPC: Conforme visto anteriormente na Seção 2.2.1, é possível extrair tais envoltórias a partir da reconstrução do sinal usando uma sequência de coeficientes LPC (de baixa ordem).

2. CEPS: Pode-se obter uma versão espectral suavizada por meio de uma filtragem passa-baixa do cepstrum estocástico.

3. INTERPOLADOR: Pode-se obter também uma versão suavizada do espectro a partir da reconstrução suavizada dos picos do espectro usando *métodos interpoladores paramétricos de base radial*(*).

O trabalho propõe a modelagem da envoltória espectral(*) enunciada no item 3 acima. Tal reconstrução exige um módulo de interpolação de funções discretas, conhecido como *interpolador radial*. O padrão adotado pela maioria das propostas encontradas na literatura é a interpolação *spline cúbica* [204]. Basicamente, tal abordagem utiliza um polinômio de ordem 3 a fim de encontrar uma sequência de valores interpolados, dados dois valores consecutivos. Visando uma versão mais fiel de interpolação aplicada especialmente aos envelopes espectrais, uma nova classe de métodos interpoladores será apresentada a seguir.

Método Interpolador por Base Radial

Um método de interpolação paramétrico encontra valores entre pontos intermediários de uma função arbitrária, usando uma *função de interpolação* auxiliar (triangular, cúbica, etc.). No caso deste trabalho, as funções de base radial, denotadas por ψ, serão usadas. Funções de base radial (*Radial Basis Functions* – RBF) são bastante utilizadas em redes neurais [236; 292]. Uma RBF é uma função real cujo valor depende somente da distância em relação a uma origem o, sendo que tradicionalmente usa-se a normalização $\psi_o(o) = 1$ e o valor $\psi_o(x)$ decresce à medida que aumenta a distância $|x - o|$.

No contexto deste trabalho é de interesse utilizar algumas RBFs com propriedades específicas de interpolação para obter os envelopes espectrais. Existem basicamente três propriedades básicas que um interpolador I definido por este trabalho deve atender. Dada uma sequência $X = \{x_1, x_2, \ldots, x_n\}$ de valores reais crescentes e uma sequência $Y = \{y_1, y_2, \ldots, y_n\}$ tal que $y_k = f(x_k),\ \forall k$, o interpolador I deve:

1. Dispor de uma base ψ_x associada a cada amostra x de X

2. Ser capaz de devolver uma representação contínua $Y_c^{\mathbf{P}}$ definida como

$$Y_c^{\mathbf{P}}(x) = \sum_{x_k \in X} \mathbf{p}_{x_k} \psi_{x_k}(x),$$

onde $\mathbf{p}_{x_k}$ é o peso da base ψ_{x_k}.

3. Manter válida a igualdade $Y_c^{\mathbf{P}}(x_k) = y_k$, para todo ponto x_k de X

4. Possuir propriedade de reconstrução (quase-)perfeita, ou seja, deve existir uma escolha particular dos pesos $\mathbf{p}$ tal que $Y_c^{\mathbf{P}}(x) \approx 1$.

Existem várias funções radiais que satisfazem tais propriedades, dentre elas a função Gaussiana e as versões truncadas das janelas de Hann, Nuttall, Blackman-Harris e Blackman-Nuttall. Tais funções são definidas pelas Equações 3.9 e 3.10, as quais serão melhor descritas em termos de largura de banda na subseção seguinte. Dentre as propriedades destas funções, destaca-se a propriedade de reconstrução (quase-)perfeita, a qual é exibida na Figura 3.9, usando as janelas de Hann, Nuttall, além da função Gaussiana.

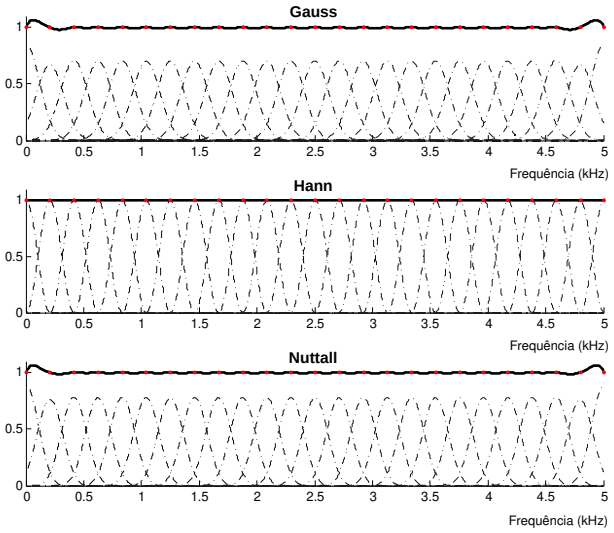

Figura 3.9: *Reconstrução da função constante $I(x) = 1$ usando diversos interpoladores radiais.*

Dentre as funções de base radial analisadas, a única que possui a propriedade de reconstrução perfeita (exata) é a função truncada de Hann, definida como

$$\psi_o(x;\sigma) = \begin{cases} \cos^2(\frac{\pi(x-o)}{2\sigma^{(l)}}), & \text{se } x \in [o-\sigma^{(l)}, 0] \\ \cos^2(\frac{\pi(x-o)}{2\sigma^{(r)}}), & \text{se } x \in [0, o+\sigma^{(r)}] \\ 0, & \text{caso contrário} \end{cases}$$

onde $[o-\sigma^l, o+\sigma^r]$ corresponde à largura de banda da função radial, fora da qual a função base é truncada.

Deve-se ter em mente que larguras de banda muito grandes podem perturbar a reconstrução do sinal causando suavização excessiva, devido à maior interferência construtiva das diversas funções base utilizadas. Sendo assim, é preferível manter as larguras de banda de cada função base o mais estreitas o possível, a fim de representar pequenas nuances da função original. Na prática,

adotaremos os próprios pontos x_k como fronteiras das funções bases, ou seja,

$$\sigma_k^{(l)} = x_k - x_{k-1} \quad \text{e} \quad \sigma_k^{(r)} = x_{k+1} - x_k, \; k = 2, \ldots, n-1. \tag{3.6}$$

Por convenção, as funções base dos pontos extremos x_1 e x_n serão simétricas, com largura definida por apenas um dos vizinhos (x_2 e x_{n-1} respectivamente). Observe que se X corresponde a uma sequência de pontos uniformemente espaçados, então $\sigma_k = \sigma_k^{(l)} = \sigma_k^{(r)}$.

Uma vez definidas as larguras e centros de cada função base, pelo Algoritmo 3.7 se obtém uma função `interpolação` que define uma versão contínua dados os conjuntos de pontos pareados $(X, Y = f(X))$. Os parâmetros de entrada X e Y são respectivamente o domínio e a imagem da função a ser interpolada nos pontos X'. A base de interpolação ψ pode ser qualquer base que satisfaça as propriedades listadas anteriormente. O valor ϵ é utilizado para garantir que a função interpolada é não-nula fora do intervalo X (vide Tabela 3.1 de variáveis globais). Neste caso, a função $@Y_c$ devolvida pelo algoritmo computa para qualquer valor $x \in [x_1, x_2, \ldots, x_n]$ o valor interpolado $y = Y_c(x)$.

Algoritmo 3.7 $@Y_c \leftarrow$ `interpolador`(X,Y,ψ,ϵ)

▷ *Inicializa as origens de cada função* ψ
1 $o \leftarrow X$

▷ *Define as larguras de banda de cada base* ψ
2 $\sigma^{(l)} \leftarrow [x_2 - x_1, x_2 - x_1, x_3 - x_2, \ldots, x_n - x_{n-1}]$
3 $\sigma^{(r)} \leftarrow [x_2 - x_1, x_3 - x_2, \ldots, x_n - x_{n-1}, x_n - x_{n-1}]$

▷ *Define os pesos de cada base*
4 $\mathbf{p} = Y - \epsilon$

▷ *Obtém a função* $Y_c(X')$ *diretamente*
5 $@Y_c \leftarrow @(F)[\epsilon + \sum_{x \in X} \sum_{f \in F} \{\mathbf{p}_x \psi_x(f)\}]$
6 **return**$(@Y_c)$

A sintaxe da Linha 5 define uma função *anônima* $@(a)[b]$, que recebe como parâmetro os argumentos em a e retorna o valor b. Como consequência, esta função devolve o conjunto de larguras de banda definidas pela Eq. 3.6.

A Figura 3.10 exemplifica interpolações de um conjunto de picos harmônicos amostrados em vermelho, com o uso do algoritmo `interpolador` para obter os envelopes harmônicos de amplitude de um segmento arbitrário de um sinal de voz usando diversas bases de interpolação.

A partir da estimação interpolada do espectro, este trabalho propõe uma nova metodologia para quantização, conforme segue.

3.3.1 Estimação do Envelope Espectral

Uma boa modelagem do envelope espectral deve corresponder o mais fielmente ao envelope espectral original. No entanto, as informações necessárias para se obter um envelope perfeito normalmente não estão disponíveis, principalmente em espectros harmônicos [138].

No caso estocástico, obter uma boa aproximação do envelope verdadeiro é mais fácil, uma vez que o padrão ruidoso do espectro estocástico fornece informação em toda a faixa de frequências utilizadas. Uma vez que os espectros de magnitude são obtidos a partir do módulo da FFT de cada

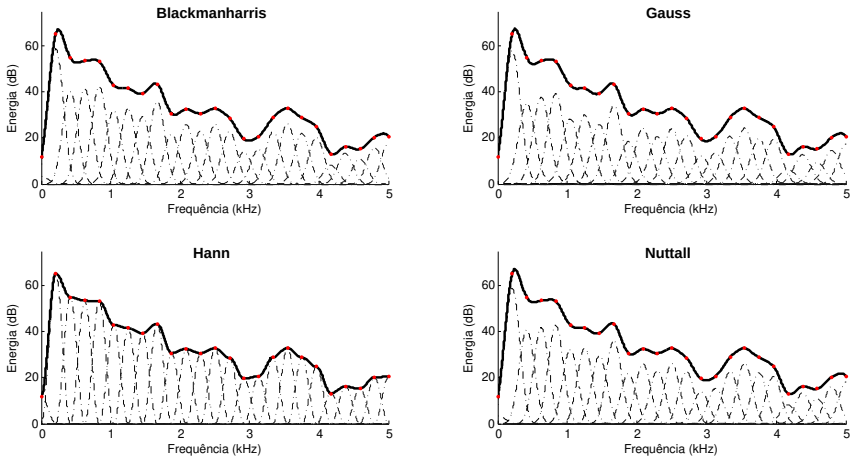

Figura 3.10: *Diversos interpoladores espectrais usando o algoritmo* `interpolador`.

segmento estocástico (cada S_k proveniente do módulo de Análise), é possível reconstruir uma versão suavizada usando um método interpolador qualquer.

Já no caso das representações espectrais harmônicas, encontrar tal envelope contínuo utilizando apenas informação local (relativa a um único segmento) é praticamente impossível. Neste caso, o módulo de Análise devolve um conjunto de amplitudes harmônicas amostradas em um conjunto restrito de frequências $0,\ f_0,\ 2f_0, \ldots,\ Lf_0$ relacionadas harmonicamente. O ideal seria obter a envoltória espectral harmônica a partir de um conjunto suficientemente grande de amostras com diferentes frequências fundamentais, mas que pertençam a um mesmo *conjunto fonético*.

Uma abordagem que utiliza um conjunto de segmentos de um mesmo fonema para reconstruir mais fielmente o padrão do envelope, conhecida como Análise Multi-Frame (do inglês, *Multi-Frame Analysis* – MFA) [227; 228; 229], toma um conjunto de amplitudes harmônicas a fim de estimar uma representação fiel da Função de Transferência do Trato Vocal (*Vocal-Tract Transfer Function* – VTTF). Este método usa o valor das amplitudes de todos os segmentos para estimar coeficientes cepstrais com um método de mínimos quadrados. Entretanto, tal abordagem não resolve o problema de estimação espectral segmento-a-segmento.

Pode-se alternativamente tomar um conjunto de segmentos vizinhos, e aplicar a estratégia AMF a cada um destes conjuntos. No caso específico de sinais de fala, a melhoria da estimação neste caso não é significativa, uma vez que a variação lenta do pitch entre segmentos consecutivos não contribui para uma boa modelagem em todo o espectro. Além disso, a quantidade de segmentos contaminados pelas mudanças de fonema prejudica a clusterização fonética do módulo seguinte (apresentado na Seção 3.4). Um pequeno experimento é realizado no Apêndice A desta tese, a fim de inspecionar alguns resultados preliminares do uso do interpolador para estimação espectral, dentro de um contexto MFA.

A Figura 3.12 exibe os envelopes harmônicos estimados pelo envelopador Gaussiano. O sinal usado nesta figura corresponde a um sinal de voz com três segmentos da vogal [a] concatenados. Cada segmento pertence a um falante distinto, os quais pronunciaram dois segundos da vogal sustentada, mantendo aproximadamente a mesma frequência.

O envelope estocástico é estimado através de filtragem espectral, a qual extrai picos (máximos) locais do sinal. A fim de eliminar a interferência causada por um eventual vazamento harmônico,

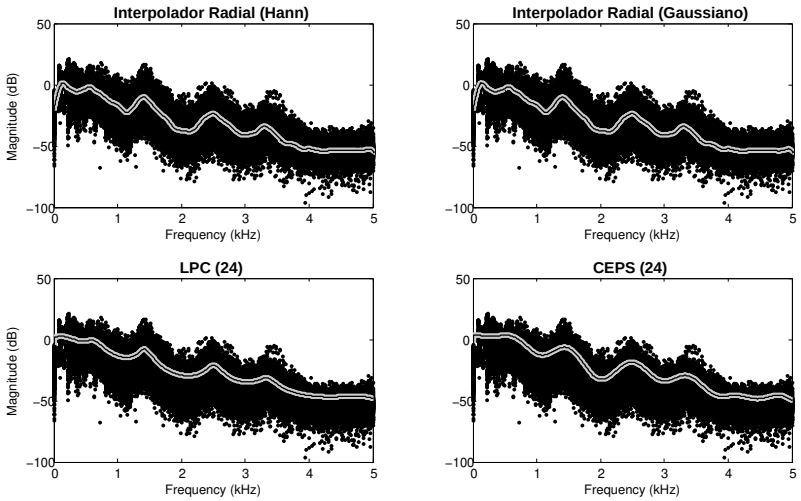

Figura 3.11: *Análise multi-frame usando alguns do envelopadores propostos.*

isto é, contaminação ocasionada por erro de estimação, uma filtragem usando valores medianos dentro uma curta janela de vizinhança (aplicada sobre o espectro) pode ser realizada com sucesso. A largura de banda ao redor de cada amostra é tomada a uma taxa definida pelo usuário entre 250 e 500 Hz, o que é suficiente para obter uma representação espectral suave do espectro estocástico. Com uma taxa de amostragem de 16 kHz, tais frequências equivalem a 4 e 8 amostras, respectivamente. O Algoritmo 3.8 implementa a função `envelopador` usada pelo módulo de Parametrização.

Algoritmo 3.8 $(@\upsilon_S, @\upsilon_A) \leftarrow$ `envelopador`(S, A, F_0)

$N_s \leftarrow$ dimensão(S)
$\sigma_s \leftarrow \frac{500 N_w}{R}$
for $k = 1 : N_s$
  if $F_0(k) > 0$ **then**
    ▷ *Envelopes harmônicos de amplitude*
    $L \leftarrow \left\lfloor \frac{F_c}{F_0(k)} \right\rfloor$
    $@\upsilon_A^k \leftarrow$ interpolador$(F_0(k)(0:L), A_k, \psi_{\text{gauss}})$
  end if
  ▷ *Envelope estocástico*
  $\sigma_c \leftarrow \lceil \frac{500 N_w}{R} \rceil$
  $[X, Y] \leftarrow$ filtro_max(S_k, σ_c)
  $@\upsilon_S^k \leftarrow$ interpolador$(X, Y, \psi_{\text{gauss}})$
end for
return(Ψ)

Como descrito anteriormente, a função `filtro_max`(S, σ_c) primeiramente estima um envelope de máximos locais M_S de S, de tal modo que

$$M_S(x) = \max\{S(x - \sigma_c : x + \sigma_c)\},\ x = [\sigma_c + 1 : N_w - \sigma_c].$$

Nos extremos tem-se $M_S(x) = S(1)$ para $x = [1, \sigma_c]$ e $M_S(x) = S(N_w)$ para $x = [N_w - \sigma_c + 1 : N_w]$. A partir de M_S, o método `filtro_max` toma os pontos x e $y = S(x)$, para os quais se verifica que $M_S(x) = S(x)$. Então, o conjunto de todos os pontos x e y são devolvidos em X e Y respectivamente,

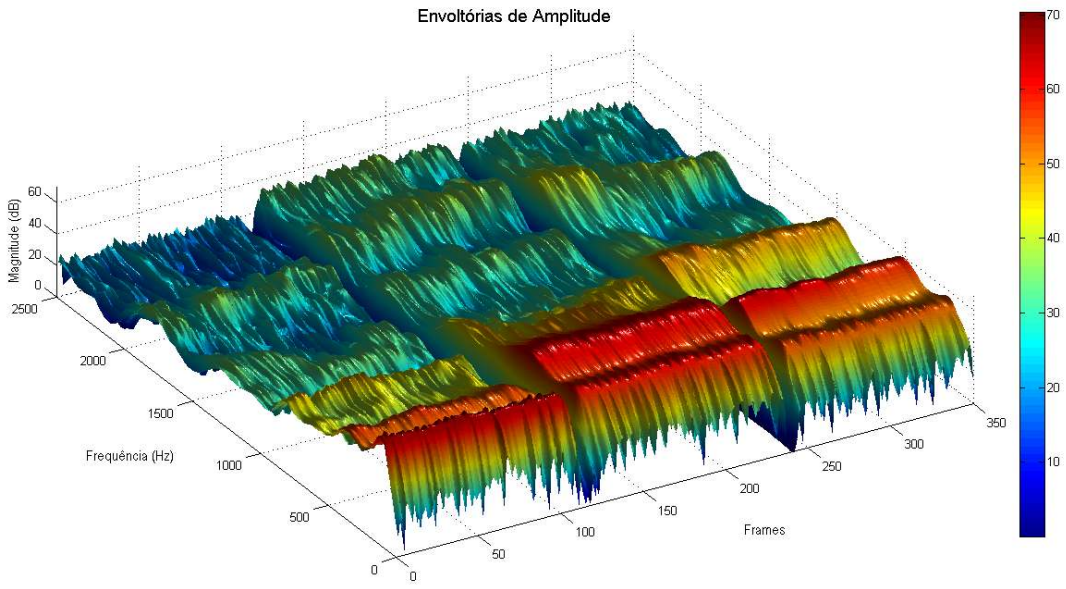

Figura 3.12: *O espectro harmônico de amplitudes envelopado pelo Interpolador Gaussiano.*

para posterior interpolação. Cabe salientar que as variáveis globais R, N_w e F_c estão definidas na Tabela 3.1.

Seria possível utilizar tais envelopes estimados como versões quantizadas de cada segmento; no entanto, o valor retornado pelo `envelopador` corresponde a um conjunto variável de coeficientes relativos às bases radiais, o que o torna inapropriado para ser usado como vetor quantizado. O ideal seria que se tivesse um tipo de interpolador, no qual os conjuntos de funções de base fossem fixos e estivessem convenientemente distribuídos dentro do espectro a ser modelado, para que posteriormente pudessem ser super-amostrados e devolvidos ao estado espectral original (com as respectivas frequências harmônicas, por exemplo).

3.3.2 Decomposição Espectral em Funções de Base

Ao longo desta seção, os dois envelopes representados pelas funções $@\upsilon_S^k$ e $@\upsilon_A^k$ referentes ao k-ésimo segmento do sinal de voz, serão denotados indistintamente por $@\upsilon$ por simplicidade de notação, embora ambos os envelopes recebam um tratamento diferenciado na fase final do processo de quantização, conforme será visto na implementação da função `quantizador` (Algoritmo 3.12). Além disso, pelo algoritmo envelopador, cada envelope $@\upsilon$ corresponde a uma versão contínua de um dos espectros suavizados (harmônico ou estocástico). Embora existam métodos estimadores de mínimos quadrados aplicados às funções contínuas (MQC), por questões de eficiência dos cálculos matriciais, uma versão discretizada E será amostrada de cada $@\upsilon$, a qual será considerada pelo método de decomposição. Sendo assim, segue a definição do método de decomposição espectral usando os mesmos tipos de funções radiais dos interpoladores radiais apresentados anteriormente.

A decomposição em função de bases paramétricas oferece pelo menos três vantagens para o processamento geral de sinais de voz:

1. Muitas vezes se deseja representar melhor algumas regiões espectrais (por exemplo as baixas frequências) em detrimento de outras, por questões psicoacústicas.

2. Em algumas transformações espectrais, requisita-se maior independência e controle de sub-bandas limitadas de frequências.

3. Como as funções modelam as localizações e amplitudes de frequências por banda, tal decomposição fornece uma representação conveniente para a realização simultânea de empenamentos de frequências (DFW) e transformações lineares da amplitude.

Considere que E corresponde a uma versão discretizada de um dos envelopes espectrais $@v$ a serem modelados. Como nosso sistema perceptual é sensível aos sons segundo uma escala logarítmica, cada envelope espectral de magnitude (harmônico ou estocástico) será ajustado a esta escala. Além do mais, para permitir o cômputo de espectros logarítmicos de amplitude, é conveniente que todos os valores estejam limitados inferiormente por uma constante positiva ϵ (definido pela Tabela 3.1) o qual delimita o 'chão' do log. Neste caso, tem-se que

$$E' = 20 \log 10(E + \epsilon) - 20 \log_{10}(\epsilon) \tag{3.7}$$

é sempre positivo. A transformação inversa a esta é definida como

$$E' = 10^{\frac{E+20 \log_{10}(\epsilon)}{20}} - \epsilon. \tag{3.8}$$

Em oposição à proposta de Goshtasby [75], no qual todo os parâmetros (a_k, μ_k, σ_k), $\forall k$ são estimados simultaneamente, a proposta a seguir trata cada função base $\psi(a_k, \mu_k, \sigma_k)$ separadamente. A ideia geral do algoritmo é construir sub-bandas E_k e inicializar os parâmetros das bases convenientemente, de acordo com a estratégia gulosa apresentada abaixo.

A soma de L^* funções paramétricas gerais é definida como

$$\hat{E}(f; a, \mu, \sigma) = \sum_{m=1}^{L^*} \psi(f; a_m, \mu_m, \sigma_m)$$

onde cada componente $\psi(f; a_m, \mu_m, \sigma_m) : \Re \longrightarrow \Re$ corresponde a uma função base contínua com amplitude a_m, frequência central μ_m e largura de banda σ_m, avaliadas na frequência $f = [0, R/2]$. O valor L^* é definido pela variável global da Tabela 3.1, o qual corresponde à ordem dos coeficientes.

Alguns exemplos de funções radiais de base são as conhecidas janelas Hann, Nuttall e Gaussiana. Estas classes particulares de funções baseadas na composição cossenoidal foram usadas neste trabalho [138], e são definidas como

$$\psi(f; a_m, \mu_m, \sigma_m) = a_m \sum_{m=1}^{L^*} d_m \cos\left(2\pi(m-1)\frac{f - \mu_m}{\sigma_m}\right) \tag{3.9}$$

se $f \in [\mu_m - \sigma_m, \mu_m + \sigma_m]$, e $\psi(f; a_m, \mu_m, \sigma_m) = 0$ caso contrário. Esta formulação nos permite selecionar um número extenso de janelas para o ajuste espectral, conforme se ajusta os parâmetros M e $\{d_m\}_{m=1}^{M}$, como pode ser vistos na Tabela 3.2.

Como exemplo, ao utilizar uma função base Gaussiana temos que a soma de L^* Gaussianas é definida como

$$\psi(f; a_m, \mu_m, \sigma_m) = a_m e^{\frac{-(f-\mu_m)^2}{2\sigma_m^2}} \tag{3.10}$$

onde a_m, μ_m e σ_m são a amplitude, posição central e desvio padrão da m-ésima componente

Tabela 3.2: *Alguns coeficientes d_m de cada função base.*

Função	d_0	d_1	d_3	d_4
Hann	0.5	−0.5		
Nuttall	0.35577	−0.48740	0.14423	−0.01260
Blackman-Harris	0.35875	−0.48829	0.14128	−0.01168
Blackman-Nuttall	0.36358	−0.48918	0.13660	−0.01064

Gaussiana, respectivamente.

Dois importantes requisitos para uma função geral de base, e que são válidos para os exemplos acima, são dados pela condição

$$\lim_{f \longrightarrow \mu_m \pm \sigma_m} \psi(f; a_m, \mu_m, \sigma_m) = 0,$$

de tal forma que a soma de todas as componentes seja contínua, e que cada componente sejam diferenciável em todos os pontos com relação aos parâmetros $[a, \mu, \sigma]$, a fim de que se possa aplicar os algoritmos iterativos de estimação. Ademais, se considera que as derivadas parciais em relação aos parâmetros a, μ e σ são calculadas em cada sub-banda $[\mu_m - \sigma_m, \mu_m + \sigma_m]$ e preenchida com zeros (*Zero Padding*) fora desta faixa de frequências.

O problema de modelagem baseado em funções de base radial é encontrar uma aproximação $\hat{E}(f; a, \mu, \sigma)$ de uma dada função discreta $E(f)$ (i.e. conhecida para um número finito de frequências) de tal modo que o erro de estimação seja mínimo. A seguir será apresentada uma heurística para obter uma aproximação inicial, e posteriormente serão discutidas técnicas de refinamento desta aproximação.

Inicialização: Ajuste Guloso

Alguns sistemas requerem que os parâmetros a serem estimados estejam dentro de intervalos fixos, especialmente no caso das posições centrais das funções base, as quais estão confinadas a um conjunto de bandas frequenciais. Considere o conjunto de L^* bandas de frequências centradas em $\{c_m\}_{m=1}^{L^*}$. No caso do espectro sonoro, é como que a escolha das frequências centrais $\{c_m\}$ levem em conta nosso sistema perceptual, por exemplo, ao se adotar um conjunto de frequências centradas nas bandas críticas da escala Bark. No entanto, visando dar mais flexibilidade quanto à escolha das bandas, o conjunto de frequências centrais usado neste trabalho estão uniformemente distribuídas de acordo com a escala MEL no caso da representação espectral harmônica, e com a escala linear convencional no caso da representação espectral da envoltória estocástica. O uso da escala linear se justifica dada a importância perceptual das altas frequências do espectro estocástico.

Dado um conjunto de frequências centrais, considera-se que a subdivisão do envelope original $E(f)$ em espectros $E_m(f)$ é definido pelo janelamento do sinal original E (envelope espectral de entrada) usando a função W_m centrada em c_m. É importante garantir que $\sum_m W_m \approx 1$, ou seja,

$$\sum_m E_m = \sum_m E \cdot W_m \approx E.$$

Uma abordagem alternativa à apresentada acima considera a subdivisão espectral dinâmica, na qual as componentes E_m são definidas equitativamente, de acordo com um critério de energia, i.e.

as frequências centrais c_m de cada E_m satisfazem

$$\int_{c_{m-1}}^{c_{m+1}} E_m = \frac{1}{L^*} \int_0^{2\pi} E.$$

Para os casos extremos do intervalo, i.e. $[0, c_1]$ e $[c_L, R]$, as energias são contadas pela metade. Cada janela W_m é definida pela frequência central c_m e fixada com largura de banda definida pelo intervalo $[c_{m-1}, c_{m+1}]$, e a amplitude a_m é estimada usando uma ajuste linear sobre E_m (estimador de mínimos quadrados). Tal abordagem tem suas particularidades, e dependendo da aplicação pode ser uma boa alternativa, uma vez que a mobilidade das frequências centrais caracterizam as regiões formânticas da voz[3].

Note que seria possível definir previamente os parâmetros de forma a serem completamente determinados por E_m. Tal abordagem é dita *gulosa*, e funciona da seguinte forma:

G 1: Num primeiro exemplo, considerando o pico global $p = (x, y)$ de cada sub-banda espectral E_m, é possível definir a_m^0 e μ_m^0 como y e x respectivamente, e atribuir a σ_m^0 um valor tal que a área sobre a curva de E_m seja a mesma da área delimitada por $\psi(f; a_m^0, \mu_m^0, \sigma_m^0)$.

G 2: Outra abordagem corresponde a tomar um conjunto fixo de bandas com larguras σ_m^0, e estimar o centroide como μ_m^0 em cada uma das sub-bandas E_m, encontrando as amplitudes ótimas a_m^0 usando o método de otimização por mínimos quadrados.

Finalmente, é a partir dos posicionamentos de μ^0 que se define a discretização da função contínua $@\upsilon$ em termos de E. Não obstante, os valores de μ^0 dependem essencialmente de E. Neste caso, uma versão temporária de E_t é gerada, a fim de obter as sub-bandas E_m e posteriormente os valores μ^0, a partir dos quais se pode definir o conjunto E com menos pontos, a ser usado na fase de otimização. Como todo o processo descrito acima (encontrar os pontos máximos μ^0) é de ordem linear $\Theta(|E_t|)$, uma super-amostragem de ordem considerável pode ser feita. Neste caso, o conjunto E_t é tomado a partir de E amostrado a $20 \times L^*$ em espectros harmônicos, e 1024 pontos em espectros estocásticos. Então, uma das duas abordagens acima é aplicada, a fim de se obter os conjuntos $(a_m^0, \mu_m^0, \sigma_m^0)$ para cada m.

Dado um fator de discretização D (normalmente entre 5 e 15), considera-se que os valores da discretização correspondem ao conjunto de amostras

$$E = \upsilon(\mu_{\text{ext}}^0), \tag{3.11}$$

onde

$$\mu_{\text{ext}}^0 = [\mu_1^0, \mu_{(1+\delta_1)}^0, \mu_{(1+2\delta_1)}^0, \ldots, \mu_{(1+D\delta_1)}^0, \mu_2^0, \mu_{(2+\delta_2)}^0, \mu_{(2+2\delta_2)}^0, \ldots, \mu_{(L^*-1+D\delta_{L^*-1})}^0, \mu_{L^*}^0]$$

é uma interpolação linear de μ^0 em sua correspondente escala de frequência, na qual

$$\delta_l = \frac{\mu_{l+1}^0 - \mu_l^0}{D+1}, \quad 1 \leq l < L^*.$$

Embora ambas as propostas gulosas acima apresentem um bom ajuste espectral, elas serão usadas como passo de inicialização do método iterativo apresentado abaixo, o qual é baseado no

[3] Tal conceito será mais amplamente abordado na Seção de Mapeamento de classes acústicas.

algoritmo de Marquardt, a fim de serem refinadas.

Algoritmo de Ajuste Iterativo

O próximo passo do método é refinar a componente $\psi(a_m, \mu_m, \sigma_m)$ para que se aproxime de sua respectiva sub-banda E_m usando o Algoritmo `ajuste_base`$_\psi$ (Alg. 3.9). A primeira linha do algoritmo realiza a discretização linear pareada, conforme definido na Eq. 3.11. As três linhas seguintes do algoritmo transformam a amplitude espectral em magnitudes logarítmicas, caso seja detectado que o envelope E corresponde ao envelope harmônico de amplitude ou ao envelope estocástico, informação esta devolvida pela função `tipo`. Entre as Linhas 6 e 9, o algoritmo realiza a conversão da escala linear de frequência para a escala MEL. Note que somente os espectros harmônicos são submetidos à esta transformação, dada a maior importância das baixas frequências na percepção harmônica (ver a Figura A.1 do Apêndice A).

Algoritmo 3.9 $(a^*, \mu^*, \sigma^*) \leftarrow$ `ajuste_base`$_\psi(\upsilon)$

1 Calcule E pela Eq. 3.11 usando υ

▷ *Normaliza o envelope E pela Eq. 3.7*

2 $E \leftarrow 20\log_{10}(E + \epsilon) - 20\log_{10}(\epsilon)$

3 **for** $m = 1 : L^*$

▷ *Define os janelamentos W_m*

4 $F_m \leftarrow F_c$

5 **if** tipo$(\upsilon) =$‘A’ **then**

▷ *Conversão para escala logarítmica MEL*

6 $F_m \leftarrow 1127.01048 \log\left(1 + \frac{F_m}{700}\right)$

7 **end if**

▷ *Encontra o centro e a largura de banda de cada janela W_k*

8 $(c_m, \text{largura}) \leftarrow \left(\frac{(m-1)(F_m)}{L^*-1}, \frac{F_m}{L^*-1}\right)$

9 $W_m \leftarrow \psi(1, c_m, \text{largura})$

▷ *Extrai a sub-banda E_k de E*

10 $E_m \leftarrow E \cdot W_m$

▷ *Obtém valores iniciais*

11 $(a_m^0, \mu_m^0, \sigma_m^0) \leftarrow (\max(E_m), \arg\max(E_m), \text{largura})$

▷ *Otimiza parâmetros da base*

12 $(a_m^*, \mu_m^*, \sigma_m^*) \leftarrow \text{otimizar\_base}(E_m, (a_m^0, \mu_m^0, \sigma_m^0), \epsilon)$

13 **end for**

14 **return**(a^*, μ^*, σ^*)

Como se pode observar, o algoritmo proposto utiliza a estratégia gulosa *G1* enumerada no tópico de inicialização visto anteriormente. No entanto, é possível facilmente adaptar a estratégia *G2* ao problema, embora isso acarrete uma perda de eficiência ocasionada pelo estimador de mínimos quadrados. Ambas as abordagens, no entanto, serão comparadas na Seção 4 de experimentos. A estratégia de subdivisão espectral dinâmica de bandas, na qual as bandas são definidas por um critério de distribuição uniforme de energias em cada banda, não foi considerada na implementação devido às menores taxas de acerto na experimentação. No entanto, tal estratégia pode ser facilmente incorporada ao algoritmo, e pode ser muito útil em aplicações nas quais o posicionamento das

frequências é mais importante que a representação espectral em si. Isto porque a estratégia não adota o critério estático, no qual a definição das sub-bandas utiliza regiões de frequências fixadas a priori, mas ao invés disso o critério de fixação recai sobre a energia em cada uma destas sub-bandas. Como resultado, espectros que concentram a maior porção energética em determinadas regiões formânticas tendem a ter um número maior de frequências próximas a estas regiões. Tal ideia será melhor explorada no módulo de Transformação apresentado na Seção 3.6.

Refinamento da Aproximação por Otimização

O ajuste da função dada ao modelo paramétrico

$$\hat{E}(f; a, \mu, \sigma) = \sum_{m=1}^{L^*} \psi(f; a_m, \mu_m, \sigma_m)$$

poderia em princípio ser obtido por qualquer método iterativo, como o método de Newton ou o gradiente descendente, que ajuste a soma de funções base a uma versão suavizada do espectro de $E(f)$. Entretanto, não é difícil encontrar situações práticas onde a parametrização obtida pelos métodos iterativos clássicos não varia suavemente entre segmentos sucessivos, mesmo que os envelopes espectrais respectivos evoluam lentamente no tempo. Este tipo de representação será então temporalmente instável. Muitas aplicações, como a síntese de fala por exemplo, exigem uma representação espectral temporalmente estável; neste caso em particular, os parâmetros (a_m, μ_m, σ_m) de cada função base correspondente a uma sub-banda espectral deveria ser uma função temporalmente estável. As propostas que seguem visam resolver exatamente este problema, a partir de restrições das posições de cada função base $\psi(a_m, \mu_m, \sigma_m)$.

Este problema é equivalente a resolver um sistema de equações não lineares de ordem $3L$ usando um estimador de mínimo quadrados não-linear. Além do mais, se é desejável encontrar uma representação que tenha uma interpretação geométrica para cada função base, é necessário supor que $a_m > 0$ para todo m. O primeiro passo do algoritmo é o isolamento das sub-bandas espectrais para o ajuste independente de cada uma das funções base usando métodos de otimização não-lineares.

Na modelagem clássica, dois métodos são tidos como bases na resolução de problema de mínimos quadrados não-lineares: o método Gauss-Newton no qual se encontra uma solução sub-ótima com um número reduzido de passos com custo elevado, ou o método dos gradientes descendentes no qual uma solução sub-ótima pode ser encontrada a partir de um número maior de passos computacionalmente mais leves. No entanto, o algortimo de Levenberg-Marquardt [147] é uma proposta que aproveita ambos os benefícios dos métodos de Gauss-Newton e do gradiente descendente. Dado um par de sequências $\{E_m(f), \psi(f; \alpha_m)\}$, o algoritmo de Levenberg-Marquardt otimiza os parâmetros α_m de modo que o erro quadrático

$$\sqrt{\sum_{\forall f} [E_m(f) - \psi(f; \alpha_m)]^2} \tag{3.12}$$

seja mínimo.

Cada função base é atualizada iterativamente conforme a equação

$$(a'_m, \mu'_m, \sigma'_m) = (a_m, \mu_m, \sigma_m) + (\delta_{a,m}, \delta_{\mu,m}, \delta_{\sigma,m})$$

a partir da atualização do parâmetro de variações δ de acordo com

$$[\mathbf{J}^T\mathbf{J}]\delta_m = \mathbf{J}^T[E_m(f) - \psi(f; a_m, \mu_m, \sigma_m)], \forall f \tag{3.13}$$

onde

$$J(a_m, \mu_m, \sigma_m) = \left[\frac{\partial\psi}{\partial a}\frac{\partial\psi}{\partial\mu}\frac{\partial\psi}{\partial\sigma}\right](a_m, \mu_m, \sigma_m) \tag{3.14}$$

é o Jacobiano de $\psi(a_m, \mu_m, \sigma_m)$. Esta atualização corresponde a ajustar otimamente o modelo linear de $\psi(a_m, \mu_m, \sigma_m)$ a E_m no espaço de representação (a, μ, σ).

O Jacobiano depende intrinsecamente da função base respectiva. Como exemplo, a função Gaussiana tem o Jacobiano definido como

$$\begin{aligned}
\frac{\partial\psi}{\partial a}(a_m, \mu_m, \sigma_m)(f) &= e^{\frac{-(f-\mu_m)^2}{2\sigma_m^2}} \\
\frac{\partial\psi}{\partial\mu}(a_m, \mu_m, \sigma_m)(f) &= \frac{a_m(f-\mu_m)}{\sigma_m^2}e^{\frac{-(n-\mu_m)^2}{2\sigma_m^2}} \\
\frac{\partial\psi}{\partial\sigma}(a_m, \mu_m, \sigma_m)(f) &= \frac{a_m(f-\mu_m)^2}{\sigma_m^3}e^{\frac{-(n-\mu_m)^2}{2\sigma_m^2}}
\end{aligned} \tag{3.15}$$

No caso das funções trigonométricas, as derivadas devem ser consideradas apenas no intervalo definido pela largura de banda. Neste caso, o Jacobiano é definido como

$$\begin{aligned}
\frac{\partial\psi}{\partial a}(a_m, \mu_m, \sigma_m)(f) &= \textstyle\sum_{m=1}^{L^*} d_m \cos\left(2\pi(m-1)\frac{f-\mu_m}{\sigma_m}\right) \\
\frac{\partial\psi}{\partial\mu}(a_m, \mu_m, \sigma_m)(f) &= a_m \textstyle\sum_{m=1}^{L^*} \frac{-2\pi d_m(m-1)}{\sigma_m} \sin\left(2\pi(m-1)\frac{f-\mu_m}{\sigma_m}\right) \\
\frac{\partial\psi}{\partial\sigma}(a_m, \mu_m, \sigma_m)(f) &= a_m \textstyle\sum_{m=1}^{L^*} \frac{2\pi d_m(m-1)(f-\mu_m)}{\sigma_m^2} \sin\left(2\pi(m-1)\frac{f-\mu_m}{\sigma_m}\right)
\end{aligned} \tag{3.16}$$

caso $f \in [\mu_m - \sigma_m, \mu_m + \sigma_m]$, e são todas definidas como 0 caso contrário.

Esta iteração carrega consigo um erro de aproximação, definido por

$$\mathcal{E}'_m = \left|E_m - \psi(a'_m, \mu'_m, \sigma'_m)\right|^2. \tag{3.17}$$

Um importante aspecto do algoritmo é o critério de parada pela busca da solução minimal, definida de acordo com as taxas de variação do erro absoluto. O processo de estimação finaliza quando o erro $\mathcal{E}_m$ converge, ou seja, dado um pequeno ϵ, a estimação é interrompida se $\Delta\mathcal{E}_m = |\mathcal{E}'_m - \mathcal{E}_m| \leq \epsilon$ é satisfeita. A contribuição de Levenberg resulta em uma versão ponderada onde se inclui um fator de aceleração de convergência não-negativo na estimativa dos coeficientes. Tal fator é ajustado a cada iteração de modo a balancear a busca pela solução minimal. O Algoritmo 3.10 sumariza estes passos.

Algoritmo 3.10 $(a_m^*, \mu_m^*, \sigma_m^*) \leftarrow$ `otimizar_base`$(E_m, (a^0, \mu^0, \sigma^0))$

```
1  E_m^0 ← |E_m − ψ(a_m^0, μ_m^0, σ_m^0)|^2
2  ΔE ← ∞
3  (a_m, μ_m, σ_m) ← (a_m^0, μ_m^0, σ_m^0)
4  while ΔE > ε
5      Calcule J(a_m, μ_m, σ_m) pela Eq. 3.15 (ou Eq. 3.16)
6      Calcule δ_m pela Eq. 3.13 usando J(a_m, μ_m, σ_m)
7      (a_m, μ_m, σ_m) ← (a_m, μ_m, σ_m) + δ_m
8      Calcule E_m pela Eq. 3.17
9      ΔE ← |E_m^0 − E_m|
10     E_m^0 ← E_m
11 end while
12 return(a_m, μ_m, σ_m)
```

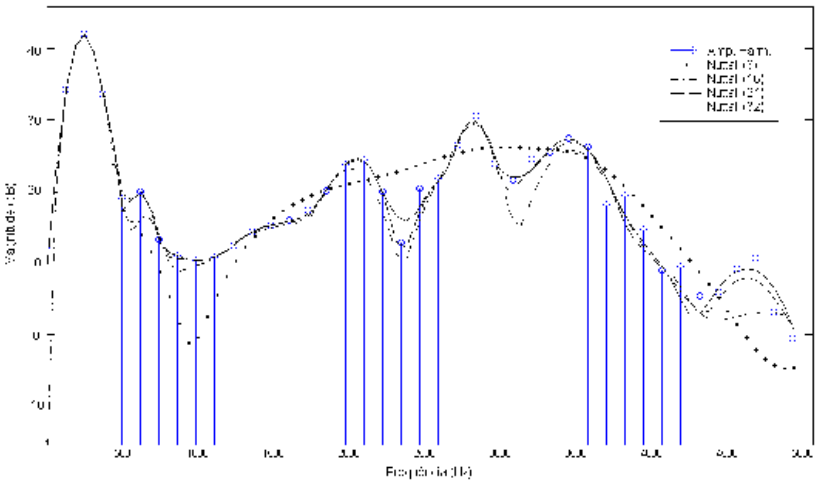

Figura 3.13: *Exemplo de estimação harmônica com 8, 16, 24 e 32 bases (Nuttall)*

.

A Figura 3.13 exibe um ajuste harmônico de amplitude de um segmento arbitrário de um sinal de voz, usando o método `otimizar_base` com a janela Nuttall e a escala linear de frequência. A figura exibe ajustes feitos usando entre 8 e 32 bases. Note que quanto mais bases, melhor o ajuste, eliminando assim, o problema de *overfitting*, como ocorre com LPC ou cepstrum. Note que a região mais grave foi melhor ajustada, por conta da utilização da escala MEL na fase de estimativa paramétrica.

A fim de representar o conjunto das L^* bases radiais, $3L^*$ parâmetros são usados, o que pode ser excessivamente caro dependendo da aplicação.

Restrição do Número de Parâmetros

A estratégia acima pode ser facilmente adaptada para produzir somas de funções base que se aproximem de $E(f)$, porém usando menos parâmetros para que seja economizado espaço. Esta representação alternativa consiste em manter os parâmetros μ e/ou σ fixos como se fossem valores globais, e então cada função base é representada por seus parâmetros livres, os quais são otimizados, usando a matriz Jacobiana da Equação 3.13 reduzida a estes parâmetros livres. Por exemplo, se μ

é fixo ($\mu_m = c_m = \texttt{center}(W_m)$) então o Jacobiano é

$$J(a_m, \sigma_m) = \left[\frac{\partial \psi}{\partial a} \frac{\partial \psi}{\partial \sigma}\right](a_m, \sigma_m),$$

tal que

$$\delta_m = (\delta_{a,m}, \delta_{\sigma,m})^T$$

é dado por

$$[J(a_m, \sigma_m)^T J(a_m, \sigma_m)]\delta_m = J(a_m, \sigma_m)^T [E_m - \psi(a_m, \mu_m, \sigma_m)].$$

As outras expressões precisam ser conformadas a esta redução de cardinalidade, por exemplo

$$\mathcal{E}_m = \left|E_m - \psi(a'_m, \sigma'_m)\right|^2.$$

Esta variante é chamada de $\texttt{ajuste\_base}_{\psi[a,\sigma]}$. A mesma pode ser analogamente definida para as variantes $\texttt{ajuste\_base}_{\psi[a,\mu]}$ e $\texttt{ajuste\_base}_{\psi[a]}$. Por motivos de comparação com estas variantes na fase experimental, o método original será denotado por $[a, \mu, \sigma]$ na seção experimental. Maiores detalhes destas abordagens podem ser encontradas em [137; 138].

Algoritmo 3.11 $E \leftarrow \texttt{desenvelopador}_{\psi[\text{tipo}]}([a], [\mu], [\sigma], F)$

for $m = 1 : L^*$

▷ *Encontra o centro e a largura de banda padrão, caso seja necessário*

$F_m \leftarrow F_c$

if tipo =‘A’ **then**

▷ *Conversão para escala logarítmica MEL*

$F_m \leftarrow 1127.01048 \log(1 + \frac{F_m}{700})$

$F \leftarrow 1127.01048 \log(1 + \frac{F}{700})$

end if

$(c_m, \text{largura}) \leftarrow \left(\frac{(m-1)(F_m)}{L^*-1}, \frac{F_m}{L^*-1}\right)$

if $\mu = \emptyset$ **then**

$\mu = c_m$

end if

if $\sigma = \emptyset$ **then**

$\sigma = \text{largura}$

end if

▷ *Retoma valores de cada base*

$E_m \leftarrow \psi(F; a_m, \mu_m, \sigma_m)$

end for

▷ *Restaura E*

$E \leftarrow \sum_{m=1}^{L^*} E_m$

▷ *Normalização Inversa do Envelope E pela Eq. 3.8*

if tipo =‘A’ OU tipo =‘S’ **then**

$E \leftarrow 10^{\frac{E+20\log_{10}(\epsilon)}{20}} - \epsilon$

end if

return(E)

Reciprocamente, pode-se ainda obter um envelope E avaliado nas frequências F (em escala li-

near), dado um subconjunto de coeficientes quantizadores do segmento, dentre $\{[a], [a, \sigma], [a, \mu], [a, \mu, \sigma]\}$, segundo as restrições adotadas pelo algoritmo. Neste caso, o algoritmo precisa ser identificado por ψ[tipo], onde tipo $= \{A, S\}$, correspondendo ao tipo do envelope em questão (A harmônico ou S estocástico).

O Algoritmo 3.11 é chamado de `desenvelopador`. Algumas particularidades deste algoritmo podem ser observadas na Linhas 5, onde o conjunto de frequências a serem amostradas são transformadas para a escala MEL, dependendo do tipo do envelope. Outro ponto importante é destacado entre as Linhas 7 e 13, nas quais se recompõem as posições centrais e/ou as larguras de bandas, caso um (ou ambos) destes argumentos não sejam passados como parâmetro à função de reconstrução. Tal recomposição é de extrema importância neste trabalho, uma vez que somente as amplitudes harmônicas e estocásticas serão utilizadas para caracterizar os segmentos de voz. Ao final do algoritmo, os contornos de energia são extraídos, tanto da parte harmônica, quanto estocástica do sinal.

A vantagem em se representar os envelopes de forma normalizada, é facilitar o processo de clusterização e mapeamento nos módulos que serão vistos posteriormente nas Seções 3.4 e 3.5.

Implementação do Quantizador

A função `quantizador` usado pelo módulo de Parametrização (ver começo da Seção, Alg. 3.5) recebe como argumentos os três envelopes modelados pelo método `envelopador`. Então o algoritmo monta um conjunto de vetores acústicos dimensionados, de acordo com um número definido pelo usuário. O método $\texttt{ajuste\_base}_{\psi[a]}$ foi selecionado neste trabalho usando um grande número de coeficientes, por apresentar um ajuste satisfatório e simples para cada componente harmônica e estocástica de cada segmento de voz (maiores justificativas podem ser encontradas na Seção 4).

Algoritmo 3.12 $(a_s, a_h, e_0) \leftarrow \texttt{quantizador}(\upsilon_S, \upsilon_A)$

1 $N_s \leftarrow$ dimensão(υ_S)
2 **for** $k = 1 : N_s$
3 **if** existe(υ_A^k) **then**
4 $(a_h^k) \leftarrow \texttt{ajuste\_base}_{\psi[a]}(\upsilon_A^k)$
5 $e_0 \leftarrow |a_h|$
6 $a_h \leftarrow \frac{a_h}{e_0}$
7 **end if**
8 $a_s^k \leftarrow \texttt{ajuste\_base}_{\psi[a]}(\upsilon_S^k)$
9 **end for**
10 **return**(a_s, a_h, e_0)

Além disso, os vetores $[a_h]$ e $[a_h]$ são estimados pelo método $\texttt{ajuste\_base}_{\psi[a]}$ usando frequências linearmente espaçadas na escala MEL. O mesmo não ocorre com a estimação estocástica, que utiliza a escala convencional (em Hertz) devido ao fato de que neste caso também se deve priorizar bandas espectrais de altas frequências. Os demais parâmetros dos envelopes, ou seja, μ e σ, são tidos como parâmetros globais do modelo.

Mediante as definições, a composição da função `quantizador` usada no módulo de Parametrização é imediata (Alg. 3.12).

Implementação do Sub_Amostrador

O algoritmo `sub_amostrador` realiza basicamente o processo inverso do `quantizador`, o qual também é usado pelo módulo de Parametrização em sua versão inversa (ver Alg. 3.6). No entanto, em vez de devolver o envelope espectral que será posteriormente re-amostrado nas frequências harmônicas (no caso dos envelopes harmônicos), opta-se por devolver diretamente estes valores pelo método `sub_amostrador`.

Sua implementação se assemelha muito à implementação do método `interpolador`, uma vez que ambos possuem a mesma estrutura de representação (por bases radiais), e está descrita no Algoritmo 3.13. A cada iteração do laço da Linha 3, o algoritmo reamostra os pulsos harmônicos de cada segmento vozeado, e juntamente, reamostra o espectro estocástico destes mesmos segmentos. Caso o segmento seja avaliado como não-vozeado, somente o espectro estocástico é recuperado.

Nas Linhas 7 e 10 a função de `desenvelopador` se encarrega de recuperar os valores de posição (μ_h, μ_S) e larguras de banda (σ_h, σ_S) para posterior subamostragem na Parametrização Inversa, de acordo com a escala usada em cada caso.

Algoritmo 3.13 $(S, A) \leftarrow$ `sub_amostrador`(V, F_0)

▷ *Restaura as variáveis a partir de* V
1 $(a_s, a_h) \leftarrow V$
2 $N_s \leftarrow$ dimensão(a_s)
3 **for** $k = 1 : N_s$
4 **if** $F_0(k) > 0$ **then**
▷ *Define as Frequências Harmônicas*
5 $L \leftarrow \left\lfloor \frac{F_c}{F_0(k)} \right\rfloor$
6 $F_h \leftarrow F_0(k)(0 : L)$
7 $A_k \leftarrow$ desenvelopador$_{\psi[\text{'}A\text{'}]}(a_h^k, \emptyset, \emptyset, F_h)$
8 **end if**
▷ *Define as Frequências Estocásticas*
9 $F_s \leftarrow R^{\frac{0:(\frac{N_w}{2}+1)}{N_w}}$
10 $S_k \leftarrow$ desenvelopador$_{\psi[\text{'}S\text{'}]}(a_S^k, \emptyset, \emptyset, F_s)$
11 **end for**
12 **return**(S, A)

Uma vez parametrizados os dados, o sistema segue para o estágio da Clusterização.

3.4 Estágio III: Clusterização dos Dados

A clusterização, também denominada análise de clusters, é uma das principais etapas em processos de análise de dados (ver Figura 3.14), realizando a classificação dos vetores acústicos e formando agrupamentos [100]. Existem várias abordagens para se agrupar ("*clusterizar*") pontos no espaço de características, das quais se destacam as técnicas de agrupamento supervisionado, tais como k-médias e k-histogramas, e não supervisionado, como a clusterização hierárquica [99]. Dado que uma ampla revisão de tais métodos pode ser encontrada no Capítulo 2, esta seção tem por objetivo apresentar apenas as propostas originais que contribuem para o desenvolvimento do sistema final de conversão de voz.

O **Módulo de Clusterização** implementa o algoritmo de classificação (**clusterizador**), o qual deve ser capaz de representar classes acústicas apropriadamente de maneira que seja possível deformar o espaço de características de um falante para outro num módulo posterior. Para que tal correspondência seja possível, o conjunto de dados deve, em primeiro lugar, ser caracterizado por um conjunto de vetores acústicos (clusters) estáveis e normalizados no espaço acústico.

Um cluster é dito ser **estável** quando um mesmo conjunto de dados é representado por somente uma configuração de classes acústicas. A estabilidade pressupõe que dados dois conjuntos de dados A e B, modelados estatisticamente e representados pelas suas respectivas funções densidade de probabilidade $\mathbb{F}_A$ e $\mathbb{F}_B$, será verdade que se $\mathbb{F}_A$ é aproximadamente igual a $\mathbb{F}_B$ segundo alguma métrica, então o conjunto das classes C^A e C^B também devem manter esta relação de proximidade de acordo com a métrica previamente estabelecida. Isto implica diretamente que, uma variação pequena na diferença entre os espaços $\mathbb{F}_A$ e $\mathbb{F}_B$ implica em uma pequena variação nos conjuntos de classes modelados.

No contexto deste trabalho, somente o subgrupo dos clusterizadores que realizam treinamento não-supervisionados serão considerados, uma vez que o conversor de voz proposto é **independente de texto** e **não-rotulado**. O objetivo do módulo de clusterização é aglutinar os parâmetros acústicos correspondentes a segmentos de voz semelhantes em torno de um ponto central, o qual caracteriza uma classe fonética artificial. A correspondência entre uma classe fonética artificial e um cluster (ou agrupamento) de vetores acústicos em torno de um ponto central no espaço de características é crucial neste sentido. A razão pela qual estas classes fonéticas são **artificiais** é justificada pela dificuldade, e em alguns casos até mesmo a impossibilidade, de rotular as classes de acordo com os fonemas reais da língua considerada.

A partir de cada agrupamento, os chamados **clusters**, o Módulo de Clusterização obtém um conjunto de parâmetros estatísticos (médias, matrizes de covariância, desvios padrões, etc...) que modelam cada classe fonética artificial. Considerando que o conjunto de todos os parâmetros acústicos podem ser vistos também como uma única (super-)classe acústica, o conjunto de parâmetros globais (do falante) também é obtido, correspondendo a momentos estatísticos ligados à prosódia (contornos de energia e pitch).

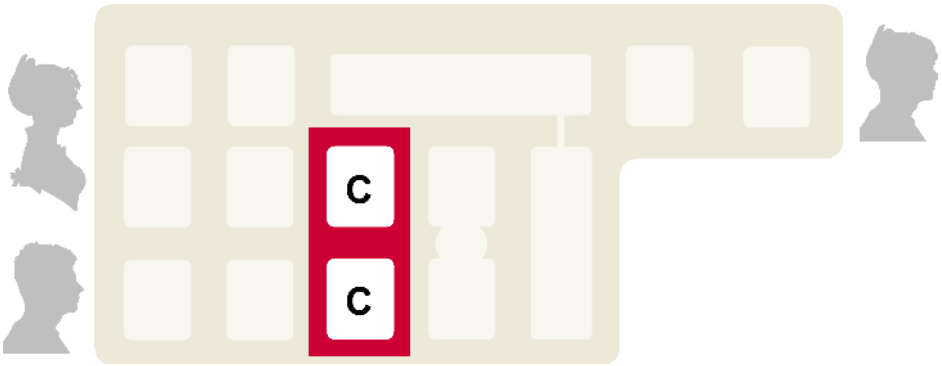

Figura 3.14: *Fase de Clusterização dos dados (C) em classes fonéticas artificiais*

Sendo assim, a seção a seguir se propõe a refinar um clusterizador estável e que modele com precisão as classes fonéticas artificiais no espaço de características.

Caracterização da Clusterização

Duas classes de clusterizadores são amplamente difundidos na literatura: os métodos baseados em redes cognitivas e os métodos estatísticos, conforme foi visto no Capítulo 2.

Os algoritmos cognitivos são compostos em sua maioria por redes neurais artificiais, sendo que a determinação dos classificadores depende de um aprendizado não supervisionado. Tais classificadores são obtidos por meio de redes de aprendizado competitivo que utilizam algum critério clássico de aprendizado (por exemplo, critérios de Kohonen [119]) para realizar a clusterização de um determinado conjunto de pontos. O problema principal desta abordagem é a falta de garantia de que os dados clusterizados atendam aos critérios de estabilidade estabelecidos, e de que haja uma correspondência fonética clara entre clusters e classes fonéticas artificiais, a fim de permitir a manipulação do espaço de características de modo consistente e perceptualmente significativo.

Igualmente, a modelagem estatística dos dados, como por exemplo a modelagem convencional usando GMM, requer que a inicialização a priori destes modelos esteja bem próxima do conjunto de classes fonéticas desejado, visto que o algoritmo clusterizador utiliza um critério de verossimilhança estabelecido pelas clássicas relações do teorema de Bayes (probabilidades condicionais) [267]. Tais condições iniciais devem ser rigorosamente controladas de modo que na clusterização final se permita o alinhamento consistente entre classes acústicas de falantes distintos. Infelizmente o modelo GMM também não atende a este critério de "estabilidade" na clusterização, ou seja, configurações iniciais (inicialização) distintas aplicadas a um mesmo conjunto de amostras podem gerar conjuntos distintos de classes. Ademais, os modelos de conversão de voz que utilizam este arcabouço exigem que os dados de entrada já estejam previamente alinhados, a fim de realizar transformações lineares ponderadas a partir da clusterização conjunta de classes pareadas. Estimações convencionais por meio de modelagens estatísticas tendem a representar os dados por meio de funções suavizadas sobre o espaço de características, as quais tendem a causar um efeito conhecido como suavização excessiva (*over smoothing*) na fase de transformação [268]. Por outro lado, outras abordagens que buscam representações mais rígidas tendem a causar descontinuidades na fase de síntese, dado o ajuste excessivo aos dados (*overfitting*) [155].

Um requisito importante do clusterizador adotado por este trabalho corresponde à normalização geométrica de cada classe acústica, a fim de eliminar incompatibilidades entre classes fonéticas artificiais (CFAs) de línguas distintas, problema este que será melhor abordado na Seção 3.5. O problema fundamental de tal abordagem é que não se conhece precisamente a função de mapeamento que deveria deformar apropriadamente uma classe de geometria complexa a uma outra dentro do espaço de características. Sendo assim, uma ideia para transpor uma classe acústica de um espaço para outro consiste em decompor cada classe por um número finito de subconjuntos normalizados em variância e magnitude, a fim de se realizar posteriormente somente translações de centroides. Sendo assim, este trabalho propõe um modelo (preliminar) de clusterização estável que se ajuste precisamente aos dados e ao mesmo tempo seja flexivelmente representado por classes fonéticas artificiais normalizadas. O Módulo de Clusterização definido neste trabalho consiste na chamada de três funções básicas as quais serão desenvolvidas ao longo da seção, e é definido a seguir.

Algoritmo 3.14 $(C, G) \leftarrow$ `Clusterização`(Ψ)

▷ *Agrupamento em classes acústicas*
1 $\tilde{C} \leftarrow$ agrupador_acústico(Ψ)

▷ *Agrupamento em classes fonéticas artificiais*
2 $[C, G] \leftarrow$ mapa_fonético$(\tilde{C}, \Psi)$

▷ *Toma momentos estatísticos locais de* Ψ
3 $C \leftarrow$ momentos_estatísticos_locais(Ψ, C)

▷ *Toma momentos estatísticos globais de* Ψ
4 $G \leftarrow$ momentos_estatísticos_globais(Ψ, G)
5 **return**(C, G)

O clusterizador se encarrega de compor um conjunto de pequenos clusters de segmentos semelhantes, os quais serão posteriormente modelados em termos de classes fonéticas artificiais representadas pelo mapa fonético do corpus. Sendo assim, o primeiro passo do Módulo de Clusterização é definir agrupamentos verossimilhantes de vetores acústicos derivados de Ψ e devidamente quantizados a fim de serem analisados a posteriori.

3.4.1 Clusterização k-Verossímil

O Clusterizador k-Verossímil é um método que toma um conjunto de vetores no espaço L-dimensional e os subdivide em classes acústicas $\tilde{C}$ com k elementos cada. Cada uma destas classes acústicas é composta por um conjunto de k vetores relativos aos segmentos de voz de um corpus qualquer, e que juntamente minimizam a média do desvio padrão entre os mesmos. Tipicamente, o valor de k é definido como $\lfloor \log(N_s) \rfloor$, onde N_s é o número de elementos do conjunto de treinamento Ψ. Assim sendo, é necessário definir para cada segmento de voz um vetor quantizado a fim de dar prosseguimento ao procedimento de clusterização.

O primeiro passo do método consiste em tomar cada vetor devolvido pelo módulo de quantização e criar uma representação que carregue consigo um conteúdo perceptualmente relevante sobre o segmento de voz respectivo inserido dentro da sentença. Neste sentido, cada vetor de magnitude harmônica é combinado com seu sucessor e antecessor imediato dentro do sinal original, multiplicados por um fator de ponderação com 50% de contribuição. Por simplificação, tanto a parte relativa às fases harmônicas quanto a parte estocástica do sinal não são consideradas neste contexto do trabalho. Deste modo, para cada segmento de voz k, o respectivo vetor v_k é definido como

$$v_k = \left[0.5\Psi_A^k + 0.25\Psi_A^{k-1} + 0.25\Psi_A^{k+1}\right],$$

onde Ψ_A^k corresponde às amplitudes harmônicas refentes ao k-ésimo segmento de voz.

Um modo alternativo de caracterizar acusticamente tais vetores utiliza uma versão similar à representação Mel-Cepstral (MFCC). Note que o vetor v_k já se encontra devidamente amostrado na escala MEL, devido ao fato que as amplitudes harmônicas A foram definidas pelo módulo quantizador nesta escala. Deste modo, para cada segmento de voz k, o respectivo vetor v_k pode também ser definido como

$$v_k^{\text{mfcc-like}} = \text{dct}\left[0.5\Psi_A^k + 0.25\Psi_A^{k-1} + 0.25\Psi_A^{k+1}\right], \tag{3.18}$$

onde dct corresponde à transformada do cosseno discreta.

Este algoritmo pode ser visto como um passo de pré-processamento para o clusterizador fonético que, embora dispensável, é bastante útil para filtrar o sinal de entrada, eliminando segmentos de voz com baixa frequência de ocorrência, bem como para obter uma base de dados mais representativa do conjunto de segmentos de voz. Outro aspecto importante é o ganho em velocidade de processamento computacional na fase de mapeamento, visto que os dados são compactados.

O clusterizador utiliza três estruturas de dados auxiliares. A primeira é uma tabela de distância entre os vetores, a segunda uma tabela de índices que relaciona cada vetor m com outros k vetores de modo a minimizar a distância euclidiana entre os mesmos, e a última estrutura contém a norma do desvio padrão de cada componente destes vetores. Especificamente, a partir de cada par de vetores (v_m, v_n) é montada a tabela de distâncias euclidianas $E(m, n)$, de modo que

$$E(m, n) = |v_m - v_n|.$$

Então, para cada linha m desta matriz é selecionado os índices dos N_k menores valores de cada linha m de E e armazenados em uma outra tabela de índices J de modo que

$$J(m, \cdot) = \arg\min_{N_k}\{E(m, \cdot)\}.$$

Finalmente, é construída uma estrutura de seleção de classes K na qual se armazena a norma euclidiana do desvio padrão dos vetores armazenados em $J(m, \cdot)$, ou seja,

$$K(m) = |\text{std}\{v_k\}|, \ \forall k \in J(m, \cdot).$$

Em sua parte final, o algoritmo de clusterização entra em um laço iterativo no qual se toma a estrutura de seleção K e se busca o índice m^* de menor valor absoluto. A partir desse índice são agregados os respectivos vetores indexados por $I = J(m^*, \cdot)$ em uma única classe. A partir de então, os valores de K indexados por I são removidos da busca. O processo se repete até que não haja mais nenhum valor em K.

O Algoritmo 3.15 implementa o algoritmo clusterizador como descrito. A Linha 5 do algoritmo quantiza os vetores acústicos a serem agrupados numa abordagem semelhante aos MFCCs. A partir daí, na Linha 10 é montado o mapa de distâncias euclidianas, do qual são tomados (na Linha 12) os índices dos N_k menores valores e armazenados em $J(m, \cdot)$ para cada m-ésima linha de E. Dados estes N_k índices, o algoritmo calcula o desvio padrão do conjunto de vetores v_k indexados por $k \in J(m, \cdot)$, e toma a norma destes desvios (Linha 13).

A segunda parte do algoritmo, a qual se inicia na Linha 15, consiste em encontrar os agrupamentos com normas de desvios padrões minimais usando uma estratégia gulosa. Entre as Linhas 18 e 21 o algoritmo seleciona o índice do mínimo global de K, a partir do qual se obtém o conjunto índice $I = J(m^*, \cdot)$ que compõe um conjunto k-verossímil. Finalmente, I é agregado ao conjunto de saída $\tilde{C}$, e posteriormente, todos estes elementos são excluídos de K, até que K esteja completamente vazio, ou que a condição de parada da Linha 16 seja alcançada.

A partir da clusterização em pequenos grupos k-verossímeis, uma filtragem opcional pode ser realizada a fim de eliminar conjuntos de vetores com desvio padrão relativamente grande. Um fator de filtragem de classes verossímeis (η) é utilizado a fim de definir um limiar para que os conjuntos de vetores com desvio padrão interno grande sejam eliminados. Esta filtragem elimina todos os

subconjuntos de $I \subseteq K$ indexados pelo vetor originário m^* nos quais se verifica que

$$K(m^*) > \mathbb{E}\{K\} + \eta \, \mathrm{std}\{K\},$$

onde $\mathbb{E}\{K\}$ e $\mathrm{std}\{K\}$ são a média e o desvio padrão de K, respectivamente (vide Linha 16). Tipicamente, o valor de η corresponde a uma pequena variação acima do valor médio do conjunto K, e neste trabalho foi utilizado o valor padrão $\eta = 0.25$.

Algoritmo 3.15 $\tilde{C} \leftarrow$ `agrupador_acústico`(Ψ)

▷ *Define o número de elementos de cada classe*
1 $N_s \leftarrow \mathrm{dimensão}(\Psi)$
2 $N_k \leftarrow \lfloor \log(N_s) \rfloor$
3 **for** $k = 1 : N_s$
4 **if** $F_0(k) > 0$ **then**

▷ *Toma o Vetor v_k*
5 $v_k \leftarrow \left[0.5\Psi_A^k + 0.25\Psi_A^{k-1} + 0.25\Psi_A^{k+1}\right]$
6 **end if**
7 **end for**

▷ *Gera o mapa de distância*
8 **for** $m = 1 : N_s$
9 **for** $n = 1 : N_s$
10 $E(m,n) \leftarrow |v_m - v_n|$
11 **end if**

▷ *Gera o mapa de índices dos N_k elementos mínimos*
12 $J(m,\cdot) = \arg\min_{N_k}\{E(m,\cdot)\}$

▷ *Toma a norma do desvio padrão do conjunto indexado em $K(m,\cdot)$*
13 $K(m) \leftarrow |\mathrm{std}\{v_k\}|, \forall k \in J(m,\cdot)$
14 **end if**
15 $\tilde{C} \leftarrow \emptyset$
16 $stop \leftarrow \mathbb{E}\{K\} + \eta \mathrm{std}\{K\}$

▷ *Encontra os agrupamentos via método guloso*
17 **while** $K \neq \emptyset$ & $\min\{K\} \leq stop$
18 $m^* = \arg\min\{K\}$
19 $I \leftarrow J(m^*,\cdot)$
20 $\tilde{C} \leftarrow \tilde{C} \cup I$
21 $K \leftarrow K \ominus K(I)$
22 **end while**
23 **return**$(\tilde{C})$

Opcionalmente, o uso de clusterizadores hierárquicos classificam automaticamente os dados por critérios de distância mínima em redor dos centroides de cada classe. Neste caso, o critério de poda da árvore de cluster é crucial para uma boa clusterização.

Na sequência, o clusterizador também devolve ao sistema um conjunto de informações globais a serem usadas posteriormente na fase de conversão.

Dado os conjuntos $\tilde{C}$, o módulo de Clusterização passa a uma fase de agrupamento fonético destes conjuntos.

3.4.2 Modelagem do Mapa Fonético Artificial

Para entender um pouco sobre o mapeamento fonético, é necessário entender algumas bases de linguística relacionadas ao posicionamento dos primeiros formantes em sons vozeados.

Conforme exposto no Capítulo 2, é sabido que o principal fator discriminante entre uma vogal e outra é a disposição de suas respectivas regiões formânticas. No contexto deste trabalho, um maior interesse na conversão de voz é direcionado aos segmentos vozeados, e em especial, às vogais, devido a uma maior concentração de energia espectral na parte harmônica nestes tipos de segmentos.

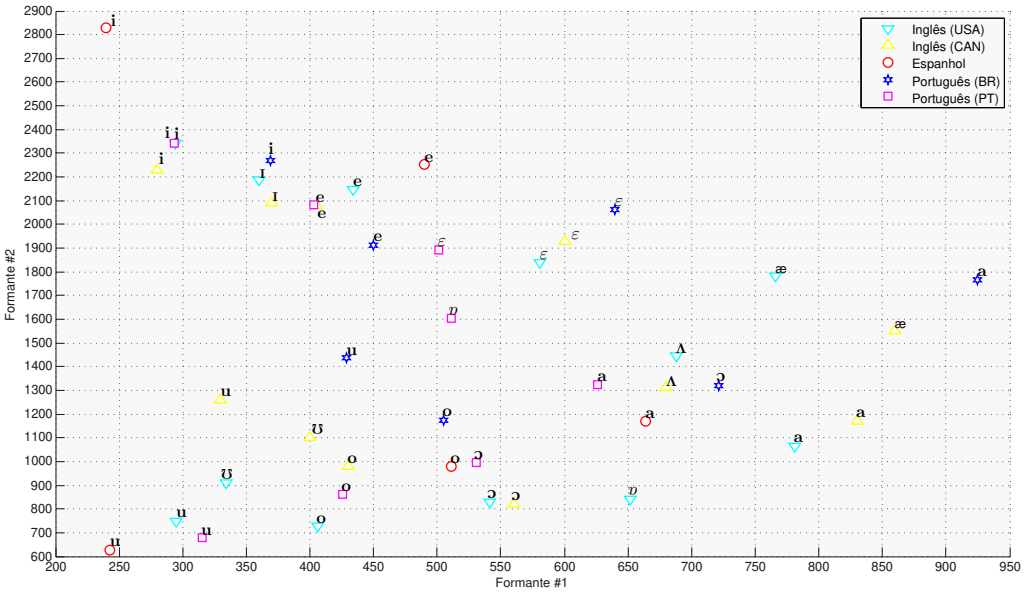

Figura 3.15: *Posições do primeiro e segundo formantes (em Hertz) para diversas vogais.*

Sabe-se que os dois primeiros formantes de um espectro vocálico são suficientes para determinar o fonema em questão. Uma importante tarefa corresponde a encontrar a região do espectro de magnitude do sinal de voz responsável por identificar fonemas da respectiva língua, ou seja, que contenha os dois primeiros formantes de uma sentença pronunciada.

Observe na Figura 3.15 as regiões onde se localizam as posições centrais (médias) do primeiro e segundo formantes, os **centros formânticos**, das principais vogais das línguas inglesa (canadense [213] e americana [86]), portuguesa (brasileira [158] e europeia [35]) e espanhola [198]. Tal mapa é conhecido como **mapa fonético** (médio) da língua.

Embora haja um consenso geral sobre a localização aproximada dos primeiros e segundos formantes para cada língua, é importante perceber que as regiões nas quais os primeiros e segundos formantes estão localizados varia não só de língua para língua, mas de indivíduo para indivíduo. Este fato se explica em função de diferenças entre homens, crianças, mulheres e idosos na constituição física do aparelho fonador.

A fim de estabelecer um ponto de partida para a clusterização de segmentos de voz em classes artificiais, o uso de um polígono com dimensões extrapoladas, o qual seguramente envolve os centros formânticos dentro do mapa fonético de um indivíduo qualquer, é usado para obtenção de um mapa fonético expandido, o qual posteriormente será filtrado, a fim de completar a clusterização do conjunto de vetores acústicos em classes fonéticas artificiais.

O polígono que envolve os centros formânticos, em geral, corresponde a uma região convexa trapezoidal ou triangular dependendo da língua correspondente. Considerando-se que o trapézio melhor modela o mapa fonético, o problema passa a ser encontrar quais as dimensões do polígono que limitam a região formântica. Encontrar um polígono que se ajuste precisamente ao mapa fonético de um específico locutor, exige que primeiramente, seja estimado um mapa fonético extrapolado, a partir do qual se provê uma largura individualizada com base no corpus. A fim de cobrir todas as frequências (na escala MEL) associadas às línguas portuguesa, espanhola e inglesa, o trapézio fonético extrapolado possui extremidades em $(300, 600)$, $(300, 2000)$, $(2000, 1100)$ e $(2000, 1600)$, valores estes amostrados na escala mel. Um valor tão grande para as abcissas dos extremos direitos do trapézio foi selecionado a fim de se agrupar as classes acústicas das vogais e eliminar aquelas que não se representam devidamente por causa da contaminação com consoantes, por exemplo.

O módulo de clusterização tem por objetivo clusterizar os dados acústicos baseado em informações fonéticas catalogadas, e por isso tal modelagem necessita de um método de composição do mapa fonético de cada subconjunto k-verossímil obtido anteriormente. O primeiro passo para obtenção de tal mapa consiste em realizar uma quantização do conjunto em termos fonéticos. Como tal mapa é discreto, é necessário amostrar o mapa fonético em M_{f1} por M_{f2} pontos na escala MEL. A Figura 3.16 exemplifica uma amostragem do mapa fonético em 9×15 bandas de frequência.

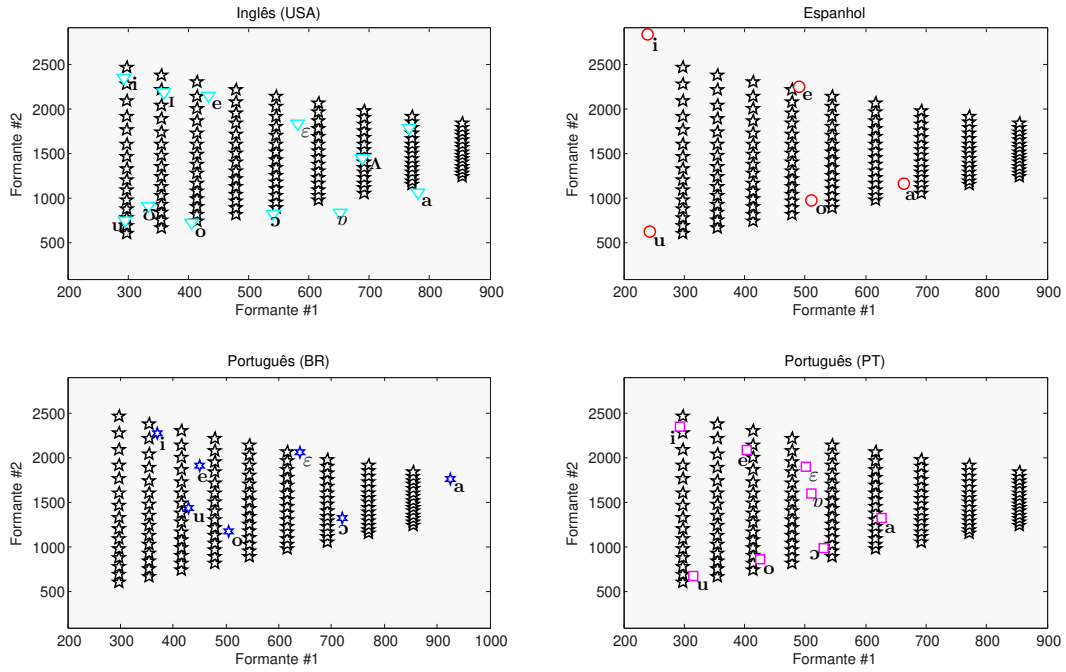

Figura 3.16: *Quantização do mapa fonético em* 9×15 *bandas.*

A partir de cada ponto posicionado em (f_1, f_2), com $f_1 \leq f_2$, o método gera um filtro digital passa-bandas composta por duas componentes triangulares, cada uma centrada em uma região formântica respectiva f_1 e f_2. A fim de evitar dois filtros demasiadamente próximos ou muito distantes um do outro, uma distância mínima de 300 mels e máxima de 1800 mels entre dois filtros consecutivos é estabelecida, a fim de controlar o posicionamento dos dois filtros sobre as regiões formânticas F_1 e F_2.

Quanto à largura de banda de cada um destes filtros, um cuidado especial deve ser tomado. Para larguras de banda demasiadamente pequenas, o método clusterizador corre o risco de atribuir

às variáveis (f_1, f_2) valores relativos a uma mesma região formântica. Por outro lado, valores muito grandes permitem que um mesmo filtro f_1 agregue duas regiões formânticas distintas. Por esta razão, as larguras de banda de um filtro triangular são definidas estaticamente como $B_f = 120$ mels, tal como definido na Tabela 3.1 de variáveis globais. A taxa de amostragem adotada para quantização do mapa fonético é $M_f = M_{f1} = M_{f2} = 35$.

O banco de filtros extrapolado, após ter sido aplicado sobre a base de dados, precisa ser refinado e, só então, pode ser utilizado na classificação fonética. A função `filtros_fonéticos`(X, Y_1, Y_2) descrita pelo Alg. 3.16 implementa tal banco de filtros, dadas as coordenadas que definem os quatro pontos do trapézio fonético envolvente. Os vértices do trapézio são definidos em sentido horário a partir do topo esquerdo como $\langle X(\min), Y_1(\max)\rangle$, $\langle X(\max), Y_2(\max)\rangle$, $\langle X(\max), Y_2(\min)\rangle$ e $\langle X(\min), Y_1(\min)\rangle$, respectivamente.

Algoritmo 3.16 *banco* $\leftarrow$ `filtros_fonéticos`(X, Y_1, Y_2)

▷ *Re-amostra a região formântica*
1 $F(\min) \leftarrow \min\{X(\min), Y_1(\min), Y_2(\min)\}$
2 $F(\max) \leftarrow \max\{X(\max), Y_1(\max), Y_2(\max)\}$
3 $F \leftarrow \left(\frac{0:N_w}{N_w}\right)(F(\max) - F(\min)) + F(\min)$

▷ *Encontra valores relativos ao primeiro e segundo formantes*
4 $x \leftarrow \left(\frac{0:M_f-1}{M_f-1}\right)(X(\max) - X(\min)) + X(\min)$
5 $y_u \leftarrow \left(\frac{0:M_f-1}{M_f-1}\right)(Y_1(\max) - Y_1(\min)) + Y_1(\min)$
6 $y_d \leftarrow \left(\frac{0:M_f-1}{M_f-1}\right)(Y_2(\max) - Y_2(\min)) + Y_2(\min)$

▷ *Determina cada par de formantes*
7 $banco \leftarrow \emptyset$
8 **for** $f_1 = 1 : M_f$
9 **for** $f_2 = 1 : M_f$
10 **if** $f_1 + 300 \leq f_2$ & $f_1 + 1800 \geq f_2$ **then**
11 $y \leftarrow \frac{f_1-1}{M_f}[y_u(f_2) - y_d(f_2)] + y_d(f_2)$
12 $filtro.f_1 = x(f_1),\ filtro.f_2 = y$
13 $filtro.v = \text{triang}(F, x(f_1), B_f) + \text{triang}(F, y, B_f)$
14 $banco \leftarrow banco \cup filtro$
15 **end if**
16 **end for**
17 **end for**
18 **return**(*banco*)

As Linhas 4, 5 e 6 do Algoritmo 3.16 devolvem interpolações lineares dos intervalos X, Y_1 e Y_2, quantizados em M_f pontos. Entre as Linha 7 e 17 são agregadas iterativamente ao banco de filtros somas de duas janelas triangulares de larguras B_f, uma centrada em $x(f_1)$ e outra centrada em y. Cada um desses filtros está amostrado dentro do intervalo F, o qual se estende desde a menor até a maior frequência nos pontos extremos do polígono, como definido nas Linha 1, 2 e 3 do algoritmo.

Dado o conjunto de bancos de filtros devolvido pelo Algoritmo 3.16 é possível obter um mapa discreto de distribuição formântica para cada conjunto k-verossímil disponível até então. Cada mapa é obtido a partir da soma dos vetores de amplitude harmônica devidamente filtrados. Deste modo, é necessário criar um vetor de amplitudes que corresponda univocamente a um conjunto k-verossímil.

Dado um conjunto $I \in \tilde{C}$ o vetor

$$V_I = \mathbb{E}\{\Psi_A^k\}, \forall k \in I$$

corresponde ao vetor médio de todos os coeficientes de amplitude harmônica dos segmentos de voz k, isto é, Ψ_A^k. Como tais valores estão linearmente espaçados na escala MEL dentro do intervalo de $X(\min)$ a $X(\max)$, é necessário fazer uma re-amostragem com N_w (ver Tabela 3.1) pontos do vetor V_I dentro do intervalo correspondente ao usado no banco de filtros, ou seja, o intervalo

$$F = \left(\frac{0 : N_w}{N_w}\right)(X(\max) - X(\min)) + X(\min).$$

Após a preparação dos vetores quantizados V_I de cada classe I, é necessário definir a função que aplica a filtragem propriamente dita. A função `filtragem_formântica`(V_I, X, Y_1, Y_2) toma o conjunto de vetores V_I e os processa pelos filtros pertencentes ao banco de filtros devolvido pelo Algoritmo 3.16, referenciado pela estrutura *banco*, conforme é mostrado no Algoritmo 3.17.

Como se pode observar, a Linha 3 do algoritmo seguinte realiza uma filtragem (multiplicação escalar) de cada vetor V_I por um filtro armazenado no campo 'v' da estrutura $filtro$, e posteriormente se toma a norma euclidiana do vetor filtrado.

Espera-se que cada um desses mapas de distribuição *distr* contenha informações fonéticas associadas ao vetor V_I. Entretanto, é evidente que cada vetor V_I contenha informações relativas ao conteúdo harmônico inerente ao indivíduo emissor, e que ademais, tal conteúdo varie pouco de um fonema para outro. Por esta razão, uma filtragem destes mapas acústicos auxilia bastante na tarefa de detecção fonética. Tal filtragem leva em consideração o fato de que o valor médio de todos os mapas de distribuição é um bom discriminante em termos fonéticos.

Algoritmo 3.17 $distr \leftarrow$ `filtragem_formântica`(V_I, X, Y_1, Y_2)

▷ *Definindo o banco de filtros*
1 $banco \leftarrow$ filtros_fonéticos(X, Y_1, Y_2)

▷ *Realiza a filtragem*
2 **foreach** $filtro \in banco$
3 $\quad distr(filtro.f1, filtro.f2) \leftarrow |V_I \cdot filtro.v|$
4 **end foreach**
5 **return**$(distr)$

A filtragem para obtenção dos mapas acústicos toma o mapa médio denotado por $\mathbb{E}\{distr\}$, e para cada mapa $distr_I$ associado à classe k-verossímil I, o mapa de distribuição normalizado de I é definido como

$$distr_I^{norm} = \text{mean}_{5\times5}\left(\mathbb{E}\{distr\} \cdot \text{std}\{distr\} \cdot (distr_I - \mathbb{E}\{distr\})^3\right),$$

onde a função $\text{mean}_{5\times5}$ calcula o valor médio local dentro de uma janela 5×5 centrada em cada ponto do mapa de distribuição. Nos casos das extremidades do mapa, os valores cujos índices excederam a fronteira são desconsiderados no cálculo da média.

A Figura 3.17 exibe exemplos de mapas de distribuição das cinco vogais dominantes da língua hispânica, pronunciada por um indivíduo de nacionalidade espanhola.

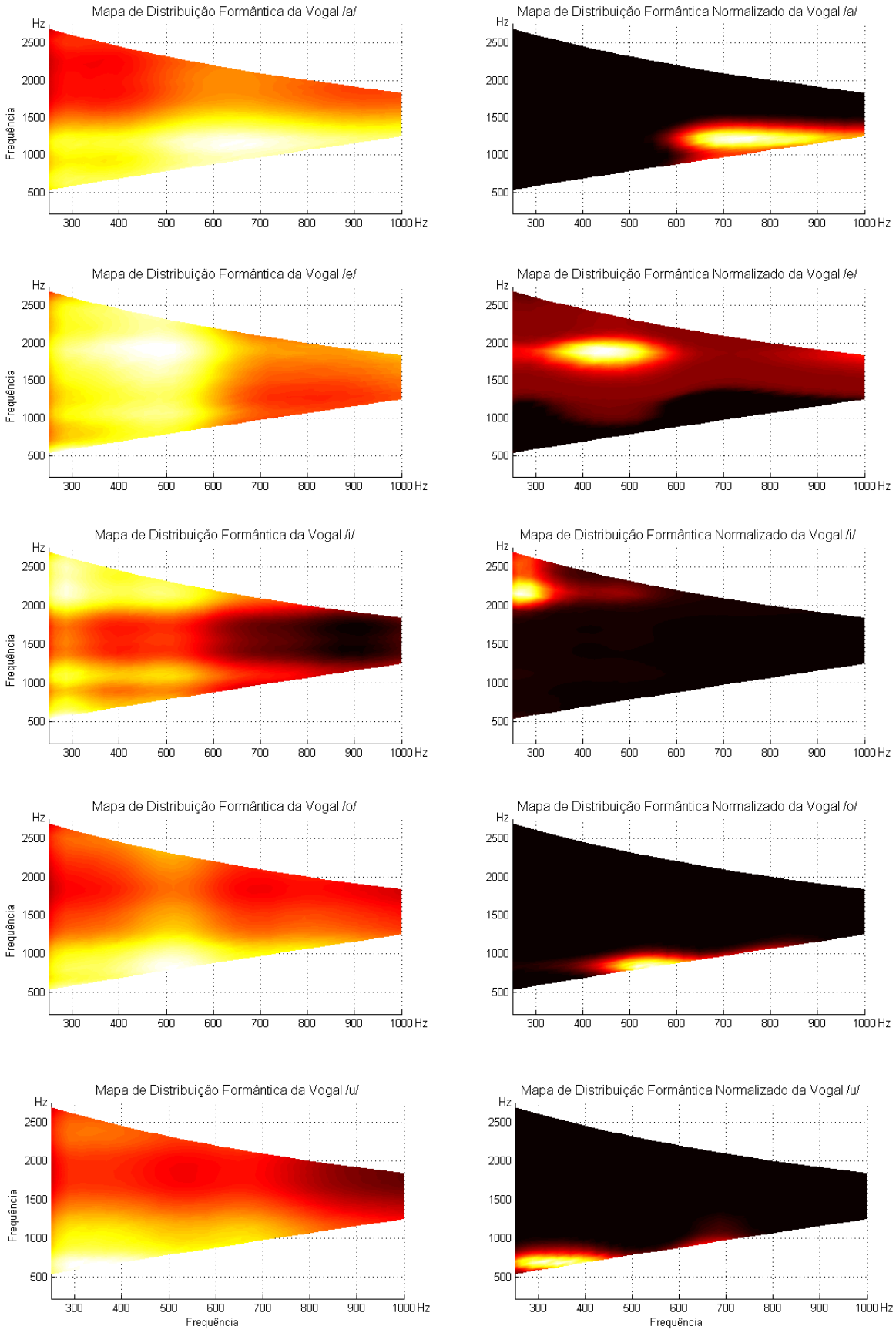

Figura 3.17: *Distribuição formântica (esq) e sua versão normalizada (dir) para cada vogal espanhola em um fragmento de voz.*

A filtragem elimina parte da componente harmônica associada à identidade sonora do indivíduo, realçando a região formântica relevante do segmento de voz como pode ser observado na Fig. 3.17. Observe o quanto estão separadas uma vogal da outra dentro do trapézio fonético definido entre 250 e 1000 Hz. Outro aspecto interessante de se perceber é a presença das componentes harmônicas associadas ao timbre do indivíduo, estando estas localizadas na região onde os formantes F_1 e F_2 são mais graves.

Para cada conjunto k-verossímil, é de extrema importância decidir a localização do seu ponto discreto correspondente dentro do mapa fonético, isto é, a determinação do par de frequências formantes dominante do segmento de voz. Existem diversas maneiras de se estimar o ponto dominante dentro do mapa fonético, como por exemplo, usando centros de massa do mapa bidimensional. No entanto, tal tarefa se aproxima muito da tarefa de reconhecimento de fala e não é de interesse deste trabalho a rotulação fonética. Assim, no contexto é realizado uma detecção de pico máximo global tomando-se os índices (f_1, f_2) do maior valor absoluto dentro do respectivo mapa de distribuição formântica.

Algoritmo 3.18 $distr \leftarrow$ `distribuição_formântica`(V_I)

▷ *Define a dimensão do mapa formântico extrapolado*
1 $(X(\min), X(\max)) \leftarrow (300, 2000)$
2 $(Y_1(\min), Y_1(\max), Y_2(\min), Y_2(\max)) \leftarrow (600, 2000, 1100, 1600)$
3 $distr_{super} \leftarrow$ filtragem_formântica(V_I, X, Y_1, Y_2)
4 $X \leftarrow$ ajuste_trapezoidal$(distr_{super})$
5 $distr \leftarrow$ filtragem_formântica(V_I, X, Y_1, Y_2)
6 **return**$(distr)$

O Algoritmo 3.18 implementa a função `distribuição_formântica`(V_I) que estima o mapa de distribuição fonético, dado um conjunto de vetores quantizados V_I de entrada. A primeira linha do mesmo define as dimensões extrapoladas do trapézio fonético. O motivo pelo qual as duas filtragens formânticas foram realizadas nas Linhas 3 e 5 se explica pelo uso da função `ajuste_trapezoidal`, implementado pelo Algoritmo 3.19, que toma um mapa de distribuição formântico com dimensões extrapoladas e encontra qual a dimensão X que melhor se ajusta aos fonemas vocálicos. Não é necessário re-estimar os valores de Y_1 e Y_2, uma vez que não oferecem problemas tão significativos na modelagem fonética quanto a primeira região formântica. Deste modo, todo as classes acústicas são novamente submetidas ao método de filtragem formântica a fim de obter o mapa de distribuição formântica final.

Para se ter uma ideia de como funciona este método, a Figura 3.18 exibe o mapa de distribuição extrapolado (super-amostrado) tomado de exemplos reais de quatro corpora de locutores de ambos os sexos (dois espanhóis e dois ingleses), os quais registraram cerca de 10 minutos de sentenças arbitrárias oriundas da base de dados rotuladas TC-STAR (vide Seção 4). Todos os mapas foram obtidos da Linha 2 do Alg. 3.18. Observe que em todos os casos existe um aglomerado de pontos imediatamente à esquerda que se assemelha a uma forma geométrica triangular (destacado em vermelho na figura).

Um corpus a ser alinhado deve possuir uma quantidade de dados de tal forma que seja possível compor pelo menos esta componente conexa. O uso de todas as componentes conexas também se apresenta como uma opção alternativa de modelagem, que no entanto não será levada em consideração neste trabalho devido à ausência de significado fonético em termos vocálicos, o que complica

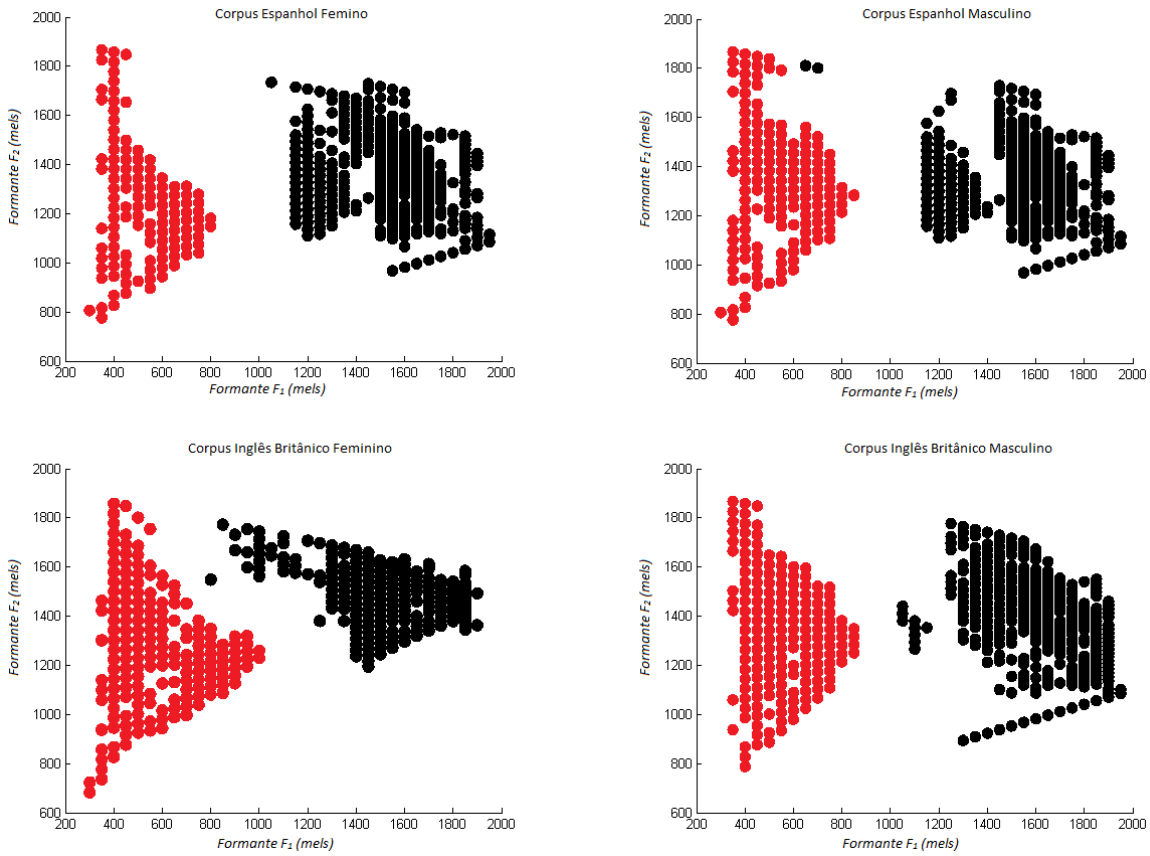

Figura 3.18: *Mapas de distribuição formântica extrapolados.*

consideravelmente o processo de mapeamento na etapa seguinte. Caso se tenha uma quantidade suficientemente grande de dados, uma abordagem alternativa à apresentada no algoritmo proposto consiste em tomar as classes definidas pela primeira estrutura conexa à esquerda do mapa, e descartar assim, as outras partes. No caso deste trabalho, toma-se tal componente conexa mais à esquerda e a partir de seus pontos extremos da direita e esquerda, devolve as dimensões X, as quais são usadas para re-estimar o mapa fonético dentro do polígono trapezoidal assim redefinido. O uso de ferramentas básicas da morfologia matemática são de grande serventia neste caso, as quais podem ser acessadas no Apêndice B.3.

Modelagem do Mapa Formântico

O mapa acústico discreto $distr_{super}$ devolvido pelo algoritmo `filtragem_formântica` pode ser descrito como um conjunto $A \subseteq \mathbb{Z}^2$ de modo que um ponto $x = \langle f_1, f_2 \rangle$ pertence a A se, e somente se, $distr_{super}(f_1, f_2)$ está definido. Deste modo, representando um mapa acústico como um subconjunto de $\mathbb{Z}$, é possível compor um operador morfológico capaz de segmentar apropriadamente a região de interesse dentro deste mapa, a saber, a estrutura formântica trapezoidal que modela o mapa fonético do respectivo falante.

Nota-se que uma componente conexa corresponde a um conjunto de pontos vizinhos que se aglutinam segundo uma relação de vizinhança definida por um elemento estruturante B. Neste trabalho, define-se $B = \{\langle -1, -1\rangle, \langle -1, 0\rangle, \langle -1, 1\rangle, \langle 0, -1\rangle, \langle 0, 0\rangle, \langle 0, 1\rangle, \langle 1, -1\rangle, \langle 1, 0\rangle, \langle 1, 1\rangle\}$ como elemento estruturante básico para segmentação das componentes conexas. A fim de eliminar o ruído

inicial do mapa A, realiza-se uma filtragem por fechamento de modo que $A' \leftarrow \phi_B(A)$. A partir de então, dois pontos vizinhos $(a_1, a_2) \in A$, isto é, pares de pontos para os quais existe um ponto $b \in B$ tal que $a_1 = a_2 + b$, são considerados de uma mesma componente conexa. Um procedimento de dilatações consecutivas anexa iterativamente pontos vizinhos a uma mesma componente usando o operador de dilatação, o qual é descrito entre as Linhas 4 e 7 do Algoritmo 3.19. Ao final, o algoritmo toma a primeira componente conexa mais significativa à esquerda, e extrai seus pontos extremos a fim de ser usada pelo Alg. 3.18.

Algoritmo 3.19 $X \leftarrow$ `ajuste_trapezoidal`$(distr_{super})$

▷ *Monta conjunto A*
1 $A \leftarrow \bigcup_{(f_1,f_2)\leftarrow \arg\max\{distr_{super}\}} \langle f_1, f_2 \rangle$

▷ *Realiza filtragem morfológica*
2 $B \leftarrow \{\langle -1,-1\rangle, \langle -1,0\rangle, \langle -1,1\rangle, \langle 0,-1\rangle, \langle 0,0\rangle, \langle 0,1\rangle, \langle 1,-1\rangle, \langle 1,0\rangle, \langle 1,1\rangle\}$
3 $A \leftarrow \phi_B(A)$

▷ *Encontra a primeira componente conexa por aglutinação*
4 $C \leftarrow \{p \in A \mid p$ seja um ponto que pertença à primeira componente$\}$
5 **while** $\exists p \in C \wedge \exists b \in B$ tal que $p + b \in A$
6 $\quad C \leftarrow \delta_B(C) \cap A$
7 **end while**

▷ *Define dimensões do trapézio envolvente*
8 $(esq, dir) \leftarrow$ extremos_laterais(C)
9 $(X(\min), X(\max)) \leftarrow (esq, dir)$
10 **return**(X)

A função `extremos_laterais` da Linha 8 estima os extremos direito e esquerdo da componente conexa a partir das quais são definidas as dimensões de X. De acordo com a Linha 4, é necessário indicar um ponto que pertença à estrutura fonética mais à esquerda. Uma forma de escolher tal ponto é selecionar o ponto mais próximo do centro do mapa formântico geral dentro do conjunto A, ou seja, $x \in A$ tal que $|x - \langle 500, 1000\rangle|$ seja mínimo, e que tenha pelo menos um vizinho definido em A de acordo com o elemento estruturante B. Como a filtragem de fechamento foi utilizada, é de se esperar que os pontos do conjunto estejam fortemente conectados. Ao final da clusterização, um passo opcional verifica a dimensão da componente conexa, a fim de verificar se o conjunto é um candidato provável a ser o trapézio envolvente das vogais no mapa fonético artificial, ou seja, $X(\max) > 500$ mels, por exemplo. Como os corpora usados no trabalho são bem controlados, tal passo de verificação é desconsiderado. Finalmente, redefinidas as dimensões do trapézio fonético, o Alg. 3.18 recalcula o mapa de distribuição formântica, a fim de relacionar cada ponto a uma CFA. Como cada conjunto k-verossímel é associado a um único ponto discreto (f_1, f_2) no mapa fonético final, o agrupamento de todos os elementos pertinentes a estes conjuntos k-verossímeis associados corresponde a uma classe fonética artificial.

Dadas todas essas considerações, o Algoritmo 3.20 toma um conjunto de agrupamentos k-verossímeis $\tilde{C}$ e os vetores quantizados Ψ devolvidos pelo Módulo de Parametrização (Seção 3.3) e devolve o mapa acústico C bidimensional, que contém os campos 'I' e 'P' correspondendo respectivamente aos índices dos elementos (segmentos) contidos em cada classe e à cardinalidade de cada um destes conjuntos.

Algoritmo 3.20 $(C, G) \leftarrow$ `mapa_fonético`$(\tilde{C}, \Psi)$

▷ *Monta o conjunto de vetores e mapa formântico de cada classe I*

1 **foreach** $I \in \tilde{C}$

2 $\quad V_I \leftarrow \mathbb{E}\left\{\bigcup_{\forall k \in I}\{\Psi_A^k\}\right\}$

3 $\quad distr_I \leftarrow$ distribuição_formântica(V_I)

4 **end foreach**

▷ *Guarda discriminantes lineares a serem usados posteriormente na conversão*

5 $G.discr.\mu \leftarrow \mathbb{E}\{distr\}$

6 $G.discr.\sigma \leftarrow \text{std}\{distr\}$

▷ *Inicializa atributos de cada classe fonética artificial*

7 $C(1 : M_{f1}, 1 : M_{f2}).\{I, P, mu, Sigma\} \leftarrow \{\emptyset, 0, \emptyset, \emptyset\}$

8 **foreach** $I \in \tilde{C}$

▷ *Normaliza mapa formântico*

9 $\quad distr_I^{norm} \leftarrow \text{mean}_{5\times5}\left(G.discr.\mu \cdot G.discr.\sigma \cdot (distr_I - G.discr.\mu)^3\right)$

▷ *Classificação fonética*

10 $\quad (f_1, f_2) \leftarrow \arg\max\{distr_I^{norm}\}$

▷ *Atualiza os atributos de índice e cardinalidade de cada classe*

11 $\quad C(f_1, f_2).I \leftarrow C(f_1, f_2).I \ \cup \ I$

12 $\quad C(f_1, f_2).P \leftarrow C(f_1, f_2).P + 1$

13 **end foreach**

14 **return**(C, G)

Este algoritmo toma os mapas de distribuição formântica de cada vetor de amplitudes média entra as Linhas 1 e 4, e atribui o mapa médio e o desvio padrão desse conjunto de mapas à estrutura global G (Linhas 4 e 5). A partir da Linha 7 o algoritmo monta a estrutura que contém as classes fonéticas artificiais do respectivo falante, usando um módulo de seleção por máximo local sobre o mapa de distribuição normalizado (Linha 9). Conforme pode ser observado, o mapa normalizado pondera as componentes de frequência que estão acima da média pela significância de cada componente, representada pela multiplicação entre o mapa global médio e seu respectivo desvio. Ao final, um filtro de média associado a uma janela quadrada 5×5 é usado para suavizar o mapa normalizado, do qual é extraído o índice de máximo global (Linha 11). A partir de cada ponto do mapa, que representa uma classe fonética artificial, o algoritmo obtém as médias e as matrizes de covariância dos elementos que compõe cada conjunto, a serem utilizadas posteriormente na fase de transformação (ver Algoritmo 3.21).

Os discriminantes lineares de distribuição formântica $G.discr.\mu$ e $G.discr.\sigma$ servem como limiares para posterior definição de chaves de seleção de uma classe, conforme é apresentado na Seção 3.6. A representação das classes fonéticas artificiais pela estrutura topológica apresentada (o mapa fonético) se faz apropriada em outros tipos de tarefas de análise e manipulação de conteúdo fonético, tais como síntese concatenativa ou reconhecimento fonético.

Modelagem dos Parâmetros Locais

A modelagem em classes fonéticas artificiais exige que um conjunto razoavelmente grande seja utilizado para povoar todo o mapa formântico. No entanto, o armazenamento da base de dados,

além de exigir um vasto espaço computacional, compromete significativamente o desempenho computacional.

Objetivando uma transformação eficiente, o sistema armazena momentos estatísticos dos dados clusterizados, a fim de serem utilizados posteriormente na fase de transformação. O algoritmo que modela estatisticamente cada CFA toma o agrupamento de conjuntos k-verossímeis que compõem uma particular CFA, e em seguida estima os vetores médios, bem como as matrizes de covariância. Segue a função `momentos_estatísticos_locais` implementada pelo Algoritmo 3.21, onde todas as médias e desvios laterais são calculadas para cada conjunto de vetores de treinamento que compõem cada classe.

Algoritmo 3.21 $C \leftarrow$ `momentos_estatísticos_locais`(C, Ψ)

1 **foreach** $(f_1, f_2) \in \arg C$

▷ *Obtém médias μ e matrizes de covariância Σ de cada classe*

2 $A_I \leftarrow \bigcup_{[\forall k \in C(f_1,f_2).I]} \{\Psi_A^k\}$

▷ *Toma momentos estatísticos de cada $v \in A_I$*

3 $C(f_1, f_2).\mu \leftarrow \mathbb{E}\{A_I\}$

4 $C(f_1, f_2).\Sigma \leftarrow \mathrm{cov}\{A_I\}$

5 **end foreach**

6 **return**(C, G)

A grande vantagem em se utilizar tais momentos estatísticos é a flexibilidade na caracterização de cada CFA, bem como o fato de possibilitar que diversos métodos, como a transformação linear, possam realizar a conversão de voz sem a necessidade de recorrer à base de dados inteira. Num âmbito mais geral, os parâmetros de controle global da sentença pronunciada, em especial a frequência fundamental média e a energia total do sinal, também podem ser modelas estatisticamente, conforme segue.

Modelagem dos Parâmetros Globais

A conversão de voz desenvolvida neste trabalho requer que o conjunto de parâmetros acústicos também seja modelado estatisticamente a nível global, isto é, usando todo o conjunto de parâmetros Ψ. Os controles globais são categorizados em duas classes distintas: **controles prosódicos**, isto é, as frequências fundamentais e energias de cada segmento harmônico de voz, uma vez que as amplitudes harmônicas respectivas se encontram normalizadas de modo que a soma seja igual a 1, e os **controles espectrais**, representados pelas médias e matrizes de covariância de todos as amplitudes harmônicas e estocásticas de cada corpus. O motivo pelo qual as fases não foram modeladas nos controles espectrais se justifica pela forte correlação entre segmentos de voz consecutivos, de modo que qualquer flutuação, por menor que seja, degrada significativamente o sinal modificado.

Embora o sistema converta localmente o conteúdo espectral do sinal, uma conversão espectral em sentido amplo é uma das alternativas a ser comparada na fase de avaliação perceptual. Vale salientar que as amplitudes estão devidamente quantizadas na escala logarítmica e amostrada na escala MEL com 48 coeficientes. Sendo assim, nenhuma conversão de escala deve ser aplicada na fase de modelagem global do sinal, assim como no caso da frequência fundamental.

Note que pelo fato de os contornos de pitch e energia serem sequências de valores atômicos (unidimensionais), a matriz de covariância dos mesmos corresponde a uma variável atômica, a variância

estatística. Nestes casos particulares, uma proposta de estimativa do desvio padrão assimétrico é apresentada visando dar maior precisão à modelagem estatística. O desvio padrão assimétrico de um conjunto $X = \{x_1, x_2, \ldots, x_n\}$ é composto por duas componentes de desvio padrão, uma associada aos valores superiores à media e outra aos valores inferiores, e são definidas como

$$\begin{aligned} \sigma_{\mathrm{u}}(X) &= \sqrt{\tfrac{1}{n-1}\textstyle\sum_{i=1}^{n} mask_{\mathrm{u}}(x_i, \bar{x}) \cdot (x_i - \bar{x})^2}, \\ \sigma_{\mathrm{d}}(X) &= \sqrt{\tfrac{1}{n-1}\textstyle\sum_{i=1}^{n} mask_{\mathrm{d}}(x_i, \bar{x}) \cdot (x_i - \bar{x})^2}, \end{aligned} \tag{3.19}$$

onde $\bar{x}$ é a média amostral de X de modo que cada componente $mask_{\mathrm{u}}^{[d]}(x_i, \bar{x})$ da máscara $mask_{\mathrm{u}}(x_i, \bar{x})$ é definida como

$$mask_{\mathrm{u}}^{[d]}(x_i, \bar{x}) = \begin{cases} 1, \text{ se } x_i \geq \bar{x} \\ 0, \text{ caso contrário} \end{cases}.$$

Analogamente, cada componente $mask_{\mathrm{d}}^{[d]}(x_i, \bar{x})$ da máscara $mask_{\mathrm{d}}(x_i, \bar{x})$ é definida como

$$mask_{\mathrm{d}}^{[d]}(x_i, \bar{x}) = \begin{cases} 1, \text{ se } x_i < \bar{x} \\ 0, \text{ caso contrário} \end{cases}.$$

Uma vez que a soma de $mask_{\mathrm{u}}^{[d]}(x_i, \bar{x})$ e $mask_{\mathrm{d}}^{[d]}(x_i, \bar{x})$ é um vetor unitário, o desvio padrão simétrico é determinado dinamicamente na fase de transformação linear de acordo com a configuração de ambas as máscaras, conforme será exibido nas seções subsequentes.

Algoritmo 3.22 $G \leftarrow$ `momentos_estatísticos_globais`(Ψ, G)

▷ *Organiza vetores de parâmetros globais*
$N_s \leftarrow$ dimensão(Ψ)
$(F_0, E_0, A, S) \leftarrow (\emptyset, \emptyset, \emptyset, \emptyset)$
for $k = 1 : N_s$
  $S \leftarrow S \cup \Psi_S^k$
  if $\Psi_{F_0}^k > 0$ **then**
    ▷ *Gera vetores contínuos para fácil modelagem*
    $F_0 \leftarrow F_0 \cup \Psi_{F_0}^k$
    $E_0 \leftarrow E_0 \cup \Psi_{E_0}^k$
    $A \leftarrow A \cup \Psi_A^k$
  end if
end for
▷ *Estima momentos estatísticos da prosódia*
$(G.F_0.\sigma_{\mathrm{u}}, G.F_0.\sigma_{\mathrm{d}}) \leftarrow$ desvios_laterais(F_0)
$G.F_0.\mu \leftarrow \mathbb{E}\{F_0\} + \frac{G.F_0.\sigma_{\mathrm{u}} - G.F_0.\sigma_{\mathrm{d}}}{2}$
$(G.E_0.\sigma_{\mathrm{u}}, G.E_0.\sigma_{\mathrm{d}}) \leftarrow$ desvios_laterais(E_0)
$G.E_0.\mu \leftarrow \mathbb{E}\{E_0\} + \frac{G.E_0.\sigma_{\mathrm{u}} - G.E_0.\sigma_{\mathrm{d}}}{2}$
▷ *Estima momentos estatísticos das componentes espectrais*
$G.A.\mu \leftarrow \mathbb{E}\{A\}$
$G.A.\Sigma \leftarrow \mathrm{cov}\{A\}$
$G.S.\mu \leftarrow \mathbb{E}\{S\}$
$G.S.\Sigma \leftarrow \mathrm{cov}\{S\}$
return(G)

É comum, ao final da estimativa dos desvios laterais, deslocar a média estimada na direção do maior desvio lateral encontrado, σ_{u} ou σ_{d}, a fim de melhor representar o conjunto modelado.

Desta maneira, a função `momentos_estatísticos_globais` segue implementada pelo Algoritmo 3.22, onde todas as médias e desvios laterais são extraídas dos vetores de treinamento armazenados na estrutura Ψ. Entre as Linhas 1 e 10 deste algoritmo é realizada a montagem dos conjuntos de frequências fundamentais F_0 e energias dos segmentos de voz E_0, para cada vetor avaliado como harmônico pela Linha 4. No caso das amplitudes harmônicas A e estocásticas S, a matriz de covariância é uma abordagem mais apropriada para a modelagem, por se tratar de vetores multidimensionais. Como se pode observar, o algoritmo utiliza a função `desvios_laterais` para modelar os contornos de energia e pitch, a qual segue implementada pelo Alg. 3.23

Algoritmo 3.23 $(\sigma_{\mathrm{u}}, \sigma_{\mathrm{d}}) \leftarrow$ `desvios_laterais`(X)

1 Calcula $\sigma_{\mathrm{u}}(X)$ e $\sigma_{\mathrm{d}}(X)$ pela Eq. 3.19
2 $\sigma_{\mathrm{u}} \leftarrow \sigma_{\mathrm{u}}(X)$
3 $\sigma_{\mathrm{d}} \leftarrow \sigma_{\mathrm{d}}(X)$
4 **return**$(\sigma_{\mathrm{u}}, \sigma_{\mathrm{d}})$

A Figura 3.19 exibe um par de exemplos nos quais se ajustou os desvios padrões laterais a um conjunto composto por uma série de frequências fundamentais amostradas na escala linear de frequências Hertz (gráfico superior), amostradas na escala MEL (gráfico central), e na escala logarítmica $20\log_{10}$ (gráfico inferior). Em todos os casos, um melhor ajuste é observado utilizando a abordagem proposta, constituindo-se o pior caso, quando ambos os desvios laterais são iguais, ou seja, equivalem ao desvio padrão clássico. Embora o valor médio deslocado não corresponda à média amostral real do conjunto, conforme apresenta a Linha 12 e 13 do Alg. 3.22, tal valor pode ser considerado uma espécie de média ponderada que se ajusta ao conjunto em questão.

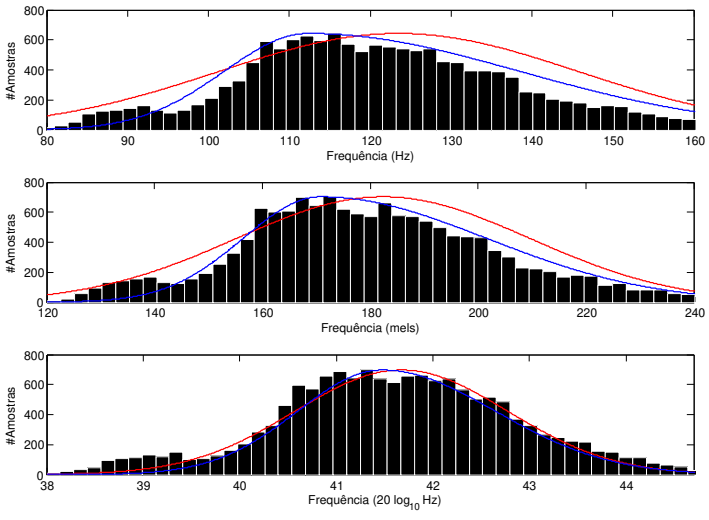

Figura 3.19: *Exemplo de modelagem por desvio padrão lateral.*

Ao finalizar o Módulo de Clusterização, o sistema passa ao Módulo de Mapeamento, no qual ambos os corpora não necessariamente paralelos são alinhados.

3.5 Estágio IV: Alinhamento dos Corpora Não-Paralelos

O estágio de alinhamento de classes acústicas de corpora não-paralelos compreende o *Módulo de Mapeamento* de sistema de conversão de voz, conforme é mostrado na Figura 3.20.

Um par de corpora é dito *paralelo* se os conjuntos de sentenças pronunciadas entre ambos os corpora estão alinhadas em pares, de acordo com um texto guia. Especificamente neste trabalho, o alinhamento de corpus realizado não precisa ser necessariamente paralelo, ou seja, o processo de alinhamento de classes acústicas devolvidas pelo módulo anterior não impõe nenhuma restrição fonética ao nível da sentença e nem sequer requer qualquer hipótese a respeito do alinhamento entre as sentenças de entrada. A fim de derivar um modelo de mapeamento acústico entre classes fonéticas artificiais devolvidas no módulo anterior, alguns temas serão discutidos para o desenvolvimento de um método de alinhamento sem corpora paralelos.

Dada a complexidade dos sinais de voz, a identidade sonora de um indivíduo está muito atrelada às características acústicas locais (fonéticas) da fala. A exigência de se alinhar os corpora de dois falantes se justifica pelo fato de que a transposição de características locais entre um par de falantes deve ser realizada localmente, ou seja, é improvável que haja uma transformação global que converta quaisquer vetores acústicos de um falante para outro, sem que esta modele localmente subconjuntos destes vetores, ou seja, sem que leve em consideração o contexto. Na prática, um corpus contém informações fonéticas relativas a segmentos de voz, recortadas e agrupadas em classes fonéticas artificiais, e informações prosódicas e espectrais globais, como momentos estatísticos de pitch e energia.

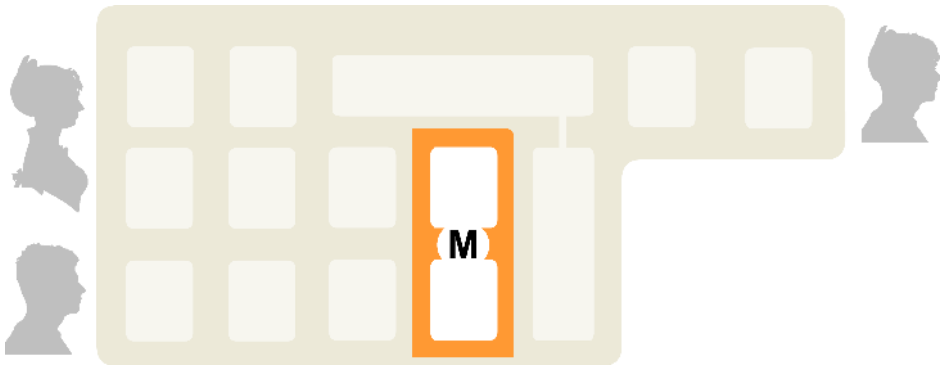

Figura 3.20: *Fase de Mapeamento de Classes Fonéticas (**M**)*

A literatura apresenta uma variedade de métodos para alinhamento de classes fonéticas, especialmente em conversão de voz convencional. Existem basicamente duas abordagens clássicas de alinhamento em conversão de voz: (1) alinhamento entre sentenças, segmento à segmento; e (2) alinhamento classe a classe do espaço acústico modelado.

No primeiro caso, o alinhamento entre sentenças pode ser feito a partir de dados rotulados e alinhados no tempo usando algoritmos clássicos como o *Dynamic Time Warping* [96]. Tal abordagem, que é dependente do texto (*text-dependent approach*), está fora do escopo do trabalho, uma vez que exige com que os corpora sejam paralelos.

Alternativamente, existe um método [49] que realiza sucessivos alinhamentos entre segmentos de dois conjuntos extensos de sentenças de treinamento, a fim de obter um alinhamento ótimo. Tais alinhamentos usam a métrica da distância euclidiana espectral, aplicada às duas direções de conversão origem→destino e vice-versa. Então o método clusterizador (o algoritmo EM [159]) do

Modelo de Misturas Gaussianas (GMM) é utilizado a fim de classificar as classes acústicas pareadas, isto é, com o alinhamento entre as classes já definido a priori. Dadas as classes conjuntas (compostas por coeficientes pareados), o método utiliza uma função de transformação espectral proveniente do arcabouço GMM, a fim de converter os vetores acústicos do falante origem para o destino. O processo é repetido até que o alinhamento se estabilize.

A partir deste alinhamento dos dados outras abordagens, apresentadas na Seção 2.4, também poderiam ser aplicadas para definir o mapeamento, como por exemplo o uso de mapas de Kohonen, classificadores Bayesianos ou qualquer outro tipo de clusterizador. No caso dos mapas neurais, o treinamento usa os vetores origem como dados de entrada e os vetores alinhados do falante destino como função objetivo.

Como mais uma alternativa, o alinhamento pode ser feito entre classes fonéticas já previamente modeladas de ambos os corpora distintos [176]. Uma maior flexibilidade desta classe de métodos é fundamentada nas seguintes observações:

- A base de dados pode ser obtida livremente, sem a necessidade de rotulação dos dados, ou pressupostos de distribuição equilibrada de fonemas entre corpora paralelos.
- O processamento classe a classe é mais eficiente em termos computacionais do que a abordagem segmento a segmento, uma vez que o conjunto de classe representante é consideravelmente menor que o conjunto total de dados.
- Corpus paralelos podem ser facilmente integrados ao sistema de conversão, estendendo o conceito de conversão de voz de '*um-para-um*' (origem-destino) para '*um-para-muitos*', dada a não-obrigatoriedade de alinhamentos custosos entre pares de corpora.
- O alinhamento classe a classe nos oferece uma interpretação perceptual em termos de linguagem, que pode ser posteriormente usado em sistemas de síntese e reconhecimento estatístico de fala.

Por estes motivos, o alinhamento classe a classe foi escolhido para o desenvolvimento do sistema de conversão de voz.

Considerações Iniciais

Em termos gerais, o problema de alinhamento entre os corpora dos falantes origem e destino usando a abordagem classe a classe consiste em mapear o espaço d-dimensional de características acústicas (as classes fonéticas artificiais) de um falante para outro, de modo a ser deformado em uma fase posterior. Neste caso, é necessário que exista uma correspondência entre cada classe fonética de um corpus para outro ao nível suficientemente local, de modo que a transposição dos espaços de características seja possível. Para tanto, o alinhamento precisa criar pontos de correspondência entre classes do espaço de características acústicas, dos quais se pode derivar funções de transferência do sinal de um falante para o outro.

Por um lado, um alinhamento entre poucas classes acústicas, ainda que densas foneticamente, ou seja, classes que representam muitos fonemas distintos, produz uma transformação suave das amostras, mantendo em grande parte a naturalidade da conversão. Mas dada a variabilidade da representação das classes, os primeiros momentos estatísticos (o valor médio e a covariância da

classe) não são capazes de moldar todos os detalhes conjunto. Neste caso, a conversão apresenta um baixo índice de similaridade.

Por outro lado, o uso de um conjunto amplo de classes acústicas tem o potencial de se ajustar precisamente ao conjunto de dados, possibilitando o alcance de altas taxas de similaridade na conversão. Não obstante, o ajuste excessivo aos dados (*overfitting*) pode ocasionar descontinuidade na transformação, uma vez que é permitido que classes vizinhas no espaço de característica sejam contaminadas com amostras ruidosas, alterando abruptamente os momentos estatísticos de baixa ordem.

Uma solução que ameniza o impacto deste problema consiste em tomar a média ponderada das transformações realizadas em cada classe e sintetizar o segmento ponderado a fim de combiná-lo com um segmento anterior. Tais transformações lineares são ponderadas com custos de concatenação e de alinhamento [45; 48; 282].

No entanto, o módulo de *Clusterização* apresentado anteriormente (Seção 3.4) nos fornece um conjunto estável e normalizado de agrupamentos acústicos, os quais atendem aos requisitos desejados e nos possibilita aproveitar ambos os benefícios de cada uma das abordagens acima:

1. O conjunto modelado corresponde a uma versão suavizada do espaço de características, tal que um envelope d-dimensional sobre os vetores acústicos de um determinado corpus corresponde a uma classe fonética artificial. Tal envelope previne o ajuste excessivo sobre os dados, o que resulta em boas taxas de naturalidade na reconstrução.

2. Há um número suficientemente grande de classes fonéticas artificiais, representadas por pequenas bases normalizadas quanto à largura de banda e espalhadas por todo o espaço de características, modelando assim detalhadamente cada nuance do falante.

Além das questões apontadas, para que se contemple um bom alinhamento entre dois corpora quaisquer, a conversão de voz inter-linguística ainda possui outros problemas que devem ser considerados e serão discutidos a seguir.

3.5.1 O Problema da Correspondência Fonética

No caso de conversão de voz inter-linguística, existe um problema adicional em relação à abordagem clássica, que é a não-correspondência fonética entre falantes de línguas distintas, uma vez que o conjunto de fonemas de uma língua é em geral distinto do de outra língua. No contexto espectral, é extremamente difícil separar aspectos da estrutura formântica relativos aos fonemas inerentes à língua dos aspectos relativos ao timbre de voz do indivíduo. Neste contexto, encontrar um mapeamento adequado é ainda mais problemático, visto que existem muitas classes fonéticas incomuns a ambas as línguas, conforme mostra a Figura 3.21. Nesta figura, as elipses contínuas representam as classes fonéticas artificiais do espaço de características do falante origem e as elipses tracejadas correspondem às classes fonéticas do falante destino.

Assim, sem qualquer tipo de informação fonética a priori, se estabelece um impasse quase que intransponível ao sistema de conversão inter-linguístico. Note que, assim como um falante de uma determinada língua tem dificuldades de pronúncia em outra língua, o sistema de conversão enfrenta o mesmo problema de mapeamento fonético. Assim, o sistema assume que o conversor de voz inter-linguístico pode ser considerado um tipo de conversor de voz convencional, o qual é independente de

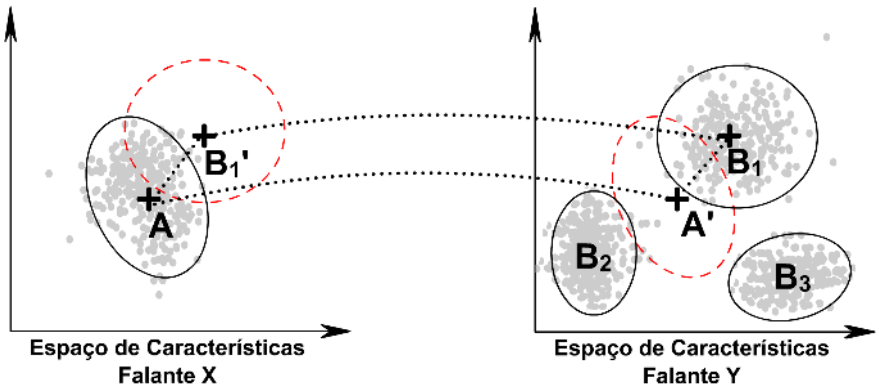

Figura 3.21: *Mapeamento entre classes fonéticas incompatíveis.*

texto na fase de treinamento, e cujos dados não são rotulados. Esta característica associada à língua em questão deverá ser levada em consideração na fase de experimentação perceptual do algoritmo.

Alguns autores contornam este problema com o uso de falantes bilíngues, no qual o falante destino deve pronunciar sentenças de ambas as línguas [2; 148]. O conjunto de sentenças de treinamento neste caso, é composto por sentenças pronunciadas na língua destino. No entanto, em função da motivação inicial do trabalho, esta abordagem não será considerada.

Pré-processamento dos Mapas Acústicos

A proposta deste trabalho a fim de contornar parte do problema anterior e amenizar o impacto da diferença linguística entre os falantes consiste em realizar uma normalização do mapa fonético devolvido pelo módulo anterior, de modo a equilibrar o nível global de energia de ambos os mapas fonéticos dos corpora em questão. Um processo de normalização inicial é aplicado aos mapas acústicos a serem pareados, e tem por objetivo criar uma representação canônica unificada de ambos os corpora envolvidos, de modo que seja possível alinhá-los mesmo no caso línguas distintas. A transformação deforma o mapa acústico de modo a distribuí-lo em torno da origem, com desvio padrão normalizado (igual a um) em todas as direções. A fim de ponderar as contribuições de cada classe, o peso de cada classe, dado pelo seu número de elementos, é agregado a um vetor artificial tridimensional $c = \langle f_1, f_2, p \rangle$ de modo que $f_1 \leq f_2$ funciona como uma chave seletora diretamente associada à classe $C(f_1, f_2)$ e p corresponde ao peso da classe. Por definição, as classes (f_1, f_2) que não possuem nenhum elemento são excluídas do conjunto de vetores artificiais.

Seja $\mathbb{C}$ o conjunto de vetores que define um mapa acústico. A normalização é realizada através da expressão

$$\bar{\mathbb{C}} = (c - \mathbb{E}\{\mathbb{C}\})(\mathrm{cov}\{\mathbb{C}\}^{-\frac{1}{2}}). \tag{3.20}$$

Opcionalmente, ao invés de se utilizar as matrizes de covariância, é possível se utilizar os desvios padrões estimados para cada componente vetorial de C. Com isso, apenas uma mudança de escala e uma translação pela média é realizada, sem deformar geometricamente o trapézio fonético original.

A Figura 3.22 exibe um exemplo no qual dois mapas fonéticos de dois locutores arbitrários foram normalizados segundo o critério de normalização descrito acima. A parte superior deste gráfico mostra um par de funções densidade de distribuição de cada segmento de voz em seus respectivos mapas fonéticos. O mapa em cores indica a quantidade de fonemas concentrados em cada ponto (f_1, f_2) relativo a uma classe fonética artificial. Quanto mais claro é o ponto no mapa,

maior a concentração de segmentos do corpus em relação à classe em questão. Ambos os corpora são provenientes de sentenças pronunciadas por locutores do gênero feminino, falantes da língua espanhola[4], os quais serão utilizados e melhor detalhados no Capítulo 4 de avaliação dos métodos propostos. A parte inferior do gráfico corresponde à versão normalizada de ambos os mapas fonéticos do corpus #1 e #2, conforme explicado no parágrafo anterior. As cores do mapa inferior indicam classes alinhadas, na qual pontos de mesma cor em ambas as classes #1 e #2 correspondem a classes alinhadas entre tais mapas. A escala do mapa de calor destes gráficos corresponde à densidade fonética no ponto respectivo, cujo valor absoluto é irrelevante nestes casos.

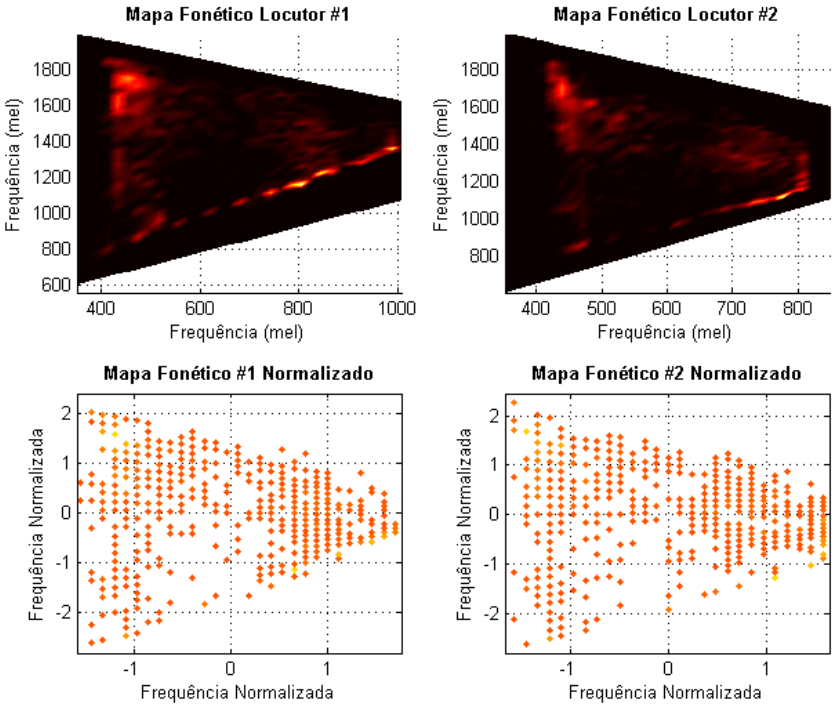

Figura 3.22: *Normalização do mapa fonético para fase de alinhamento.*

Dados dois conjuntos de classes fonéticas artificiais (ou mapas fonéticos) de ambos os corpora C^X e C^Y representados pelos respectivos conjuntos de chaves seletoras normalizadas $\bar{\mathbb{C}}^X$ e $\bar{\mathbb{C}}^Y$, o objetivo do alinhamento é encontrar uma função de mapeamento **bijetora** $@M : \bar{\mathbb{C}}^X \rightarrow \bar{\mathbb{C}}^Y$ definida como

$$@M(x_i) = y_i, \text{ tal que } \left\{ \sum_{x_i \in \bar{\mathbb{C}}^X} \mathbf{d}(x_i, y_i) \mid y_i \in \bar{\mathbb{C}}^Y \right\} \text{ seja mínimo,} \tag{3.21}$$

onde $@\mathbf{d}$ corresponde à função de distância euclidiana. No entanto, qualquer outra função distância, como por exemplo as que levam em consideração nosso sistema perceptual auditivo, poderia ser utilizada.

A importância da premissa de mapeamento bijetor se justifica pelo fato de que se deseja que para todas as classes do corpus origem haja uma classe correspondente no corpus destino (hipótese sobrejetora). Além disso, se espera que o maior número possível de classes do corpus destino seja

[4]os corpora 'ES_75' e 'ES_76' da base de dados TC-STAR foram gentilmente cedidos pelo prof. Dr. Antonio Bonafonte em nome da Universidade Politècnica da Catalúnya.

correspondida, a fim de evitar falta de cobertura fonética, estabelecendo assim uma relação de um-para-um entre os mapas fonéticos alinhados (hipótese injetora).

Um ponto relevante no alinhamento é a típica diferença de cardinalidade entre os conjuntos $\bar{\mathbb{C}}^X$ e $\bar{\mathbb{C}}^Y$, tornando impossível encontrar um mapeamento bijetor. Neste caso, o maior dos conjuntos será reduzido a fim de forçar a existência de uma correspondência biunívoca entre as classes. Esta redução é feita após a tentativa de pareamento, quando então as classes que não foram pareadas são excluídas da base de dados.

Note que a estratégia gulosa na qual se toma para cada ponto $x_i \in \bar{\mathbb{C}}^X$ um valor $y_i \in \bar{\mathbb{C}}^Y$ tal que $\mathbf{d}(x_i, y_i)$ seja mínimo pode produzir uma associação entre classes que contradiz a hipótese injetora, dado que uma mesma classe do mapa destino pode ser associada a mais de um x_i. No entanto, dados dois conjuntos de chaves seletoras $\bar{\mathbb{C}}^X$ e $\bar{\mathbb{C}}^Y$ podemos determinar um emparelhamento ótimo de classes entre dois falantes distintos, usando ferramentas disponíveis no compêndio teórico da teoria dos grafos.

3.5.2 Mapeamento usando Bases da Teoria dos Grafos

A grande vantagem de se abordar o problema do mapeamento usando teoria dos grafos é poder utilizar soluções formais muito bem amparadas pela literatura. Tipicamente, um grafo $\mathbb{G} = (V, A)$ é representado como um conjunto de vértices V e um conjunto de arestas A. Para cada aresta $a_{i,j} \in A$, existem um par de vértices v_i e v_j contidos em V tal que $a_{i,j} = (v_i, v_j)$.

Dadas classes fonéticas C^X e C^Y, o problema do mapeamento acústico é modelado a partir de um grafo bi-partido completo $\mathbb{G} = (V, A)$, com partição $V = V^x \cup V^y$ onde $V^x = \bar{\mathbb{C}}^X$, $V^y = \bar{\mathbb{C}}^Y$ e $A = \{(x_i, y_j) \mid \forall x_i \forall y_j\}$, sendo que toda aresta $a_{i,j} = (x_i, y_j)$ de A tem um custo (peso) definido como

$$w(a_{i,j}) = |x_i - y_j|,$$

que é a distância euclidiana entre x_i e y_j.

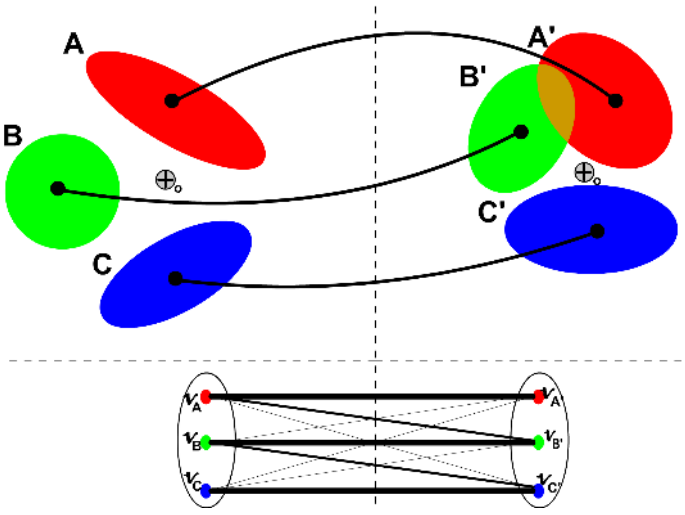

Figura 3.23: *Relação existente entre o problema do mapeamento ótimo e o emparelhamento perfeito.*

O *Problema do Mapeamento Acústico Ótimo* entre classes fonéticas de falantes distintos é então reduzido ao *Problema do Emparelhamento Máximo de Custo Mínimo – EMCM* em grafos bi-partidos. Neste caso, dois vértices x_i e y_j estão pareados no mapeamento acústico se, e somente

se, a aresta $a_{i,j}$ pertence ao emparelhamento perfeito, conforme mostra a Figura 3.23.

Alguns conceitos básicos sobre a Teoria de Grafos pode ser encontrada no Apêndice B.2, os quais embasam o algoritmo que resolve o problema do EMCM.

Sabe-se que este problema é o problema dual do Emparelhamento Máximo de Peso Máximo – EMPM, uma vez que dada uma atribuição de pesos negativos às arestas de um problema, a solução do EMPM e do EMCM são iguais. Ambos os problemas podem ser eficientemente resolvidos pelo clássico *Algoritmo de Hopcroft e Karp* [95] em tempo $\mathcal{O}(|A|\sqrt{|V|})$. Seu método antecessor, o Algoritmo Húngaro [124], originalmente desenvolvido para resolver o problema da atribuição de tarefas [202], também resolve o problema do emparelhamento máximo. Tal algoritmo encontra um EMPM através de sequências de caminhos aumentativos em tempo $\mathcal{O}(|A||V|)$. A diferença básica entre ambos os métodos é que o Algoritmo Húngaro encontra um caminho aumentativo por iteração, enquanto que o de Hopcroft e Karp encontra um conjunto maximal de caminhos aumentativos curtos por iteração, reduzindo a complexidade computacional.

Uma vez que a complexidade computacional do algoritmo está diretamente ligada com a quantidade de arestas (por ser em número muito maior), uma "poda" no conjunto de arestas é realizada a fim de eliminar casos improváveis (distâncias maiores que a média global), ganhando assim tempo de processamento computacional. Assim, passemos a uma descrição algorítmica do método de Hopcroft e Karp.

Algoritmo de Hopcroft e Karp

Considere o grafo $\mathbb{G} = (V_1 \cup V_2, A)$ definido pela redução do problema de mapeamento de classes fonéticas ao problema do EMCM. O algoritmo consiste em buscar caminhos aumentativos que sejam alternantes de um nível de busca para outro. O algoritmo é bastante simples, e é composto por quatro passos básicos, os quais seguem:

1. Inicialmente, os vértices livres em V_1 são usados como pontos de partida na busca. (a) Num primeiro nível de busca, somente as arestas que não foram marcadas podem ser visitadas; (b) Nos níveis subsequentes da busca, arestas já visitadas podem ser revisitadas. Porém, o algoritmo realiza uma busca por caminhos intercalados de arestas visitadas e não-visitadas anteriormente. A busca termina num k-ésimo nível quando um ou mais vértices livres em V_2 forem encontrados (ou seja, um ou mais caminhos aumentativos forem encontrados).

2. Posteriormente, todos os vértices de V_2 no nível k são agrupados em F. Cada vértice destes é um extremo de um caminho aumentativo.

3. Então, o algoritmo passa a uma fase de busca em profundidade por um conjunto de vértices disjuntos que compõem caminhos aumentativos de comprimento k (na camada k). O sentido da busca parte de F em direção aos vértices livres de V_1. Note que o primeiro nível de busca permite que a busca em profundidade leve a vértices não usados na camada anterior, ou seja, a busca constrói um caminho alternante aumentativo o qual deve necessariamente envolver algum vértice de F (não usado anteriormente). A cada caminho aumentativo encontrado, o algoritmo toma um novo vértice de F, e o processo de busca em profundidade é reiniciado.

4. A cada caminho aumentativo encontrado, o conjunto de emparelhamentos M é atualizado. Quando não é possível encontrar mais caminhos aumentativos, o algoritmo para.

O Algoritmo `hopcroft_karp`[5] toma como entrada um grafo $\mathbb{G}$ e devolve um conjunto de arestas E do emparelhamento máximo de custo mínimo, conforme foi detalhado no tópico anterior. A Linha 3 do algoritmo corresponde aos itens 1 e 2 do método enunciado no tópico anterior. A Linha 4 computa a diferença simétrica entre os conjuntos E e $\bigcup\mathcal{P}$, a qual realiza a alternância entre os caminhos aumentativos.

Algoritmo 3.24 $E \leftarrow$ `hopcroft_karp`$(\mathbb{G})$

▷ *Inicializa* E
1 $E \leftarrow \emptyset$
2 **repeat**
3 $\quad \mathcal{P} \leftarrow \{P_1, P_2, \ldots, P_k\}$ *onde* $\mathcal{P}$ *é um conjunto maximal de caminhos aumentativos mais curtos com vértices disjuntos.*
4 $\quad E \leftarrow E \ominus \bigcup\mathcal{P}$
5 **until** $\mathcal{P} = \emptyset$
6 **return**(E)

Para ilustrar o resultado do emparelhamento entre dois corpora distintos, um exemplo de emparelhamento entre ambos os corpora usados na Figura 3.22 é exibido na Figura 3.24.

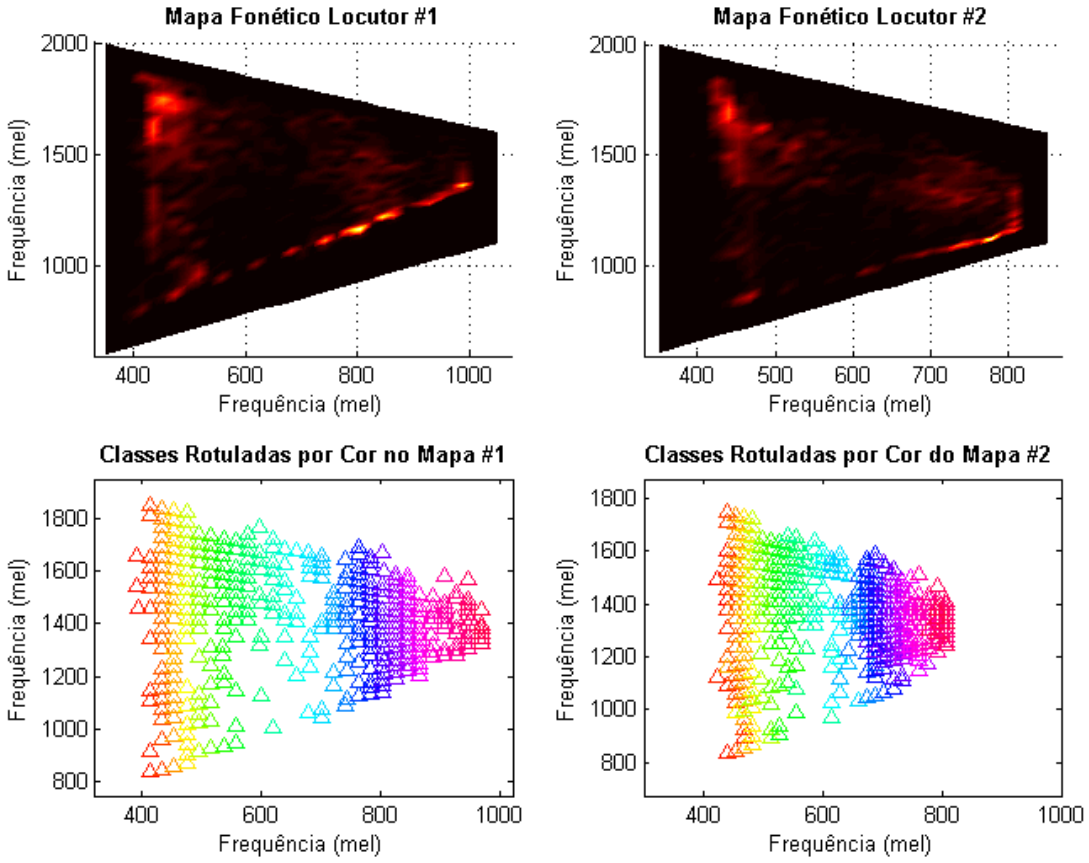

Figura 3.24: *O método de alinhamento aplicado ao alinhamento de sentenças paralelas.*

Neste gráfico, pontos de mesma cor estão alinhados entre ambos os corpora não-paralelos. Dada uma sentença pronunciada pelo falante #1, espera-se que as associações criadas entre as classes do

[5] O algoritmo foi obtido no sítio http://en.wikipedia.org/wiki/Hopcroft-Karp_algorithm.

corpus #1 e as classes do corpus #2 (de mesma cor) permitam definir uma função de transformação linear (em partes) que deforme apropriadamente o conteúdo espectral da sentença para se ajustar ao mapa #2.

Algoritmo 3.25 $@M \leftarrow$ `Mapeamento`(C^X, C^Y)

▷ *Toma os Centroides de cada Classe*
1 **for** $n = 1 : N$
2 Calcule $\bar{\mathbb{C}}^X$ pela Eq. 3.20 usando C^X
3 Calcule $\bar{\mathbb{C}}^Y$ pela Eq. 3.20 usando C^Y
4 **end for**

▷ *Define os Conjuntos de Vértices*
5 $V_1 \leftarrow \bar{\mathbb{C}}^X$
6 $V_2 \leftarrow \bar{\mathbb{C}}^Y$

▷ *Define o Conjunto de Arestas*
7 $A_{i,j} \leftarrow \mathbf{d}(x_i, y_j),\ \forall x_i \in V_1,\ y_j \in V_2$

▷ *Realiza o Emparelhamento*
8 Defina Grafo $\mathbb{G} = (V_1 \cup V_2, A)$
9 $E \leftarrow$ hopcroft_karp$(\mathbb{G})$
10 **foreach** $a_{i,j} \in E$
11 Obtém Classes $c_i \in \bar{\mathbb{C}}^X$ e $c_j \in \bar{\mathbb{C}}^Y$ correspondentes à $x_i \in V_1$ e $y_j \in V_2$
12 $@M \leftarrow @(c_i)[c_j]$
13 **end foreach**
14 **return**$(@M)$

O algoritmo base do Módulo de Mapeamento, o denominado Algoritmo `Mapeamento` (Alg. 3.25) é apresentado usando os conceito de grafos anteriormente enunciados. Como entrada, o algoritmo recebe dois conjuntos de classes acústicas C^X e C^Y (ambas de tamanho N) dos corpora origem e destino, respectivamente, e devolve uma função que associa cada classe C^X a uma classe C^Y, segundo critérios de minimização de distância global. Como definido anteriormente, as classes fonéticas artificiais são representadas pelas suas respectivas chaves normalizadas $\bar{\mathbb{C}}^X$ e $\bar{\mathbb{C}}^Y$, as quais serão usados como chaves de seleção nas Linhas 2 e 3.

Embora não esteja no algoritmo acima, uma filtragem opcional pode ser realizada, a fim de eliminar potenciais erros de alinhamento entre pares classes acústicas distantes entre si. Um processo de filtragem idêntico ao adotado para filtragem de classes k-verossímeis (vide Seção 3.4) pode ser efetivado, de modo que todos os pares de classes (c_i, c_j) pertencentes ao emparelhamento são eliminados do corpus, uma vez constado que $w(A_{i,j}) > \mathbb{E}\{w(A)\} + \eta \text{std}\{w(A)\}$, onde $w(A_{i,j})$ corresponde ao peso da aresta $A_{i,j}$, a qual representa a associação entre a classe c_i e c_j; $\mathbb{E}\{w(A)\}$ é o valor esperado dos pesos de todas as arestas de A, assim como $\text{std}\{w(A)\}$ representa o desvio padrão amostral de $w(A)$. O valor η *default* assumido no trabalho foi $\eta = 0.25$.

Note que o método de emparelhamento também poderia ser utilizado para alinhar segmentos de voz em sentenças pronunciadas em corpora paralelos. Neste caso, uma quantização de cada segmento de voz é suficiente para ser usada como chave de seleção. A Figura 3.25 exibe um caso simples onde duas sentenças paralelas foram alinhadas, nas quais foram pronunciadas as vogais [a], [e], [i], [o], [u] por falantes hispânicos distintos. Neste gráfico, um ponto qualquer (x, y) é negro se os segmentos x e y foram alinhados pelo algoritmo, e preto caso contrário. Cada segmento de

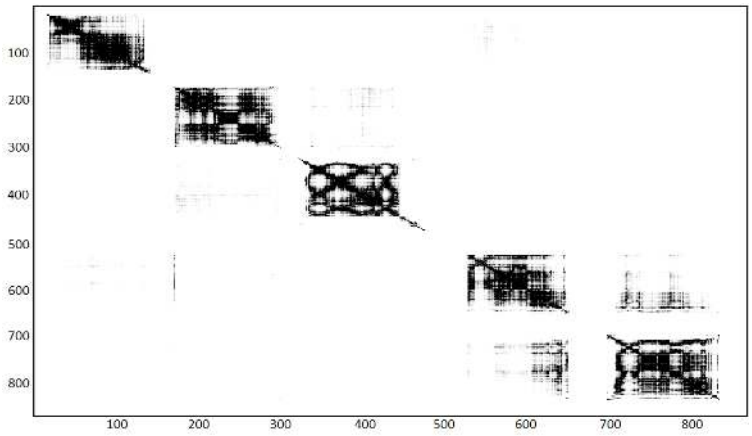

Figura 3.25: *O método de alinhamento aplicado ao alinhamento de sentenças paralelas.*

voz foi quantizado tomando-se somente as amplitudes harmônicas dos envelopes quantizados na fase de Parametrização (ver Seção 3.3). Como ambos os corpora se encontram (quase)-alinhados no tempo, espera-se que os pontos de maior magnitude no gráfico se concentrem em torno de quadrados centrados na diagonal principal. Cada quadrado corresponde a uma das 5 vogais pronunciadas.

A fim de completar as módulos do sistema, o estágio seguinte toma um conjunto de classes acústicas e alguns parâmetros de controle global da sentença, a fim de transformar parâmetros acústicos de um falante para o outro.

3.6 Estágio V: Transformação dos Parâmetros Acústicos

A etapa de transformação (Figura 3.26) corresponde ao último estágio dentro da fase de treinamento, a qual devolve uma função de transferência a ser utilizada na fase de conversão propriamente dita. Sendo assim, é necessário de antemão definir uma função de transferência que manipule coeficientes espectrais de uma sentença pronunciada de modo que as informações acústicas sejam convertidas de um indivíduo para outro, seguindo padrões de conversão estipulados pelos alinhamentos entre classes fonéticas artificiais de um falante para o outro.

Em termos algorítmicos, o **Módulo de Transformação** se divide em duas partes: a primeira corresponde à definição da função dentro da fase de treinamento, a qual toma ambos os mapas fonéticos do corpus origem e destino, e os caracteriza acusticamente em nível **global** e **local**. Tal caracterização devolve ao módulo uma função de conversão $\mathcal{T}$ que carrega consigo um conjunto de parâmetros estatísticos, os quais são utilizados na fase de conversão. O Algoritmo 3.26 implementa o módulo de Transformação, conforme descrito acima, e devolve a função $\mathcal{T}$ a ser utilizada na fase de conversão.

Algoritmo 3.26 $@\mathcal{T} \leftarrow$ `Transformação`$(@M, C^X, C^Y, G^X, G^Y)$

▷ *Composição de Transformações Locais e Globais*

1 $@\mathcal{T} \leftarrow @(\Psi)\left[\text{transformação\_local}\left(\text{transformação\_global}(\Psi, G^X, G^Y), @M, C^X, C^Y, G^X, G^Y\right)\right]$

2 **return**$(@\mathcal{T})$

Definidos os mapas entre as classes fonéticas artificiais devidamente pareadas, o sistema deve ser capaz de, dada uma sentença de voz pronunciada pelo falante origem, tomar cada segmento de voz e associá-lo à sua correspondente classe fonética no corpus origem. De tal associação se estima uma

transformação que deforma o conteúdo espectral da sentença de modo que seus segmentos internos sejam parametrizados, classificados e avaliados como se pertencessem ao corpus destino.

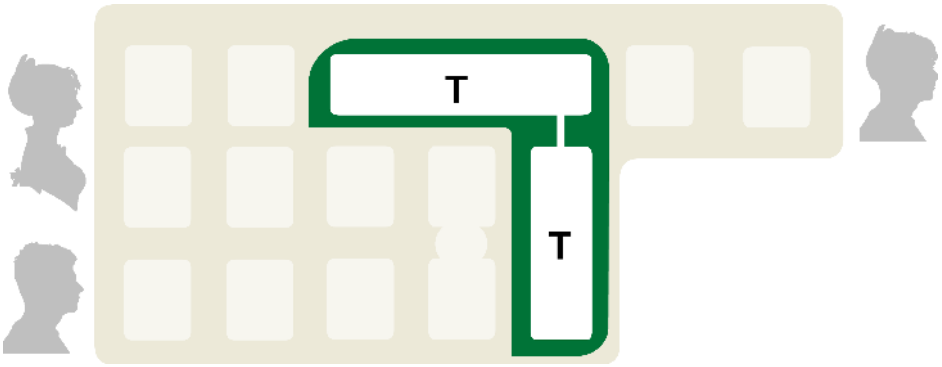

Figura 3.26: *Definição da função de Transformação (**T**)*

Dois tipos de transformações são aplicadas a uma sentença de voz. O primeiro tipo pressupõe que a função de transferência aplicada sobre cada segmento de voz possui parâmetros globais ao nível da sentença; por outro lado, as transformações locais estão definidas num escopo segmental, ou seja, usam como suporte as classes fonéticas artificiais dinamicamente atribuídas a cada segmento de voz, de acordo com critérios de proximidade em relação aos centroides dessas classes.

3.6.1 Transformações Globais ao Nível da Sentença

Conforme visto no Módulo de Clusterização (Seção 3.4), a estrutura G contém médias e desvios laterais dos contornos de energia e pitch, bem como médias e matrizes de covariância das amplitudes harmônicas e estocásticas a nível global. Sendo assim, é conveniente separar a discussão em duas partes, uma relacionada à conversão da prosódia e outra à conversão espectral.

Quanto à conversão de prosódia, uma transformação linear de cada parâmetro de prosódia F_0 e E_0 é suficiente para transpor de maneira mais ampla os contornos de energia e pitch de uma sentença pronunciada. Os momentos estatísticos suficientes para realizar tal conversão de prosódia foram estimados apropriadamente no Módulo de Clusterização (Seção 3.4), entre as Linhas 11 e 14 da função `momentos_estatísticos_globais` implementada pelo Alg. 3.22. Note que os dois contornos de prosódia foram modelados pelas suas respectivas médias e desvios padrões laterais.

Dados dois conjuntos de parâmetros globais G^X e G^Y, a conversão do contorno de um coeficiente de prosódia $P_0 \in \{F_0, E_0\}$ toma cada parâmetro de entrada associado a um dado segmento de voz k e atualiza $\Psi^k_{P_0}$ de modo que

$$[\Psi^k_{P_0}]' \leftarrow G^Y.P_0.\mu + \frac{\sigma^Y_{P_0}}{\sigma^X_{P_0}}\left(\Psi^k_{P_0} - G^X.P_0.\mu\right),$$

onde cada $\Sigma^X_{P_0}$ e $\Sigma^Y_{P_0}$ é montando dinamicamente de acordo com $\Psi^k_{P_0}$ de modo que

$$\sigma^X_{P_0} = \begin{cases} G^X.P_0.\sigma_{\mathrm{u}}, & \text{se } \Psi^k_{P_0} \geq G^X.P_0.\mu \\ G^X.P_0.\sigma_{\mathrm{d}}, & \text{se } \Psi^k_{P_0} < G^X.P_0.\mu \end{cases}$$

assim como

$$\sigma^Y_{P_0} = \begin{cases} G^Y.P_0.\sigma_{\mathrm{u}}, & \text{se } \Psi^k_{P_0} \geq G^X.P_0.\mu \\ G^Y.P_0.\sigma_{\mathrm{d}}, & \text{se } \Psi^k_{P_0} < G^X.P_0.\mu \end{cases}.$$

Analogamente, as transformações (opcionais) de amplitudes estocásticas e harmônicas seguem o mesmo padrão de transformação, ou seja, dado um espectro de magnitude $M \in \{S, A\}$ do k-ésimo segmento de voz Ψ_M^k, o resultado da conversão é obtido como

$$[\Psi_M^k]' \leftarrow G_M^Y.\mu + (G_M^Y.\Sigma)^{0.5}(G_M^X.\Sigma)^{-0.5}\left(\Psi_M^k - G_M^X.\mu\right),$$

onde todas as médias e matrizes de covariância contidas em G_M^X e G_M^Y foram anteriormente calculadas pelo módulo de Clusterização. As amplitudes estocásticas são submetidas apenas a este tipo de transformação. Já no caso particular de conversão de amplitudes harmônicas, a conversão global não é suficiente para converter todas as nuances espectrais de um corpus para outro. Neste caso particular, uma conversão local posterior a esta é realizada ao nível das classes fonéticas artificiais, desconsiderando assim a transformação feita pelo algoritmo a priori. Ao final da transformação local, os momentos estatísticos globais extraídos das amplitudes harmônicas serão utilizados para enfatizar a transformação espectral harmônica ao nível da sentença.

A função de `transformação_global` é então implementada pelo Algoritmo 3.27 conforme observado a seguir.

Algoritmo 3.27 $\Psi \leftarrow$ `transformação_global`(Ψ, G^X, G^Y)

1 $N_s \leftarrow$ dimensão(Ψ)
2 **for** $k = 1 : N_s$

▷ *Conversão da prosódia*
3 **if** $\Psi_{F_0}^k > 0$ **then**
4 **if** $\Psi_{F_0}^k \geq G^X.F_0.\mu$ **then**
5 $\Psi_{F_0}^k \leftarrow G^Y.F_0.\mu + \frac{G^Y.F_0.\sigma_u}{G^X.F_0.\sigma_u}\left(\Psi_{F_0}^k - G^X.F_0.\mu\right)$
6 **else**
7 $\Psi_{F_0}^k \leftarrow G^Y.F_0.\mu + \frac{G^Y.F_0.\sigma_d}{G^X.F_0.\sigma_d}\left(\Psi_{F_0}^k - G^X.F_0.\mu\right)$
8 **end if**
9 **if** $\Psi_{E_0}^k \geq G^X.E_0.\mu$ **then**
10 $\Psi_{E_0}^k \leftarrow G^Y.E_0.\mu + \frac{G^Y.E_0.\sigma_u}{G^X.E_0.\sigma_u}\left(\Psi_{E_0}^k - G^X.E_0.\mu\right)$
11 **else**
12 $\Psi_{E_0}^k \leftarrow G^Y.E_0.\mu + \frac{G^Y.E_0.\sigma_d}{G^X.E_0.\sigma_d}\left(\Psi_{E_0}^k - G^X.E_0.\mu\right)$
13 **end if**
14 **end if**

▷ *Conversão das amplitudes estocásticas*
15 $\Psi_S^k \leftarrow G^Y.S.\mu + G^Y.S.\Sigma^{0.5}(G^X.S.\Sigma)^{-0.5}\left(\Psi_S^k - G^X.S.\mu\right)$
16 **end for**
17 **return**(Ψ)

O algoritmo acima toma uma sentença de voz representada por uma sequência de vetores acústicos relativos a cada segmento de voz e os converte usando parâmetros estatísticos globais. Conforme comentado anteriormente, uma conversão de voz da parte harmônica baseada na aplicação da conversão global não é capaz de transformar todas as nuances espectrais, devido à forte correlação entre as médias e desvios padrões de ambos os corpora destino e origem, conforme pode ser observado na Figura 3.27. A parte superior deste gráfico mostra a relação entre as médias e os desvios padrões entre ambos os corpora tanto para amplitudes harmônicas (à esquerda) quanto para amplitudes estocásticas (à direita). A parte inferior do gráfico mostra um exemplo onde a conversão do es-

pectro harmônico pontilhado em negro é feita sem o ajuste local de classe a classe. Neste caso, o espectro transformado (pontilhado em vermelho) é alterado somente numa direção devido ao fato de as médias e desvios padrões serem constantes a nível local. Além disso, quanto maior a correlação entre os parâmetros globais de ambos os corpora, menor é o efeito da conversão, como acontece com locutores de uma mesma língua e mesmo sexo. Como resultado, a conversão de voz não modifica apropriadamente a sentença pronunciada, conforme será subjetivamente avaliado na Seção 4.

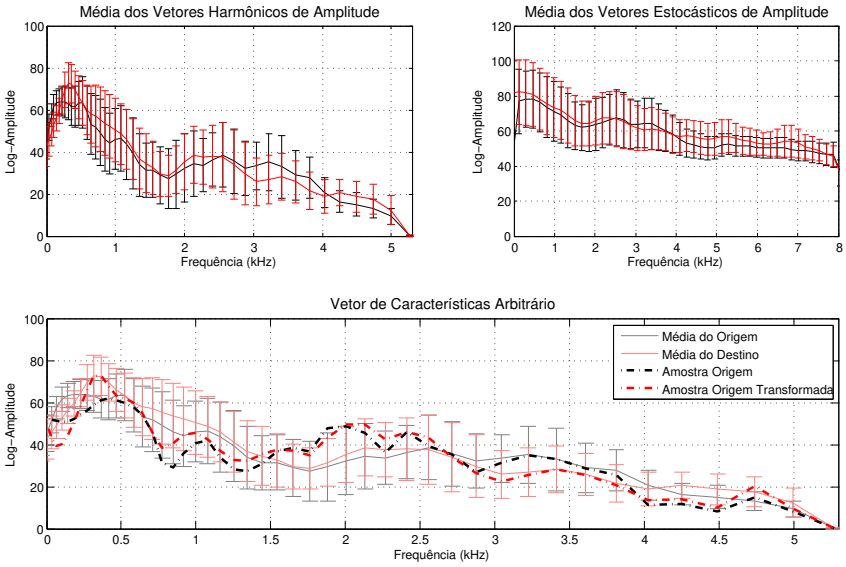

Figura 3.27: *Conversão de um segmento de voz usando apenas parâmetros globais.*

Já no caso das amplitudes estocásticas, o trabalho se propôs a realizar uma conversão apenas a nível local. A justificativa se dá diante da necessidade de modelar tais componentes em termos de classes fonéticas artificiais, o que aumenta significativamente o custo computacional do sistema sem acarretar uma melhoria significativa da conversão. Por esta razão, somente a conversão de voz da parte harmônica é realizada a nível local, conforme segue.

3.6.2 Transformações Locais ao Nível do Segmento

A ideia central da transformação local é tomar cada segmento quantizado da entrada (do falante origem) e encontrar a classe fonética artificial (ou um conjunto delas) que melhor caracteriza o segmento avaliado. Então, a partir da função de mapeamento $@M$ devolvida pelo Módulo de Mapeamento, se define uma transformação de conversão, como a transformação global vista anteriormente.

Em primeiro lugar, o acesso a estas classes se dá por meio de chaves seletoras únicas para cada classe. Uma **chave seletora** de uma classe é um conjunto de parâmetros acústicos suficientes para discriminar uma classe das outras. Algumas abordagens de seleção de classes fonéticas artificiais são consideradas por este trabalho. A escolha do método de seleção será realizada com base em resultados experimentais realizados no Capítulo 4.

Uma primeira alternativa consiste em empregar os métodos utilizados em fases anteriores do sistema. Uma vez que as classes fonéticas artificiais provenientes do estágio de clusterização já estão quantizadas em um banco de filtros, o mesmo método de quantização pode ser aplicado ao vetor v. O método de quantização neste caso corresponde ao Algoritmo 3.17, que retorna um mapa de

distribuição formântica $distr_v$, o qual normalizado constitui uma chave de seleção de modo que

$$distr_v^{norm} \leftarrow G.discr.\mu \cdot G.discr.\sigma \cdot (distr_v - G.discr.\mu)^3, \tag{3.22}$$

onde G é a estrutura global do corpus origem (X) definida no Módulo de Clusterização. Como cada classe fonética artificial $C(f_1, f_2)$ está biunivocamente associada a um ponto $c = (f_1, f_2)$ do mapa $distr_v$, uma lista decrescentemente ordenada pelas magnitudes de cada ponto do mapa corresponde à lista de classes candidatas a se associar à v. Além disso, cada magnitude w_c associada ao ponto c do mapa corresponde ao grau de pertinência do vetor v em relação à classe associada $C(f_1, f_2)$, mais especificamente, ao peso

$$\mathbf{w}_D(v, C) = w_c^2.$$

Alternativamente, uma das formas mais comuns de selecionar classes em geral consiste na utilização de centros de massa das classes fonéticas artificiais de um determinado espaço de características acústicas, os bem conhecidos **centroides**. A escolha de uma função de distância apropriada é de crucial importância no processo de seleção. A distorção espectral [13], por exemplo, é caracterizada pela distância euclidiana logarítmica do espectro de magnitude. Ao invés de usar o centroide outras também poderiam contribuir para a especificação das chaves seletoras, derivadas e outras informações estatísticas, como por exemplo medidas de dispersão, assimetria e curtose, entre outras. Como definido anteriormente, cada classe acústica é representada pela média pela matriz de covariância dos coeficientes relativos às amplitudes harmônicas normalizadas, conforme descrito nas linhas finais do Algoritmo 3.20.

Dada uma classe fonética C representada pela sua respectiva média μ_C e matriz de covariância de Σ_C, e dado o k-ésimo vetor acústico $v = \Psi_A^k$, a distância de Mahalanobis [142] é uma clássica alternativa para definir uma relação entre um vetor v e uma classe C. Neste caso, quanto menor a distância

$$\mathbf{d}_M(v, C) = \sqrt{(v - \mu_C)^T (\Sigma_C)^{-1} (v - \mu_C)}, \tag{3.23}$$

maior a significância da associação entre a classe C e o vetor v. Dada tal métrica, a lista de classes candidatas a se associar a v é organizada crescentemente segundo as distâncias de Mahalanobis, de modo a quantificar a pertinência de v em cada conjunto C do corpus. O grau de pertinência do vetor v em relação a uma classe C é definido como

$$\mathbf{w}_d(v, C) = \left(1 - \frac{\mathbf{d}_M(v, C) - \min_{\forall k}\{\mathbf{d}_M(v, C_k)\}}{\max_{\forall k}\{\mathbf{d}_M(v, C_k)\} - \min_{\forall k}\{\mathbf{d}_M(v, C_k)\}}\right)^4, \tag{3.24}$$

onde $L^* = 48$ é a dimensão de v e o conjunto de todas as classes C_k compõe o conjunto das classes fonéticas artificiais do respectivo corpus. Um cuidado especial deve ser tomado em corpora com baixa densidade de dados, devido à insuficiente modelagem estatística, refletindo principalmente uma representação imprecisa na matriz de covariância.

Na Equação 3.24, em vez de se utilizar a distância de Mahalanobis ($\mathbf{d}_M(v, C_k)$), qualquer outra distância (ou métrica) poderia ser conveniente adaptada neste módulo de seleção de classe fonética artificial do corpus origem, como por exemplo, a distância Euclidiana dos espectros logarítmicos, a conhecida distorção espectral na escala MEL [123], ou mesmo outras distância perceptuais como a distorção espectral Bark [296; 297], ou a distância de Itakura-Saito [200], entre outras. Frequentemente, os coeficientes Mel-Cepstrais são utilizados para a caracterização de vetores acústicos com

altas taxas de acerto, principalmente em sistemas de reconhecimento de fala [167; 188]. No capítulo experimental (Capítulo 4) serão avaliadas algumas destas medidas a fim de detectar uma boa alternativa a ser considerada no sistema final apresentado por este trabalho, bem como formulações da função de ponderação $@\mathbf{w}_d$.

A partir destas listas de candidatos, as transformações locais são realizadas usando os primeiros m candidatos mais significativos com respeito ao grau de pertinência. A composição de transformações locais ponderadas é um recurso usado comumente em transformações lineares baseadas em GMM, e evita descontinuidades na transformação [298]. Tal abordagem é usada no contexto deste trabalho, com um número de classes a serem interpoladas proporcional à dimensão do mapa acústico. Supondo que v pertence às classes $C_1, C_2, \ldots, C_m$ com graus de pertinência $w_1, w_2, \ldots, w_m$, respectivamente, a composição de transformações ponderadas é definida como

$$v' = \frac{\sum_{i=1}^{m} w_i \mathcal{T}_{loc}(v, C_i, M(C_i))}{\sum_{j=1}^{m} w_j}, \tag{3.25}$$

onde $\mathcal{T}_{loc}$ é uma função de conversão arbitrária e $@M$ corresponde à função de mapeamento entre classes devolvida no Módulo de Mapeamento (Seção 3.5).

Algoritmo 3.28 $\Psi \leftarrow$ `transformação_local`$\big(\Psi, @M, C^X, C^Y, G^X, G^Y\big)$

1 $N_s \leftarrow$ dimensão(Ψ)
2 **for** $k = 1 : N_s$
3 **if** $\Psi^k_{F_0} > 0$ **then**
▷ *Gera lista de graus de pertinência* $\mathbf{w}$
4 **foreach** $c \in C^X$
5 $w(c) \leftarrow \mathbf{w}_d(v, C)$
6 **end foreach**
▷ *Toma as* M_c *classes que obtiveram maiores* $\mathbf{w}$
7 $C^* \leftarrow \arg\min_{M_c}\{w\}$
▷ *Guarda valores de* Ψ^k_A *para posterior reajuste final de energia*
8 $a(k, \cdot) \leftarrow \Psi^k_{E_0} \cdot \Psi^k_A$
▷ *Realiza transformações locais ponderadas*
9 $\Psi^k_A \leftarrow \frac{\sum_{c \in C^*} w(c) \mathcal{T}_{loc}(\Psi^k_A, c, M(c))}{\sum_{c \in C^*} w(c)}$
10 **end if**
11 **end for**
▷ *Fase de pós-processamento*
12 $\Psi \leftarrow$ pós_processamento(Ψ, a, G^X, G^Y)
13 **return**(Ψ)

O valor de m adotado pelo trabalho foi $M_c = 15$, conforme definido na Tabela de Parâmetros Globais (Seção 3.1). Tal valor foi empiricamente ajustado. No entanto, é válido observar que, quanto maior M_c, maior a contribuição de classes alinhadas consideradas na função de transformação, e consequentemente, menor o impacto percebido na transformação local de um fonema em particular. Em contrapartida, quanto menor M_c, melhor é o ajuste local, e assim, a transformação a nível local (em termos fonéticos) é bem mais precisa, o que acarreta uma reconstrução mais descontínua no tempo. Em outras palavras, maiores valores de M_c tendem a realizar uma conversão de voz em um

escopo mais global, proporcionando mais naturalidade ao sinal convertido e menor similaridade em relação ao falante destino. Por outro lado, menores valores de M_c tendem a converter a sentença em um escopo mais local, provendo taxas maiores de similaridade, mas comprometendo a naturalidade do sinal transformado.

O Algoritmo 3.28 implementa a função de transformação local de amplitudes harmônicas que toma como entrada a sequência de vetores acústicos Ψ do falante origem (X), as classes acústicas C^X e C^Y dos corpora origem e destino, respectivamente, além dos correspondentes parâmetros de controles globais G^X e G^Y, juntamente com a função de mapeamento $@M$ entre tais classes.

Conforme se pode observar na Linha 5 do Algoritmo 3.28, qualquer uma das duas métricas de distância podem ser utilizadas para selecionar as classes do corpus origem. De qualquer forma, na Linha 7, as classes mais pontuadas em termos de grau de pertinência são selecionadas a fim de se realizar a composição de transformações locais ponderadas na Linha 9. Note a presença de uma função genérica $\mathcal{T}_{loc}$ nesta mesma linha, que é a função de transferência espectral definida a seguir. Ao final, a função `pós_processamento`(Ψ, a) tem o objetivo de suavizar a transição entre segmentos consecutivos, eliminando assim ruídos introduzidos por eventuais erros de mapeamento e/ou transformação, e aumentando o grau de naturalidade do sinal reconstruído.

Propostas Clássicas de Transformação Local

O sistema básico de conversão adota como método de transformação local o mesmo método usado para realizar as transformações globais, a saber, a transformação linear usando matrizes de covariâncias completas de cada uma das classes acústicas envolvidas. Esta abordagem implementa a função genérica de modo que

$$\mathcal{T}_{loc}^{LT\text{-}full}(v, X, Y) = \mu_Y + \Sigma_Y^{0.5}(\Sigma_X^{-0.5})(v - \mu_X), \tag{3.26}$$

onde μ_X e μ_Y são respectivamente as médias dos vetores das classes origem e destino, assim como Σ_X e Σ_Y são suas respectivas matrizes de covariância. O uso das matrizes de covariância permite que os dados sejam deformados no espaço de características acústicas, enquanto que as médias possibilitam a translação destes vetores. O problema desta abordagem é que a mesma requer um número suficientemente grande de dados, de modo a evitar problemas de mal-condicionamento da matriz, como matrizes (quase-)singulares, por exemplo. Experimentos mostram que artefatos ruidosos são inevitavelmente introduzidos no sinal de saída, quando se trata de corpora com uma base de dados que não é grande o suficiente (o caso deste trabalho).

Uma solução alternativa para este caso é o uso de matrizes diagonais (as variâncias) adicionadas de um valor ϵ pequeno na transformação definida como

$$\mathcal{T}_{loc}^{LT\text{-}diag}(v, X, Y) = \mu_Y + \mathrm{diag}(\Sigma_Y + \epsilon)^{0.5}(\mathrm{diag}(\Sigma_X + \epsilon)^{-0.5})(v - \mu_X), \tag{3.27}$$

onde `diag` converte uma matriz completa em uma matriz diagonal.

Neste caso, a introdução de ϵ diminui o impacto de um valor próximo a zero em Σ_X. No entanto a propriedade de rotação dos vetores no espaço de características acústicas é perdida. Além disso, a propriedade de escala é comprometida com o uso de valores ϵ relativamente grandes. O problema da definição da matriz de covariância é ainda mais crítico em casos nos quais uma componente de Σ_X é muito menor que a respectiva componente em Σ_Y. Neste caso, pequenas flutuações ruidosas

que se excederem ao desvio padrão no falante origem geram grandes flutuações ruidosas no sinal transformado. Para contornar este problema, uma nova proposta é introduzida neste trabalho, nomeada de Deformação em Frequência Normalizada.

Deformação em Frequência Normalizada

A Deformação em Frequência Normalizada (ou também, Empenamento em Frequência Normalizada) é uma função de transferência não-linear entre duas funções reais positivas. A maioria dos métodos de deformação, ou empenamento de frequência [270] utilizam métodos de seleção de segmentos com máxima similaridade usando técnicas de otimização, tais como programação dinâmica por exemplo (ver Seção 2.4.2 do Capítulo 2).

Este trabalho propõe um conceito de deformação em frequência baseado em **distribuições em frequência normalizada**. A motivação para o uso desta abordagem parte do princípio de que as regiões formânticas do sinal de voz concentram sua maior parte energética em picos do envelope espectral, e este fato não depende da língua em particular. Por esta razão, não é razoável transformar linearmente uma envoltória espectral discreta, sem que antes haja um re-alinhamento energético do sinal usando como padrão os envelopes espectrais médios das duas classes alinhadas para conversão. Tal deformação em frequência distorce levemente a escala de frequência, ou seja, realiza um empenamento em frequência (no Inglês, *Frequency Warping*) entre um envelope espectral e outro. Esta estratégia é bastante comum entre os trabalhos de conversão de voz inter-linguística [241].

Vale ressalvar que o espaço de características acústicas de um sinal de voz é composto por um conjunto de vetores quantizados, que correspondem a cada envelope espectral harmônico de seus respectivos segmentos de voz. No contexto deste trabalho, um k-ésimo envelope espectral harmônico Ψ_A^k é representado pela soma de L^* funções contínuas $\psi(f, a_l, \mu_l, \sigma_l)$, onde a, μ e σ são amplitudes, frequências centrais e larguras de banda da base radial ψ, conforme definido na Seção 3.3. Vale salientar que os valores μ e σ são fixos em todos os segmentos modelados, como definido anteriormente. No caso de conversão inter-linguística, é importante criar uma relação recíproca entre a distribuição de energia espectral e a escala de frequências associada (linear, Bark ou MEL), visto que uma parcela dos fonemas de uma língua é inexistente na outra.

Uma distribuição de energia espectral, ou simplesmente, **distribuição energética do espectro** é um método que caracteriza o espectro por meio de percentuais energéticos, nos quais a localização de pontos divisores do espectro é um tipo de discriminador acústico. O primeiro passo do método é encontrar as **Funções de Distribuição Energética Acumulada** (DEA) que representem o envelope espectral E.

A DEA de um sinal contínuo $E(\omega) = \sum_{\forall k} \psi(\omega, a_k, \mu_k, \sigma_k)$ corresponde ao sinal definido em cada frequência ω, definido por

$$\mathrm{DEA}_E(\omega) = \frac{\int_{\omega_{\min}}^{\omega} E(\omega) df}{\int_{\omega_{\min}}^{\omega_{\max}} E(\omega) df}, \tag{3.28}$$

onde $\omega_{\min}$ e $\omega_{\max}$ são os valores mínimo e máximo da representação espectral E, respectivamente. O cálculo de uma DEA é razoavelmente simples, uma vez que as bases radiais ψ (especialmente as trigonométricas truncadas, já que $\int \cos(ax)dx = \frac{1}{a}\sin(ax)$) são funções analíticas facilmente integráveis. Entretanto, para o cálculo de bases mais complexas, ou mesmo, sobre sinais discretos com $f_{\max} - f_{\min} + 1$ amostras para os quais não se conhece uma representação contínua, uma versão

discreta é apresentada abaixo:

$$\mathrm{DEA}_E(f) = \frac{\sum_{k=f_{\min}}^{f} \{E(k)\} - E(f_{\min})}{\sum_{k=f_{\min}}^{f_{\max}} \{E(k)\} - E(f_{\min})}, \tag{3.29}$$

onde $f_{\min}$ e $f_{\max}$ são respectivamente as frequências mínima e máxima de E.

Note que em ambos os casos uma DEA_E definida em cada valor de frequência respectiva ao ponto f corresponde a um valor percentual de energia espectral acumulada entre 0 e 1 desde a frequência $f_{\min}$ até f. Por esta razão se exige que a função E a ser transformada seja estritamente positiva.

Por outro lado, a distribuição em frequência normalizada (DFN) associa cada valor de energia normalizada $\varepsilon \in [0, 1]$ a uma frequência ω de modo que

$$\mathrm{DFN}(\varepsilon) = \frac{\int_0^{\varepsilon} \mathrm{DEA}(\varepsilon)d\varepsilon}{\int_0^{1} \mathrm{DEA}(\varepsilon)d\varepsilon}(\omega_{\max} - \omega_{\min}) + \omega_{\min}.$$

Analogamente, uma versão discretizada da equação acima é derivada como

$$\mathrm{DFN}(e) = \frac{\sum_{k=e_0}^{e} \mathrm{DEA}(k)}{\sum_{k=e_0}^{e_1} \mathrm{DEA}(k)}(f_{\max} - f_{\min}) + f_{\min}. \tag{3.30}$$

No caso discreto, as amostras entre e_0 e e_1 correspondem a energias linearmente espaçadas entre 0 e 1. A fim de derivar uma versão interpolada da DFN discreta, os valores de e podem ser obtidos por meio de interpolação linear entre valores adjacentes.

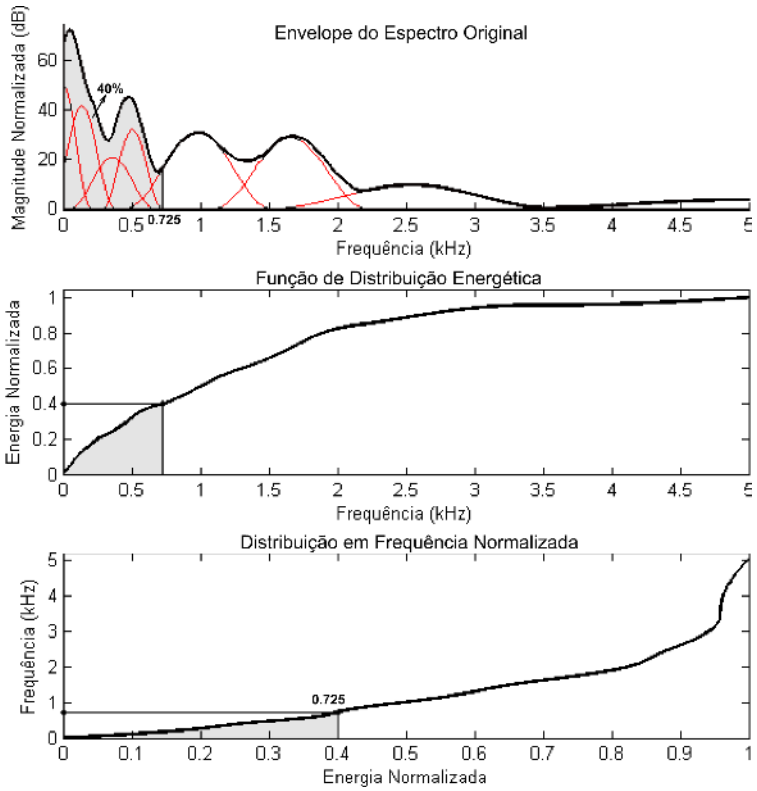

Figura 3.28: *Relação entre a função energética acumulada e o espectro original.*

Tomando-se uma sequência discreta de percentuais energéticos $[e_0, e_1, \ldots, e_n]$ de valores linear-

mente espaçados entre 0 e 1, a Distância em Frequência Normalizada entre dois envelopes discretos E_x e E_y é definido como a distância euclidiana

$$\mathbf{d}_F(x,y) = \sum_{e=e_0}^{e_n} (\mathrm{DFN}_x(e) - \mathrm{DFN}_y(e))^2, \tag{3.31}$$

onde $\mathrm{DFN}_x(e)$ e $\mathrm{DFN}_y(e)$ são, respectivamente, a DFN de S_x e S_y.

Tais transformações associam um percentual energético do sinal original E desde sua origem ao valor em frequência correspondente no limiar desta proporção, conforme é observado na Figura 3.28. Neste caso, $f_e = 0.5 \approx 750$ Hz corresponde a uma frequência na qual a energia entre 0 e f_e é aproximadamente 40% da energia total do espectro.

Com base nestes conceitos introduzidos, o método de Deformação em Frequência Normalizada (NFW) entre dois espectros positivos é implementado pelo Algoritmo 3.29 seguinte. Tal método se fundamenta em transformar um vetor a baseado no alinhamento entre proporções energéticas em frequências normalizadas em X e Y, a fim de criar uma função de empenamento para posterior interpolação.

Algoritmo 3.29 $a \leftarrow \mathtt{NFW}(v, X, Y)$

▷ *Gera vetores artificiais válidos iniciados com* 0
1 $E_X \leftarrow [0, X - \min\{X\}]$
2 $E_Y \leftarrow [0, Y - \min\{Y\}]$

▷ *Toma as distribuições em frequência normalizada de cada E*
3 **for** $f = 1 : L^* + 1$
4 $\quad F_X(f) \leftarrow \mathrm{DFN}_{E_X}(f)$
5 $\quad F_Y(f) \leftarrow \mathrm{DFN}_{E_Y}(f)$
6 **end for**

▷ *Obtém o núcleo da função de empenamento*
7 $F = \text{interpolação}(F_X, F_Y, 1 : L^* + 1)$

▷ *Realiza a deformação em frequência*
8 $a \leftarrow \text{interpolação}(F, v, 2 : L^* + 1)$
9 **return**(a)

Graças à normalização das Linhas 1 e 2, a transformação admite funções negativas como padrões (X e Y). O método de deformação em frequência basicamente realiza duas interpolações seguidas, uma para definir a função F e outra para realizar a deformação em frequência propriamente. O fato de a deformação ocorrer em frequência se explica pela definição do núcleo de empenamento definido na Linha 7 e posteriormente aplicado na Linha 8.

Dada a função de empenamento, é possível se derivar pelo menos um método de empenamento aplicado ao problema de conversão de voz. Esta proposta implementa a função genérica $\mathcal{T}_{loc}$ chamada no Alg. 3.28.

Como observado, a função NFW necessita de padrões a fim de realizar a transformação espectral. As médias das classes origem X e destino Y são tidas como tais padrões, as quais conformarão os coeficientes espectrais armazenados no vetor harmônico v pertencente à classe X, de modo que o resultado da transformação o categorize como se fosse pertencente à classe Y.

Infelizmente, a aplicação da função NFW pura e simplesmente não é capaz de realizar uma transformação espectral com alto grau de confiança, devido ao fato que o simples empenamento de frequências não ajusta as amplitudes (harmônicas) de cada componente do vetor. A fim de solucionar este problema, a deformação em frequência é aplicada sobre a diferença entre o vetor v e a média de X, de modo que

$$\mathcal{T}_{loc}^{\mathrm{NFW}}(v, X, Y) = \mu_Y + NFW\left(v - \mu_X, \mu_X, \mu_Y\right). \tag{3.32}$$

Neste caso, as amplitudes espectrais são convertidas via transformação linear, enquanto que os desvios são empenados em frequência.

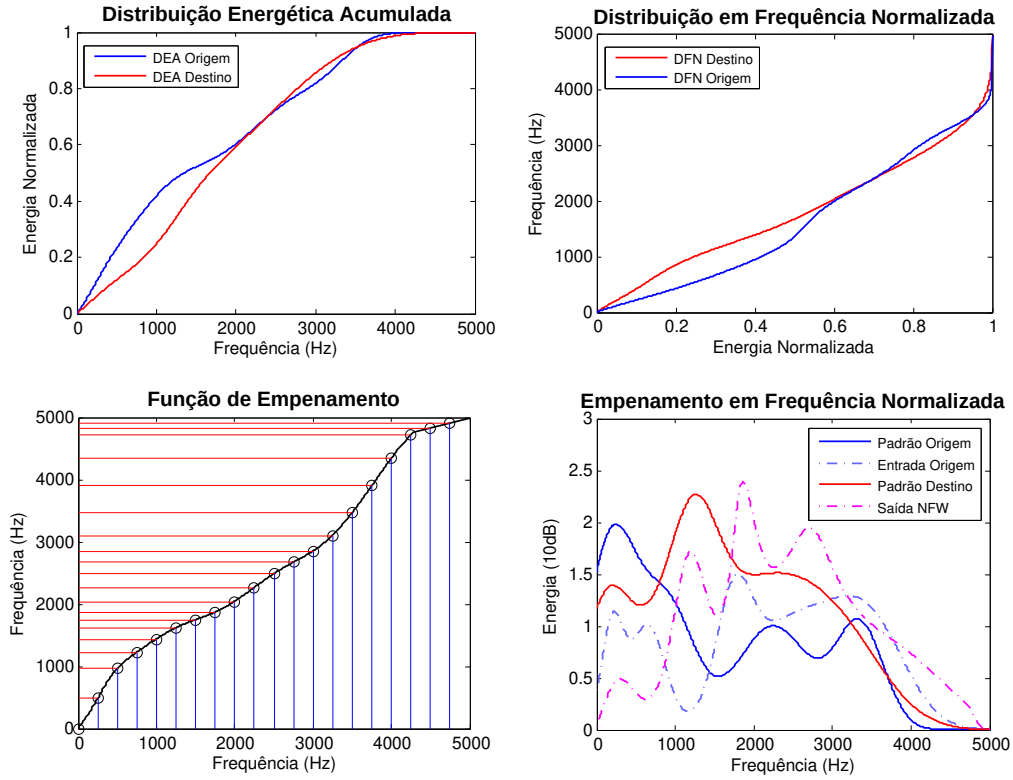

Figura 3.29: *Conversão local usando deformação (empenamento) em frequência normalizada.*

A Figura 3.29 exibe um exemplo no qual ambas as transformações NFW são utilizadas para converter um envelope espectral sintético. Esta figura ilustra todo o processo de transformação espectral, desde a fase de alinhamento energético (ambos os gráficos superiores da figura), a qual define a função de empenamento em frequência (inferior-esquerda), que segundo a Eq. 3.32, é aplicada sobre os desvios do sinal de entrada usando os padrão descriminados pela legenda. Note que o sinal de saída se relaciona ao padrão destino de modo similar ao sinal de entrada com relação ao padrão origem. Tal conversão aplicada sobre a parte harmônica não interage com a energia total de cada segmento. Assim, um estágio de pós-processamento no qual as energias são equilibradas é realizado no último procedimento do método de transformação.

Pós-Processamento Corretivo ao Nível da Sentença

Ao final das transformações, três tipos de atualizações do sinal são aplicadas ao vetor de amplitudes harmônicas (no último caso, também sobre as amplitudes estocásticas) a fim de garantir a reconstrução natural do segmento de saída.

O primeiro ajuste trata de uma interpolação ponderada entre os coeficientes harmônicos, assim como a energia total de cada segmento usando os M_s segmentos anteriores a este. Neste caso, a contribuição dos segmentos anteriores decai exponencialmente (na base 2). Tipicamente, poucos segmentos consecutivos são utilizados, e neste trabalho em particular serão usados os 5 segmentos anteriores ao segmento corrente. Evidentemente, o uso de um decaimento linear, mesmo diante de um número não muito grande de segmentos anteriores, já degradaria substancialmente o vetor final por causa do efeito de suavização excessiva que seria introduzido no sinal modificado. Por esta razão a função de decaimento exponencial foi utilizada.

O segundo ajuste corresponde ao reajuste da variância global do sinal transformado ao nível da sentença. Pelo fato de que as amplitudes harmônicas Ψ_A estão dissociadas das energias totais (armazenadas pelo vetor Ψ_{E_0}) destas componentes harmônicas, as transformações locais dessas componentes harmônicas podem devolver espectros harmônicos descontínuos o longo do tempo. Tal provável descontinuidade temporal entre segmentos consecutivos deve ser corrigida a fim de aumentar o grau de naturalidade do sinal reconstruído.

Para contornar este problema, os vetores

$$a(k,\cdot) = \Psi_{E_0}^k \cdot \Psi_A^k$$

são tomados para todo k, antes do estágio de conversão local. Então, ao final deste estágio, outro vetor

$$b(k,\cdot) = [\Psi_{E_0}^k]' \cdot [\Psi_A^k]'$$

é novamente tomado para todo k, no entanto, com os valores atualizados de energia total e amplitudes harmônica para cada segmento. A partir de então, um ajuste global ao nível do segmento é aplicado sobre os envelopes espectrais, tomando-se os desvios padrão de a e b de modo que

$$[\Psi_A^k]'' \leftarrow \frac{\mathbb{E}\{b\} + \sigma_a(\sigma_b)^{-1}\left(b(k,\cdot) - \mathbb{E}\{b\}\right)}{[\Psi_{E_0}^k]'}, \tag{3.33}$$

onde

$$\sigma_a = \mathrm{diag}(\mathrm{std}\{a\}) \qquad \text{e} \qquad \sigma_b = \mathrm{diag}(\mathrm{std}\{b\})$$

são os desvios padrão dos conjuntos a e b representados por matrizes diagonais, respectivamente. Neste caso, $[\Psi_A^k]''$ corresponde ao vetor final de amplitudes harmônicas do segundo pós-processamento.

Um último estágio da normalização nesta fase de pós-processamento implica em re-equilibrar a razão energética entre as partes harmônicas e estocásticas do sinal reconstruídos, pelo menos ao nível da sentença. Neste caso, a razão entre a soma das energias relativas às partes harmônicas de um sinal de voz antes e depois da conversão, é suficiente para atualizar a parte estocástica a fim de manter inalterado o percentual energético harmônico-estocástico. Como ambas estas partes são armazenadas por coeficientes espectrais harmônicos (a e b), ao final de todas as transformações, um

ajuste energético é aplicado à parte estocástica do sinal de voz, de modo que

$$[\Psi_S^k]'' \leftarrow \frac{\sum_{\forall x}\sum_{\forall y} b(x,y)}{\sum_{\forall x}\sum_{\forall y} a(x,y)} \cdot [\Psi_S^k]', \tag{3.34}$$

onde $[\Psi_S^k]'$ corresponde aos coeficientes estocásticos do k-ésimo segmento de voz após a conversão (global). Deste modo, o equilíbrio energético entre as partes harmônicas e estocásticas do sinal convertido é reestabelecido.

A implementação da função de pós-processamento conforme descrita anteriormente é realizada pela Algoritmo 3.30.

Algoritmo 3.30 $\Psi \leftarrow$ `pós_processamento`(Ψ, a, G^X, G^Y)

$M_s \leftarrow 3$
$N_s \leftarrow$ dimensão(Ψ)

▷ *Obtém vetor b*
for $k = 1 : N_s$
  if $\Psi_{F_0}^k > 0$ **then**
    $b(k,\cdot) \leftarrow \Psi_{E_0}^k \cdot \Psi_A^k$
  end if
end for

▷ *Primeira normalização: Interpolação usando segmentos consecutivos*
for $k = 1 : N_s$
  if $\Psi_{F_0}^k > 0$ **then**
    $\Psi_A^k \leftarrow \frac{2}{M_s(M_s+1)} \sum_{i=1}^{M_s} (M_s - i + 1)\Psi_A^{k-i}$
    $\Psi_{E_0}^k \leftarrow \frac{2}{M_s(M_s+1)} \sum_{i=1}^{M_s} (M_s - i + 1)\Psi_{E_0}^{k-i}$
  end if
end for

▷ *Segunda normalização: Restaura variância global da sentença*
for $k = 1 : N_s$
  if $\Psi_{F_0}^k > 0$ **then**
    $\sigma_a = \text{diag}(\text{std}\{a\})$
    $\sigma_b = \text{diag}(\text{std}\{b\})$
    $\Psi_A^k \leftarrow \frac{1}{\Psi_{E_0}^k}\mathbb{E}\{b\} + \sigma_a(\sigma_b)^{-1}\left(b(k,\cdot) - \mathbb{E}\{b\}\right)$
  end if
end for

▷ *Terceira normalização: Reequilíbrio energético entre as partes harmônica e estocástica do sinal.*
$\Psi_S^k \leftarrow \frac{\sum_{\forall x}\sum_{\forall y} b(x,y)}{\sum_{\forall x}\sum_{\forall y} a(x,y)} \cdot [\Psi_S^k],\ \forall k$
return(Ψ)

Cada uma destas normalizações desempenha um papel específico na finalização do sinal transformado. A primeira, por exemplo, garante uma transição dos níveis de energia mais suavemente entre segmentos adjacentes, enquanto que a segunda balanceia as contribuições de cada uma das componentes Ψ_A^k.

Um conjunto de métodos foram apresentados e necessitam ser validados experimentalmente, tanto sob o aspecto quantitativo usando métricas objetivas quanto sob o aspecto qualitativo a partir de entrevistas de opinião. Este é o tema do capítulo subsequente.

Capítulo 4

Resultados Experimentais

A validação experimental é um importante estágio dentro de um projeto computacional, e especialmente no caso da conversão de voz, devido à carência de métodos e métricas puramente teóricas capazes de avaliar apropriadamente os resultados da conversão em termos perceptuais. Por via de regra, todos os experimentos foram realizados dentro de um cenário hipotético controlado, o qual envolve as duas línguas mais faladas no mundo: o inglês e o espanhol.

O problema da conversão de voz é composto por uma série de outros subproblemas inter-ligados em série (alinhamento de classes, seleção de classes dado um vetor, transformação, entre outras), de modo que a avaliação é dividida em partes independentes, relativas a cada subproblema. Desta feita, este capítulo é composto por três seções, sendo uma delas simplesmente introdutória, a qual apresenta os dados (os corpora) usados no experimento. As outras duas seções são especializadas em realizar experimentos objetivos (métricos) e subjetivos (perceptuais). Esta seção está organizada de modo que

- A Seção 4.1 introduz o módulo experimental, especificando principalmente os conjuntos de dados usados no experimento.
- A Seção 4.2 trata de descrever todos os experimentos sob o ponto de vista objetivo usando medidas de distância entre as amostras comparadas. Nesta seção se testará as principais funcionalidades do sistema isoladamente: a representação do sinal de voz, a clusterização/-mapeamento dos dados, assim como a conversão propriamente dita. Neste caso, para que seja possível o uso das métricas objetivas, as bases de dados utilizadas são paralelas.
- A Seção 4.3 apresenta a parte subjetiva dos experimentos, onde os resultados finais da conversão de voz inter-linguística são submetidos a uma bateria de testes de opinião.

4.1 Definições Gerais

Duas questões fundamentais estão relacionadas com a avaliação da conversão de voz: a primeira se associa à ideia de **similaridade** entre o timbre da voz do falante origem convertido e o timbre de voz do falante destino; e a segunda está relacionada com a **qualidade** do sinal transformado, envolvendo aspectos sonoros como a inteligibilidade e a naturalidade da sentença convertida, entre outros. Este projeto se propõe a avaliar tais questões objetivamente e subjetivamente.

Avaliar aspectos de similaridade espectral, dado que se possa estabelecer pares de sentenças alinhadas no tempo, é uma tarefa que pode ser realizada com sucesso a partir de medidas de distância perceptuais, o que define o conjunto de testes **objetivos**. Entretanto, avaliar outros aspectos mais gerais, tais como a naturalidade da voz sintética, é uma tarefa eminentemente humana, ou seja, exige que seja realizada uma bateria de testes **subjetivos**.

Basicamente, os testes objetivos são responsáveis por determinar quais as melhores opções de configuração dos sistemas de conversão de voz a serem comparados posteriormente na fase de testes subjetivos. Em ambos os casos, uma base de dados é requerida. No caso dos testes objetivos, o banco de dados de voz ainda deve ser rotulado para fins de verificação. A influência da prosódia na avaliação (principalmente subjetiva) precisa ser de certa forma isolada do experimento, e neste caso, o projeto propôs realizar os experimentos dentro de um cenário controlado em termos prosódicos. Esse cenário hipotético conta com um conjunto de sentenças com pouca variação no contorno do pitch e pouca expressividade de intonação.

Este trabalho contou com o apoio da Universidade Politécnica da Catalunha, a qual disponibilizou um corpus privado utilizado no projeto TC-STAR [23], concedido gentilmente pelo Dr. Antonio Bonafonte. O projeto TC-STAR é um projeto europeu interdisciplinar que envolveu variados sistemas de processamento de voz. A grande vantagem deste corpus é que as sentenças que o compõem foram gravadas paralelamente entre os diversos falantes, de modo a facilitar o alinhamento, com um controle prosódico bem rigoroso no qual cada falante ouviu a mesma sentença gravada por outros falantes, e tentou imitar o padrão melódico dos outros. Além disso, tal banco fornece um conjunto de sentenças rotuladas automaticamente por um software de reconhecimento de voz usado no projeto TC-STAR. Uma outra peculiaridade desta base de voz é que suas respectivas sentenças foram gravadas em três idiomas distintos sob as mesmas condições: Inglês Europeu, Espanhol Europeu e Mandarim. Infelizmente, não foi possível encontrar uma base de dados com especificações semelhantes para o português. Por motivos de familiaridade nos testes subjetivos, somente as partes **espanhola** e **inglesa** do corpus TC-STAR foram utilizadas.

Totalizando oito corpora, categorizados em locutores bilíngues (Espanhol e Inglês) dois do sexo feminino e dois do sexo masculino, o corpus TC-STAR contém aproximadamente 10 horas de gravações de sentenças em áudio a uma taxa de amostragem de 96 kHz e quantizadas a 24 bits/amostra, praticamente sem ruído de fundo (nos experimentos seguintes os sinais foram re-amostrados a uma taxa de 16 kHz por questões de eficiência no cálculo dos coeficientes espectrais).

Os corpora foram catalogados de modo que os dois locutores do sexo feminino são numerados como '75' e '76', e os dois do sexo masculino como '79' e '80'; as línguas por sua vez são discriminadas como 'ES' (Espanhol) e 'EN' (Inglês). Assim, o corpus 'ES_75' corresponde ao corpus em Espanhol da locutora '75'.

Toda vez em que o texto se referir ao uso dos corpus da base TC-STAR dentro dos experimentos, considere que 50 sentenças de aproximadamente 5 segundos foram tomadas para compor o mesmo. Tal restrição da base de dados pequena visa melhor qualificar o sistema em situações extremas com poucos dados de treinamento.

4.2 Avaliação Objetiva

A realização de tais avaliações experimentais objetivas tem por meta demonstrar empiricamente quais dentre os métodos propostos são melhor qualificados a compor a parte subjetiva do experimento. Basicamente, três classes de métodos são considerados fundamentais no processo de conversão de voz: (1) a parametrização e quantização dos vetores acústicos dos segmentos de voz processados na base de treinamento, (2) a clusterização destes vetores em classes acústicas para posterior mapeamento entre corpus distintos e (3) a conversão de segmentos de voz de um sinal de entrada com base nas classes acústicas previamente alinhadas. Evidentemente, cada bateria de testes é aplicada em cada uma destas classes de métodos separadamente.

4.2.1 Parametrização e Quantização

Conforme apresentado na Seção 3.3 do capítulo anterior, a decomposição espectral proposta por este trabalho toma cada segmento de voz e o representa como uma soma de bases radiais (banco de filtros). Sendo assim, o seguinte experimento tem por meta não apenas mostrar a acurácia do ajuste espectral obtido pelos métodos propostos, mas também mostrar a estabilidade temporal dos parâmetros que representam cada função de base $\psi(a_k, \mu_k, \sigma_k)$, onde a_k, μ_k e σ_k são as amplitudes, posições e larguras de banda do k-ésimo segmento de voz quantizado. Algumas propriedades de clusterização do sinal de fala também serão investigadas usando estes parâmetros.

Três sinais de voz foram gravados à parte para estes testes particulares. Cada sinal de voz é composto por cinco segmentos com as principais vogais da língua espanhola ([a], [e], [i], [o] e [u]) pronunciadas por três locutores diferentes (do sexo masculino). O sistema segmentou cada sinal de entrada em pequenos trechos de 256 amostras cada, para os quais foram estimados os envelopes harmônicos e estocásticos a partir do Modelo Harmônico/Estocástico (HSM) apresentado na Seção 3.2. Aproximadamente 900 segmentos vozeados, para cada locutor, foram considerados nestes experimentos. Ao final, a média e o desvio padrão entre os três locutores foram tomadas e amostradas em cada gráfico para cada medida. A fim de comparar com outros modelos clássicos, os coeficientes LPC e os coeficientes Cepstrais foram também estimados. Em ambos os casos, os envelopes harmônicos e estocásticos são estimados diretamente dos picos harmônicos e do log do espectro de magnitude, respectivamente. No caso particular deste experimento, a estimação dos coeficientes LPC é realizada a partir da resolução do seguinte sistema de equações:

$$\begin{pmatrix} R_0 & R_1 & \cdots & R_{p-1} \\ R_1 & R_0 & \cdots & R_{p-2} \\ \vdots & \vdots & \ddots & \vdots \\ R_{p-1} & R_{p-2} & \cdots & R_0 \end{pmatrix} \cdot \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{pmatrix} = \begin{pmatrix} -R_1 \\ -R_2 \\ \vdots \\ -R_p \end{pmatrix} \tag{4.1}$$

onde se sabe que os valores de R podem ser aproximados de modo que

$$R_n \approx \int_{-\pi}^{\pi} |X(\omega)|^2 e^{j\omega n} d\omega = \frac{1}{2} \sum_{l=-L}^{L} A_l^2 \cos(l\omega_0 n), \forall n,$$

dada uma decomposição harmônica de um espectro original X com amplitudes A_l e frequência fundamental ω_0.

De modo parecido, os coeficientes cepstrais c_i são determinados a partir da equação

$$\log|X(f)| = c_0 + 2\sum_{i=1}^{p}\cos(\frac{2\pi if}{SR}),$$

onde SR é a taxa de amostragem. Costuma-se eliminar a componente c_0 pelo fato de que a mesma é desnecessária na modelagem da forma do espectro.

Em ambos os casos (LPC e Cepstrum), os respectivos coeficientes podem ser obtidos através da otimização por mínimos quadrados. Além disso, pelo fato da matriz de autocorrelação apresentada na abordagem LPC ser uma matriz Toeplitz, a inversa da mesma pode ser eficientemente obtida pelo método de Levinson-Durbin [168].

Quanto aos métodos propostos, as seguintes janelas são usadas como bases paramétricas: *Hann*, *Nuttall*, *Blackman-Harris*, *Blackman-Nuttall* e Gaussiana. Os métodos usados na comparação são o Método Guloso de estimação paramétrica de bases com larguras de banda e posições fixas, o qual obtém as amplitudes via otimização por mínimos quadrados, bem como os métodos com os parâmetros livres $[a, \mu, \sigma]$, $[a, \mu]$, $[a, \sigma]$ e $[a]$ (Seção 3.3). O parâmetro de limiar que estabelece o critério de parada, utilizado por cada um esses métodos no algoritmo de ajuste de base, foi de $\epsilon = 10^{-5}$. Assim, os envelopes espectrais obtidos por estes métodos foram comparados com os envelopes espectrais estimados via LPC e Cepstrum. O número de coeficientes utilizados em todos os métodos foi de 24 coeficientes cada, com uma taxa de amostragem do sinal de 16 kHz.

O projeto experimental levou em conta os seguinte aspectos:

1. a medida de acurácia, tomada como a distância entre o log do espectro de magnitude reconstruído e o envelope harmônico interpolado originalmente;
2. a estabilidade na evolução dinâmica temporal de cada parâmetro do modelo;
3. as distâncias entre segmentos que pertencem a uma mesma classe fonética artificial; e
4. as distâncias entre os centroides de cada uma dessas classes.

As comparações foram feitas segmento a segmento entre o sinal reconstruído $\hat{s}$ e o sinal original s. A medida de acurácia considera a média de todos os ajustes espectrais de cada envelope modelado em relação ao envelope real interpolado, o qual foi obtido da parte harmônica proveniente do modelo HSM. Neste experimento, uma versão normalizada da distorção espectral adaptada à escala MEL foi tomada entre tais envelopes. Considere que S^k corresponde ao espectro original do k-ésimo segmento de voz do sinal de entrada s, com $k \in \{1, \ldots, N\}$. Então, a acurácia do ajuste espectral é calculada de modo que

$$\varepsilon_{SD} = \frac{1}{N}\sum_{k=1}^{N}\left|S^k_{\log} - \hat{S}^k_{\log}\right|$$

onde $S_{\log}$ é uma normalização logarítmica de S definida como

$$S_{\log} = 10\log_{10}(S + \varepsilon_0)$$

onde ε_0 é um valor que define o chão da normalização, o qual foi definido como 10^{-5}.

A estabilidade da evolução temporal de cada parâmetro é medida como a soma de distâncias entre valores consecutivos dos parâmetros em cada par de segmentos adjacentes. Esta medida está

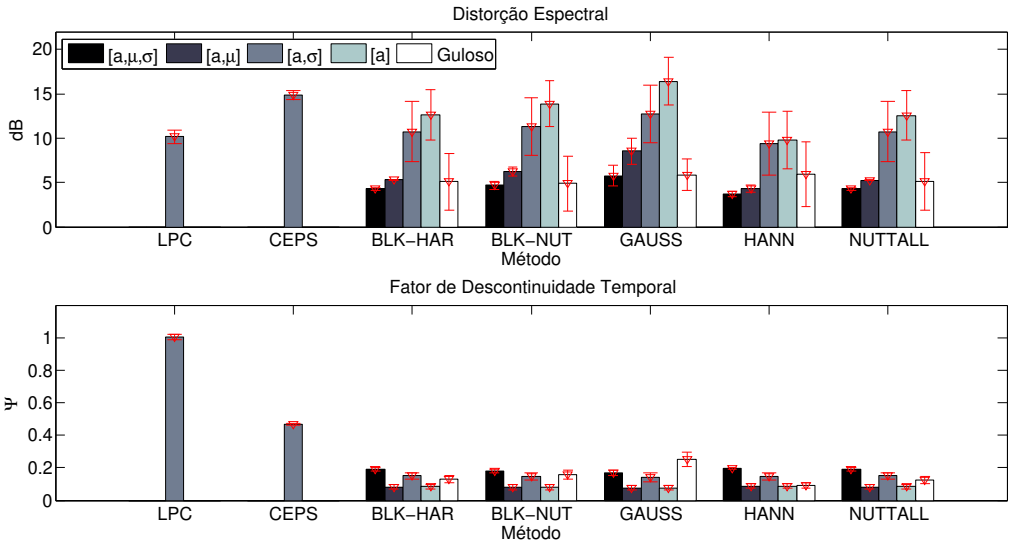

Figura 4.1: *Distorções Espectrais e Taxas de Descontinuidade do Sinal de Entrada*

associada com a taxa de descontinuidade do sinal ao longo de sua evolução temporal. Supondo que $\mathbf{w}^k = (a^T, \mu^T, \sigma^T)^T$ seja um vetor de amplitudes, frequências centrais e larguras de banda relativas ao segmento de voz $\hat{S}^k$, então a taxa de descontinuidade do sinal s é medida como

$$\Psi = \frac{1}{N-1} \sum_{k=2}^{N} \left| \mathbf{w}^k - \mathbf{w}^{k-1} \right|,$$

ou seja, uma maior estabilidade temporal corresponde a valores menores de Ψ.

A Figura 4.1 apresenta valores de distorção espectral ε_{SD} e a taxa de estabilidade Ψ para cada método. Dentre as bases avaliadas, um visível destaque é dado à janela de *Hann*. A Figura 4.1 indica que esta base combinada ao método com $[a, \mu, \sigma]$ livres apresenta melhor ajuste, com baixas taxas de distorção espectral e estabilidade temporal aceitável. Além disso, se observa que o método $[a, \mu]$ é uma alternativa mais acurada que $[a, \sigma]$ em representações usando poucos coeficientes. Embora o método guloso não tenha obtido bons valores, o mesmo merece destaque dado o baixo custo computacional necessário para realizar a estimação.

Uma outra propriedade esperada destes modelos de representação está relacionada ao seu uso na forma de vetores de característica quantizados na tarefa de clusterização em classes fonéticas artificiais. Tipicamente uma classe fonética artificial usa o centroide espectral como chave primária. A medida denominada *within-class scatter* [44] para uma classe com centroide $\mathcal{C}_i$ é definida como

$$\gamma_i = \sum_{k=1}^{N_i} \left| \mathbf{w}^k - \mathcal{C}_i \right|^2.$$

A fim de medir o quão longe está uma classe fonética de outra, uma variante do *Discriminante Linear de Fisher* foi adotada, a qual é definida como

$$\mathbb{J} = \frac{\sum_{i=1}^{I} N_i \left| \mathcal{C}_i - \mathbf{m} \right|^2}{\sum_{j=1}^{I} \gamma_j},$$

onde $\mathbf{m}$ é a média global de todas as amostras. Observe que maiores valores de $\mathbb{J}$ implicam em

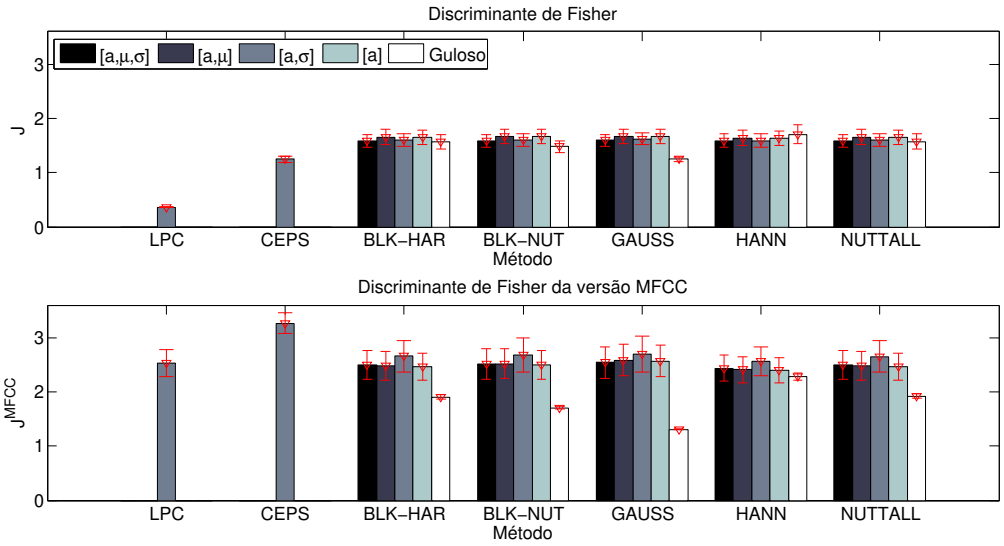

Figura 4.2: *Discriminantes de Fisher dos coeficientes espectrais e suas versões adaptadas à abordagem MFCC.*

maior separabilidade entre as classes $\mathcal{C}$.

Uma vez que em classificação de fala os coeficientes MFCC são frequentemente usados, dada uma versão MFCC dos vetores $\mathbf{w}^k$, espera-se que algum ganho seja obtido na discriminação linear. A Figura 4.2 exibe uma versão *MFCC-like* derivada da representação paramétrica proposta por este trabalho. Neste caso, os valores de amplitude a_k (os quais já se encontram amostrados na escala logarítmica) são multiplicados por σ_k a fim de obter um fator fortemente correlacionado com os valores de energia por banda da abordagem padrão MFCC. Finalmente, a transformada do cosseno discreta (DCT) destes valores é calculada. Note na figura que houve um aumento significativo dos valores de $\mathbb{J}$ em relação ao discriminante obtido pelos dados originais. O melhor discriminante dentre as propostas deste trabalho foram obtidas com o método $[a, \sigma]$, no qual não se percebe uma diferença significativa com relação à base utilizada.

4.2.2 Clusterização e Mapeamento

A clusterização, bem como o mapeamento de classes fonéticas artificiais, são estágios do procedimento de conversão de voz os quais requerem uma atenção particular, uma vez que no contexto do trabalho os corpus não necessariamente são alinhados e também não são rotulados. A dificuldade em alinhar fonemas correspondentes entre dois corpora quaisquer é o principal entrave na reconstrução de uma sentença de voz convertida com alto grau de similaridade e naturalidade em relação ao sinal objetivo (destino).

É de se esperar que as classes acústicas obtidas na primeira etapa do módulo clusterizador correspondam diretamente à agrupamentos que possuam algum significado em termos fonéticos. Neste caso, uma visualização de cada classe acústica obtida em termos de fonemas reais é uma ferramenta essencial para o controle da experimentação do módulo de alinhamento.

Uma vez que o uso de um corpus rotulado é requerido, foi utilizado um par de corpora femininos em Espanhol: 'ES_75' (acima) e 'ES_76' (abaixo) da base TC-STAR para este teste. A Figura 4.3 apresenta cada classe fonética artificial como um 'gráfico de pizza' que exibe as respectivas proporções em fonemas rotulados de vogais simples. A fim de facilitar a visualização, somente as cinco

vogais dominantes da língua espanhola foram exibidas. Observe que existe uma forte correlação entre as regiões formânticas apresentadas na Figura 3.15 do capítulo anterior e tais gráficos.

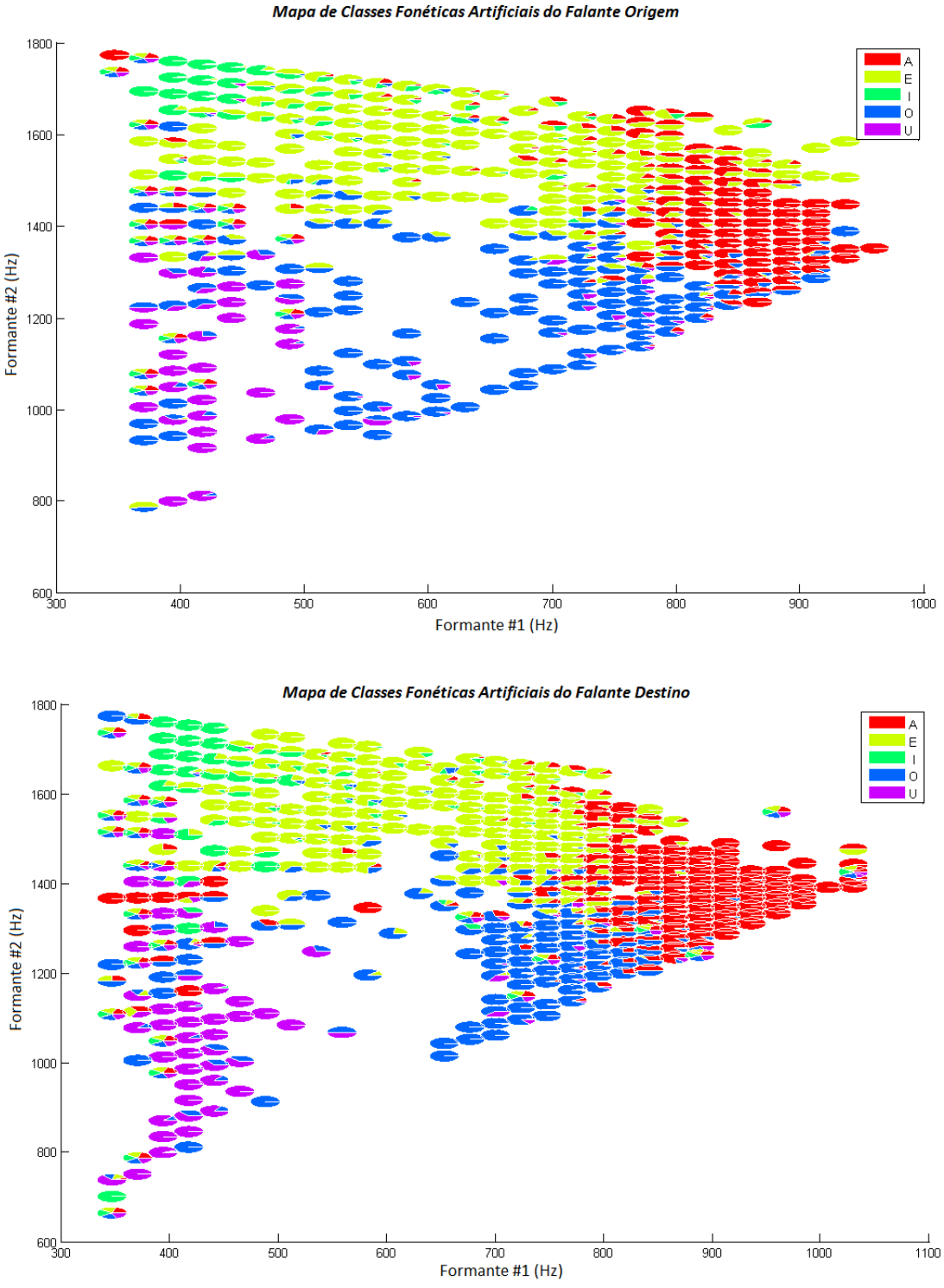

Figura 4.3: *Principais regiões formânticas dos corpus origem 'ES_75' e destino 'ES_76'.*

Um dos problemas do alinhamento não-rotulado é determinar as relações de reciprocidade entre classes fonéticas de ambos os mapas. Note que o mapa de cima tem uma distribuição da vogal [o] mais esparsa do que o inferior, o que dificulta muito o alinhamento.

	Corpus Origem (valores em %)						Corpus Destino (valores em %)						α %
	A	E	I	O	U	peso	A	E	I	O	U	peso	
CFA #1	**94.61**	0.81	–	4.58	–	*3.09*	**97.33**	–	–	2.67	–	*3.42*	**98.91**
CFA #2	**98.28**	1.15	–	0.57	–	*2.90*	**99.26**	0.37	–	0.37	–	*2.48*	**99.61**
CFA #3	**93.91**	1.60	–	4.49	–	*2.60*	**99.19**	–	–	0.81	–	*2.26*	**97.89**
CFA #4	**97.74**	0.97	–	1.29	–	*2.58*	**92.83**	–	–	7.17	–	*2.42*	**97.65**
CFA #5	0.38	–	–	**98.09**	1.53	*2.18*	–	–	–	**91.61**	8.39	*1.42*	**97.26**
CFA #6	–	0.42	**99.58**	–	–	*1.96*	–	1.28	**98.72**	–	–	*2.15*	**99.66**
CFA #7	**91.42**	0.86	–	7.73	–	*1.94*	**96.43**	0.45	–	3.13	–	*2.05*	**98.00**
CFA #8	–	–	**100.00**	–	–	*1.91*	–	1.94	**98.06**	–	–	*1.42*	**99.23**
CFA #9	–	1.33	**95.11**	0.44	3.11	*1.87*	–	2.59	**97.41**	–	–	*2.12*	**98.58**
CFA #10	**87.02**	1.44	–	11.54	–	*1.73*	**97.74**	–	–	2.26	–	*2.02*	**95.71**
CFA #11	**97.52**	1.98	–	0.50	–	*1.68*	**100.00**	–	–	–	–	*1.50*	**99.01**
CFA #12	**88.44**	1.51	–	7.54	2.51	*1.66*	0.88	–	–	**94.69**	4.42	*1.03*	**64.37**
CFA #13	**70.16**	–	–	26.18	3.66	*1.59*	**87.74**	–	–	12.26	–	*1.42*	**92.97**
CFA #14	34.88	1.16	–	**59.88**	4.07	*1.43*	**100.00**	–	–	–	–	*1.18*	**73.95**
CFA #15	**60.98**	0.61	–	32.32	6.10	*1.36*	**88.50**	–	–	9.73	1.77	*1.03*	**88.99**
CFA #16	**98.75**	1.25	–	–	–	*1.33*	**97.33**	2.67	–	–	–	*1.37*	**99.43**
CFA #17	6.45	–	–	**83.23**	10.32	*1.29*	**77.14**	–	–	14.29	8.57	*0.64*	**71.72**
CFA #18	**98.56**	1.44	–	–	–	*1.16*	**100.00**	–	–	–	–	*0.92*	**99.42**
CFA #19	3.73	–	–	**84.33**	11.94	*1.11*	6.90	–	–	**88.51**	4.60	*0.80*	**97.06**
CFA #20	–	3.05	**95.42**	–	1.53	*1.09*	–	0.58	**99.42**	–	–	*1.56*	**98.40**
CFA #21	–	–	**96.12**	3.88	–	*1.07*	–	3.61	**96.39**	–	–	*1.52*	**98.45**
CFA #22	**55.04**	–	–	42.64	2.33	*1.07*	1.25	–	–	**92.50**	6.25	*0.73*	**78.48**
CFA #23	18.42	–	–	**66.67**	14.91	*0.95*	35.19	–	–	**64.81**	–	*0.49*	**93.29**
CFA #24	–	–	–	**94.34**	5.66	*0.88*	–	–	–	**100.00**	–	*0.63*	**97.74**
CFA #25	1.03	**98.97**	–	–	–	*0.81*	5.95	**85.71**	8.33	–	–	*0.77*	**94.70**
CFA #26	**78.95**	–	–	20.00	1.05	*0.79*	**100.00**	–	–	–	–	*0.88*	**91.58**
CFA #27	**57.45**	42.55	–	–	–	*0.78*	**95.65**	4.35	–	–	–	*0.84*	**84.72**
CFA #28	2.22	**60.00**	37.78	–	–	*0.75*	–	**83.18**	16.82	–	–	*0.98*	**90.73**
CFA #29	–	14.29	**74.03**	1.30	10.39	*0.64*	–	1.83	**98.17**	–	–	*1.00*	**90.34**
CFA #30	34.21	**65.79**	–	–	–	*0.63*	**97.06**	2.94	–	–	–	*0.62*	**74.86**
CFA #31	**82.35**	17.65	–	–	–	*0.57*	**82.14**	17.86	–	–	–	*0.51*	**99.92**
CFA #32	–	–	–	**100.00**	–	*0.56*	–	–	–	**91.49**	8.51	*0.43*	**96.60**
CFA #33	–	1.52	**98.48**	–	–	*0.55*	–	2.74	**97.26**	–	–	*0.67*	**99.51**
CFA #34	–	**98.48**	–	1.52	–	*0.55*	–	**93.55**	6.45	–	–	*0.57*	**97.42**
CFA #35	**95.45**	4.55	–	–	–	*0.55*	**96.36**	3.64	–	–	–	*0.50*	**99.64**
CFA #36	–	**76.56**	23.44	–	–	*0.53*	6.06	**68.18**	25.76	–	–	*0.60*	**96.65**
CFA #37	–	–	–	**98.33**	1.67	*0.50*	4.17	–	–	**75.00**	20.83	*0.44*	**90.67**
CFA #38	5.45	**94.55**	–	–	–	*0.46*	2.27	**97.73**	–	–	–	*0.40*	**98.73**
CFA #39	**77.78**	9.26	–	12.96	–	*0.45*	32.26	–	–	**67.74**	–	*0.28*	**78.09**
CFA #40	44.44	**55.56**	–	–	–	*0.45*	37.93	**58.62**	3.45	–	–	*0.27*	**97.39**
CFA #41	–	13.46	**84.62**	–	1.92	*0.43*	–	**89.90**	10.10	–	–	*0.90*	**69.43**
CFA #42	1.96	**88.24**	9.80	–	–	*0.42*	–	**85.25**	14.75	–	–	*0.56*	**98.02**
CFA #43	–	–	–	**94.00**	6.00	*0.42*	–	–	–	**96.67**	3.33	*0.27*	**98.93**
CFA #44	4.08	–	–	**89.80**	6.12	*0.41*	–	–	–	**100.00**	–	*0.23*	**95.92**
CFA #45	**97.96**	2.04	–	–	–	*0.41*	**95.24**	4.76	–	–	–	*0.19*	**98.91**
CFA #46	2.08	**97.92**	–	–	–	*0.40*	–	**100.00**	–	–	–	*0.60*	**99.17**
CFA #47	–	6.52	**93.48**	–	–	*0.38*	–	31.82	**68.18**	–	–	*0.40*	**89.88**
CFA #48	46.51	**53.49**	–	–	–	*0.36*	**80.00**	20.00	–	–	–	*0.27*	**86.60**
CFA #49	2.38	**90.48**	7.14	–	–	*0.35*	–	**98.31**	1.69	–	–	*0.54*	**96.87**
CFA #50	11.90	33.33	–	**47.62**	7.14	*0.35*	–	17.39	8.70	**73.91**	–	*0.21*	**86.00**
CFA #51	**80.95**	14.29	–	4.76	–	*0.35*	**93.33**	6.67	–	–	–	*0.27*	**95.05**
CFA #52	**78.57**	19.05	–	–	2.38	*0.35*	**90.91**	9.09	–	–	–	*0.30*	**95.06**
CFA #53	2.50	–	–	**97.50**	–	*0.33*	**60.61**	–	–	39.39	–	*0.30*	**76.76**
CFA #54	**94.87**	2.56	–	–	2.56	*0.32*	**97.30**	2.70	–	–	–	*0.68*	**98.97**
CFA #55	–	**97.37**	2.63	–	–	*0.32*	3.23	**70.97**	25.81	–	–	*0.28*	**89.44**
CFA #56	–	**86.84**	5.26	7.89	–	*0.32*	–	**94.59**	2.70	2.70	–	*0.34*	**96.90**
CFA #57	–	**64.86**	29.73	5.41	–	*0.31*	–	2.33	**97.67**	–	–	*1.18*	**72.82**
CFA #58	5.41	**89.19**	5.41	–	–	*0.31*	–	**92.50**	7.50	–	–	*0.37*	**97.84**
CFA #59	2.70	2.70	–	**83.78**	10.81	*0.31*	11.54	–	–	**84.62**	3.85	*0.24*	**96.13**
CFA #60	27.03	**72.97**	–	–	–	*0.31*	25.93	**55.56**	18.52	–	–	*0.25*	**92.59**
CFA #61	–	11.43	**88.57**	–	–	*0.29*	–	27.36	**71.70**	0.94	–	*0.97*	**93.25**
CFA #62	–	**100.00**	–	–	–	*0.29*	–	**98.18**	1.82	–	–	*0.50*	**99.27**
CFA #63	–	**79.41**	8.82	11.76	–	*0.28*	–	**98.31**	1.69	–	–	*0.54*	**92.44**
CFA #64	48.48	–	–	**51.52**	–	*0.27*	**55.17**	–	–	44.83	–	*0.27*	**97.32**
CFA #65	–	**78.13**	21.88	–	–	*0.27*	2.08	**87.50**	10.42	–	–	*0.44*	**95.42**
CFA #66	–	**50.00**	**50.00**	–	–	*0.27*	4.76	**52.38**	42.86	–	–	*0.19*	**97.14**
CFA #67	3.13	3.13	–	**87.50**	6.25	*0.27*	11.11	–	–	**88.89**	–	*0.25*	**96.25**
CFA #68	–	**100.00**	–	–	–	*0.26*	–	**90.32**	–	9.68	–	*0.28*	**96.13**
CFA #69	–	**96.67**	3.33	–	–	*0.25*	–	**97.37**	2.63	–	–	*0.35*	**99.72**
CFA #70	3.33	**96.67**	–	–	–	*0.25*	–	**94.74**	5.26	–	–	*0.17*	**97.89**
CFA #71	10.00	6.67	3.33	**80.00**	–	*0.25*	11.11	**44.44**	–	**44.44**	–	*0.08*	**84.44**
CFA #72	**76.67**	10.00	–	13.33	–	*0.25*	**100.00**	–	–	–	–	*0.20*	**90.67**
CFA #73	3.45	**93.10**	3.45	–	–	*0.24*	–	48.15	**51.85**	–	–	*0.25*	**80.64**
CFA #74	–	**89.66**	6.90	3.45	–	*0.24*	–	**94.37**	5.63	–	–	*0.65*	**98.12**
CFA #75	–	**100.00**	–	–	–	*0.24*	10.34	**79.31**	10.34	–	–	*0.27*	**91.72**
CFA #76	–	**96.55**	3.45	–	–	*0.24*	7.69	**61.54**	30.77	–	–	*0.24*	**85.99**
CFA #77	3.45	**86.21**	10.34	–	–	*0.24*	10.00	**70.00**	20.00	–	–	*0.09*	**93.52**
CFA #78	6.90	**93.10**	–	–	–	*0.24*	–	14.29	14.29	**71.43**	–	*0.06*	**65.71**
CFA #79	6.90	**93.10**	–	–	–	*0.24*	4.35	**95.65**	–	–	–	*0.21*	**98.98**
CFA #80	20.69	**72.41**	–	6.90	–	*0.24*	5.00	**60.00**	10.00	25.00	–	*0.18*	**88.76**
CFA #81	**72.41**	27.59	–	–	–	*0.24*	13.04	**82.61**	4.35	–	–	*0.21*	**76.25**
CFA #82	**100.00**	–	–	–	–	*0.24*	**100.00**	–	–	–	–	*0.16*	**100.00**
CFA #83	–	–	**100.00**	–	–	*0.23*	–	–	**100.00**	–	–	*0.33*	**100.00**
CFA #84	–	**100.00**	–	–	–	*0.23*	7.41	**77.78**	14.81	–	–	*0.25*	**91.11**
CFA #85	3.57	**89.29**	7.14	–	–	*0.23*	–	**88.89**	–	11.11	–	*0.08*	**95.56**
CFA #86	–	–	–	**100.00**	–	*0.23*	–	–	–	**100.00**	–	*0.08*	**100.00**
CFA #87	–	3.57	–	**85.71**	10.71	*0.23*	–	31.25	–	**62.50**	6.25	*0.15*	**88.93**
CFA #88	–	**85.19**	7.41	7.41	–	*0.22*	–	**68.00**	24.00	8.00	–	*0.23*	**93.13**
CFA #89	**44.44**	37.04	–	18.52	–	*0.22*	**68.75**	31.25	–	–	–	*0.15*	**90.28**
CFA #90	18.52	**77.78**	3.70	–	–	*0.22*	**45.83**	**45.83**	8.33	–	–	*0.22*	**87.22**
CFA #91	**59.26**	22.22	–	7.41	11.11	*0.22*	**86.67**	13.33	–	–	–	*0.14*	**89.04**
CFA #92	**77.78**	22.22	–	–	–	*0.22*	**68.00**	32.00	–	–	–	*0.23*	**96.09**
CFA #93	22.22	**77.78**	–	–	–	*0.22*	**94.74**	5.26	–	–	–	*0.35*	**70.99**
CFA #94	–	**84.62**	3.85	11.54	–	*0.22*	–	**95.24**	4.76	–	–	*0.38*	**95.38**
CFA #95	–	**92.31**	7.69	–	–	*0.22*	–	**100.00**	–	–	–	*0.30*	**96.92**
												β_1	**89.30**
												β_2	**95.17**

Tabela 4.1: *Indicadores de alinhamento entre classes fonéticas artificiais (CFA) de ambos os corpora origem e destino.*

Outro fato interessante pode ser percebido pela compactação do trapézio formântico nos casos de vozes masculinas, o qual pode ser acompanhado pela Figura 4.4. Nesta figura é possível visualizar a maior parte dos pontos no mapa abaixo dos 700 Hz. Este fenômeno ocorre porque a anatomia do aparelho fonador masculino (como a região supraglótica e caixas de ressonância, por exemplo) modela o som de forma a concentrar um porção maior de energia sonora nas regiões graves, fato este que reflete em uma conversão de voz com menos similaridade na fase final.

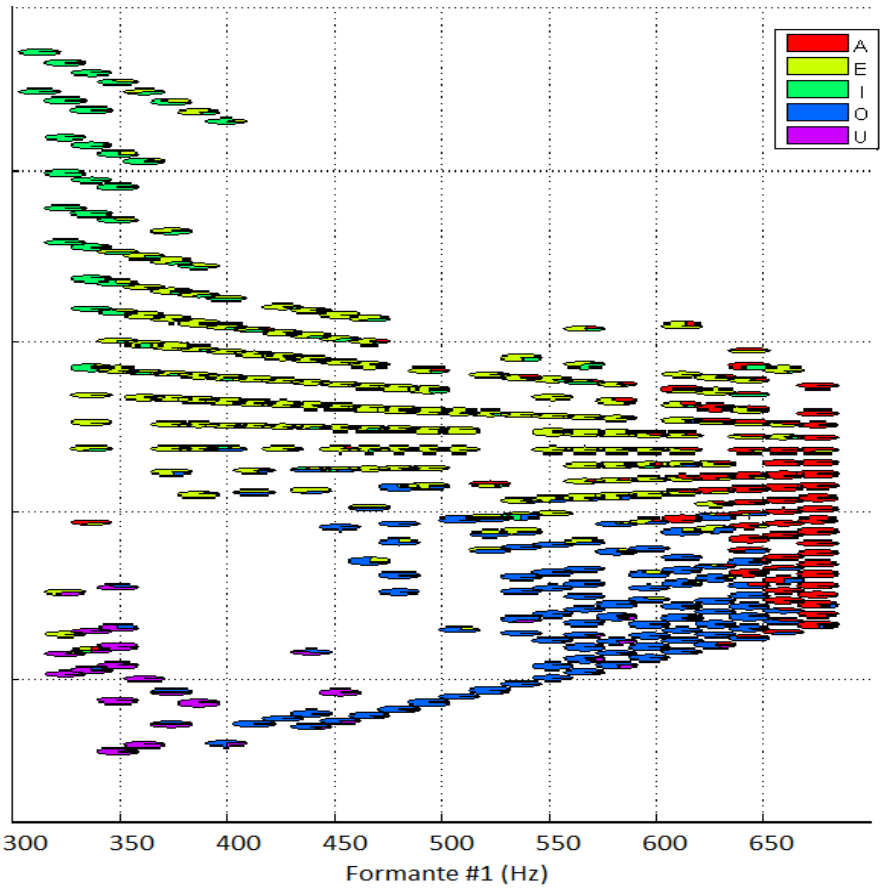

Figura 4.4: *Principais regiões formânticas do corpus 'ES_ 79'.*

Neste caso, o alinhamento de mapas fonéticos é ainda mais complicado, uma vez que ambos os mapas não são normalizados quanto ao trato vocal. Espera-se que o método de emparelhamento aplicado sobre uma versão normalizada dos mapas, conforme proposto na Seção 3.3, contorne parte deste problema.

Para avaliar a qualidade do emparelhamento, principalmente em casos de falantes de sexo diferente, dois indicadores são propostos para a avaliação da acurácia do alinhamento. O primeiro indicador corresponde à taxa de acerto de alinhamento interno, denotada por α, entre uma classe origem O e uma classe destino D, e definida como

$$\alpha(C,D) = 1 - \frac{\sum_{\forall f \in fnm} (\mathbf{p}(C.\{f\}) - \mathbf{p}(D.\{f\}))}{\sum_{\forall f \in fnm} \mathbf{p}(C.\{f\})}, \tag{4.2}$$

onde $\mathbf{p}(C.\{f\})$ e $\mathbf{p}(D.\{f\})$ correspondem às proporções do mesmo fonema f (rotulado) dentro das classes C e D, respectivamente. Como as proporções de fonemas consistem em um vetor normalizado de valores, ou seja, $\sum_{\forall f \in fnm} \mathbf{p}(C.\{f\}) = 1$, o denominador da Eq. 4.2 é desconsiderado.

A Tabela 4.1 mostra a correspondência entre cada classe fonética artificial CFA #k e as respectivas proporções de fonemas rotulados que as compõem. Tais classes estão devidamente alinhadas entre os corpora espanhóis 'ES_76' e 'ES_79' (feminino para masculino). A tabela está dividida em três partes, separadas por uma barra dupla. Informações de uma mesma linha correspondem às informações relativas às correspondentes classes pareadas. O primeiro grupo de colunas agrupadas corresponde às informações do corpus origem, enquanto que o segundo grupo é referente ao corpus destino. A última coluna da tabela (denominada α) apresenta a taxa de acerto do alinhamento conforme definida na Eq. 4.2. Todos os valores tabelados estão amostrados em valores percentuais.

Cada célula interna da Tabela 4.1 relaciona uma CFA a um vetor de proporções relativas às vogais rotuladas. Em especial, as Colunas 6 e 12, nomeadas *peso*, mostram as densidades de cada classe fonética artificial, definidas como a quantidade de vetores em cada classe (cardinalidade da classe) dividida pelo número total de vetores que compõem todo o corpus. Por razões de visualização, somente as 95 classes fonéticas artificiais (CFA) mais densas (em relação ao corpus origem) foram exibidas. Ao observar as duas últimas linhas da tabela, se nota a presença de duas taxas de alinhamento global entre ambos os corpora: $\beta_1 = \mathbf{89.30}$ e $\beta_2 = \mathbf{95.17}$. Tais indicadores estão relacionados com o acerto do alinhamento global entre todas as classes de dois corpora distintos. O primeiro indicador é definido como a média de todos os alinhamentos internos α, ou seja,

$$\beta_1 = \mathbb{E}\{\alpha\},$$

enquanto que o segundo corresponde à mesma média em uma versão ponderada pelas densidades de cada CFA do corpus origem, ou seja,

$$\beta_2 = \frac{\sum_{\forall k=1}^{N} (w(C_k) \cdot w(D_k) \cdot \alpha_k)}{\sum_{\forall k=1}^{N} (w(D_k) \cdot w(C_k))},$$

onde $w(C_k)$ ($w(D_k)$) é a densidade da classe C_k (D_k) associada à coluna 'peso' da Tabela 4.1, com $k = [N]$. Vale salientar que, em ambos os casos, as taxas β_1 e β_2 foram calculadas usando todas as classes do corpus, e não somente as 95 exibidas na Tabela 4.1.

Se ao invés de realizar o alinhamento de classes fonéticas artificiais o módulo de mapeamento realizasse o alinhamento de classes k-verossímeis, as quais foram calculadas no capítulo anterior (Seção 3.4), o resultado do alinhamento seria $\beta_1 = \mathbf{84.32}$ e $\beta_2 = \mathbf{85.63}$[1]. A razão pela qual estes resultados são piores se fundamenta muito provavelmente no fato de haver um número muito maior de disparidades entre classes k-verossímeis do que entre CFA, levando a mais erros de mapeamento fonético. Por exemplo, se um corpus possui mais fonemas da vogal [i] que um outro corpus, possivelmente os fonemas excedentes sejam alinhados a fonemas da vogal [e], aumentando o erro de alinhamento global. A filtragem do alinhamento proposto na Seção 3.4.2, a qual elimina pares de classes distantes, não foi considerada no experimento.

Outra observação importante se refere ao tipo do alinhamento. Considere que em vez de realizar o alinhamento a partir do algoritmo de emparelhamento (ver Seção 3.5), o módulo de mapeamento optasse por utilizar uma estratégia gulosa de busca por mínimos globais, isto é, para cada classe representada pelo ponto $x \in C_O$ no corpus origem encontrar uma classe representada por $y \in C_D$ da classe destino tal que a distância Euclidiana $\mathbf{d}(x, y)$ fosse mínima. Em outras palavras,

[1] Estes valores foram devidamente calculados usando exatamente os mesmos corpora utilizados na Tabela 4.1.

o procedimento do alinhamento guloso primeiramente normaliza o conjunto dos dados quanto à média e matriz de covariância de modo análogo ao método proposto com o emparelhamento, e iterativamente atribui para cada CFA do corpus origem uma CFA do destino a qual melhor se ajusta segundo critérios de erro mínimo.

As taxas de alinhamento globais[1] usando o método guloso acima descrito são $\beta_1 = 89.23$ e $\beta_2 = 94.87$ para CFA e $\beta_1 = 87.55$ e $\beta_2 = 89.67$ para alinhamento entre conjuntos k-verossímeis. As taxas relativamente altas obtidas pelo método guloso são explicadas devido ao fato de que a hipótese com respeito ao alinhamento bijetor entre ambas as classes foi desprezada, conforme comentado na Seção 3.5. Por esta razão, somente o alinhamento de CFA por emparelhamento foi considerado neste trabalho.

Os bons resultados indicam um alinhamento satisfatório, se tratando de um método para alinhamento de corpus de gêneros distintos. Sendo assim, passemos à experimentação algorítmica de cada método do módulo de transformação.

4.2.3 Transformação Espectral

Para a realização deste tipo de experimento é necessário que se tenha em mão tanto a sentença convertida quanto a sentença do falante destino correspondente. Neste tipo de teste se alinham no tempo ambas as sentenças convertida e destino, de modo que seja possível calcular um tipo de distância espectral entre pares de segmentos de voz correspondentes foneticamente. O algoritmo usado no alinhamento dos dados é o clássico algoritmo Dynamic Time Warping – DTW [15] adaptado ao sistema HSM. Deste modo, considere que para cada **segmento de voz da entrada**, caracterizado pelo vetor Ψ, existe um respectivo **segmento ótimo da saída**, representado por $\bar{\Psi}$, o qual pertence à sentença pronunciada paralelamente pelo locutor destino.

Basicamente, dois indicadores são imprescindíveis nesta etapa de avaliação, os quais analisam os dois principais aspectos da fase de transformação:

- A taxa de erro de classificação do **módulo de seleção de classes fonéticas artificiais**, que é definida a partir de um vetor acústico quantizado do corpus origem. Neste caso, o erro entre uma classe fonética e um vetor acústico é menor à medida que tal vetor é "caracterizado" pela classe.

- A taxa de erro do **método de transformação**, que considera a diferença entre o vetor transformado e o vetor ótimo de saída.

A taxa de qualidade da conversão dos parâmetros de prosódia, ou seja, os contornos de pitch e energia do sinal, não serão levados em consideração neste trabalho, por se tratarem de medidas difíceis de quantificar matematicamente e aferir objetivamente, além de não pertencerem ao escopo central do trabalho, que é a conversão espectral.

Módulo de Seleção de Classes

O objetivo do módulo de conversão é, dado um vetor acústico quantizado $v = \Psi_A^k$ referente ao k-ésimo segmento de voz do sinal de entrada, encontrar um conjunto de classes fonéticas artificiais C_o do corpus origem que melhor representa v. Cada classe $c \in C_o$, por sua vez, possui um peso proporcional à proximidade de v em relação à c, denominado **grau de pertinência** de v em relação

a c e denotado por $\mathbf{w}(v, c)$. O valor de $\mathbf{w}(v, c)$ está diretamente ligado à função de distância usada neste método de seleção.

Como definido no capítulo anterior na Seção 3.6, um número variado de funções de distância podem ser utilizadas no módulo de seleção, tais como a distância de Mahalanobis [142]

$$\mathbf{d}_M(v, c) = \sqrt{(v - \mu_c)^T (\Sigma_c)^{-1} (v - \mu_c)}, \tag{4.3}$$

a distorção espectral obtida pela norma Euclidiana

$$\mathbf{d}_S(v, c) = |\text{mel2freq}(v - \mu_c)|, \tag{4.4}$$

onde tanto v quanto μ_c, por estarem quantizados na escala logarítmica, necessitam ser re-amostrados na escala linear em frequência (usando uma função `mel2freq`), e também a distorção mel-cepstral adaptada, definida como

$$\mathbf{d}_C(v, c) = |\text{dct}[v] - \text{dct}[v]|, \tag{4.5}$$

análoga à formulação da Eq. 3.18 do capítulo anterior.

Vale salientar que, dos vetores quantizados via MFCC, são considerados somente a primeira terça parte (16 neste caso) dos coeficientes, com exceção do primeiro, que corresponde à componente DC do espectro. Tal poda é justificada pelo fato de que se deseja comparar apenas os coeficientes mais relevantes do espectro, referentes ao formato mais suavizado do espectro, desconsiderando assim, detalhes irrelevantes como pequenas variações/descontinuidades espectrais ou mesmo o nível DC do espectro.

Uma associação direta de cada vetor acústico v a uma classe c também pode ser realizada, a qual usa todo o ferramental de distribuição formântica e mapas fonéticos, conforme descrito na Seção 3.5. Nesta abordagem, o valor $w_v(c)$ corresponde ao fator de participação de cada classe c (associada a um ponto) no mapa fonético artificial do vetor v.

No caso da seleção via mapa fonético, o grau de pertinência é definido de forma direta $\mathbf{w}(v, c) = \mathbf{d}(v, c)^2$. Já no caso das associações por meio de distâncias, uma normalização é necessária, uma vez que se deseja atribuir maiores graus de pertinências a menores valores de distância. Neste caso, o grau de pertinência do vetor v em relação a uma classe c é definido como

$$\mathbf{w}(v, c) = N(v, c)^4 = \left(1 - \frac{\mathbf{d}_{[C|S]}(v, c) - \min_{\forall k}\{\mathbf{d}_{[C|S]}(v, C_k)\}}{\max_{\forall k}\{\mathbf{d}_{[C|S]}(v, C_k)\} - \min_{\forall k}\{\mathbf{d}_{[C|S]}(v, C_k)\}}\right)^4, \tag{4.6}$$

onde $[C|S]$ corresponde ou à distância Euclidiana ou à distorção mel-cepstral, $L^* = 48$ é a dimensão de v, para todas as classes C_k que compõem o corpus origem. A distância Euclidiana da formulação acima está aplicada sobre vetores amostrados na escala logarítmica em amplitude, ou seja, corresponde à distorção espectral clássica.

Particularmente, pelo fato da distância de Mahalanobis levar em consideração a matriz de covariância, é conveniente adotar outro modo de calcular o grau de pertinência, normalizando a distância segundo a norma dessa matriz, de modo que

$$\mathbf{w}^{Mahal}(v, c) = \frac{1}{2\pi^{\frac{L^*}{2}} |\Sigma_c|^{\frac{1}{2}}} e^{-0.5\mathbf{d}_M(v,c)}, \tag{4.7}$$

Desta forma, espera-se que os problemas de suavização excessiva da transformação sejam controlados pela função exponencial normalizada desta formulação. A função de normalização adotada neste trabalho foi a formulação quártica que define a Equação 4.6, escolhida a partir de um pequeno experimento realizado no Apêndice A.3. No entanto, qualquer outro tipo de função de normalização pode ser considerada.

Um tópico importante para seleção da classe fonética que melhor representa v_k consiste em saber qual dentre as funções de distância é a mais apropriada para definir o grau de pertinência de um vetor acústico qualquer em relação a uma classe fonética artificial. As funções a serem comparadas são:

(1) Mahal$^{\exp}$: A função de Mahalanobis (Eq. 4.3) com a normalização exponencial que leva em conta a matriz de covariância de cada conjunto.

(2) Mahal^4: A função de Mahalanobis com a normalização quártica padrão.

(3) SDist: A distorção espectral (Eq. 4.4) com normalização quártica adaptada aos vetores parametrizados na escala linear.

(4) MFCCa: A distorção mel-cepstral (Eq. 4.5) com normalização quártica sobre a escala MEL.

(5) MpFon: Seleção direta via mapa fonético artificial.

Para a realização do teste, todos os corpora ingleses ('EN_75', 'EN_76', 'EN_79' e 'EN_80') da base TC-STAR foram submetidos a um procedimento experimental básico. Tal procedimento experimental realiza buscas por classes c_i, $i \in [M_c]$ no corpus origem que apresentem maiores graus de pertinência $\mathbf{w}(v_k, c_i)$, para todos os vetores $v_k = \Psi_A^k$ do sinal de entrada. A partir das classes c_i, um vetor artificial $\bar{v_k}$ relativo a v_k é gerado de tal forma que

$$\bar{v_k} = \frac{\sum_{i=1}^{M_c} \mathbf{w}(v_k, c_i) \cdot \mu_{c_i}}{\sum_{i=1}^{M_c} \mathbf{w}(v_k, c_i)}.$$

Finalmente, tal vetor é utilizado para calcular a taxa de erro do módulo de seleção, definida como $\Gamma(M_c) = \sum_{k=1}^{N} \gamma(v_k, M_c)$, onde N é o número de segmentos harmônicos do sinal de entrada e

$$\gamma(v_k, M_c) = |\bar{v_k} - v_k|,$$

com $v^k = \Psi_A^k$ correspondente ao próprio vetor quantizado da sentença origem. A fim de se equiparar os resultados de todos os corpora e compor a média de todos eles, uma versão normalizada é considerada de modo que $\Gamma'(M_c) = \frac{\Gamma(M_c)}{N}$. O motivo pelo qual as bases inglesas foram tomadas se justifica pelo maior grau de dificuldade de seleção das classes, dada a riqueza fonética da língua inglesa.

Pela Tabela 4.2 é possível deduzir que a distorção mel-cepstral MFCCa é, com certa margem, a métrica mais apropriada para realizar a seleção de classes fonéticas dado um vetor de característica quantizado, confirmando assim a preferência de alguns autores que executam tarefas afins [167; 190].

M_c	$\Gamma(M_c)$ para cada métrica de seleção				
	Mahal^{exp}	Mahal^{4}	SDist	MFCCa	MpFon
1	45.5967	45.8756	35.5571	32.0711	44.0629
2	44.8892	44.6060	35.1184	31.8614	43.7563
3	44.6455	44.3611	35.2614	31.9912	43.6502
4	44.5856	44.3007	35.4691	32.1928	43.5885
5	44.5709	44.4285	35.7422	32.4670	43.4317
10	44.5490	45.2606	36.8632	33.7320	43.1257
15	44.5381	46.2456	37.8503	34.5775	43.1152
20	44.5368	47.0980	38.7029	35.2879	43.1139
25	44.5363	47.8089	39.2716	35.9990	43.1134
30	44.5360	48.5201	39.9828	36.5679	43.1131
35	44.5359	49.0891	40.4096	36.9947	43.1130
40	44.5358	49.6581	40.9786	37.4215	43.1129
45	44.5358	50.2273	41.4055	37.8483	43.1129
50	44.5358	50.5118	41.8323	38.1329	43.1129
MÉDIA	44.5705	47.6695	39.0200	**35.6351**	43.1910

Tabela 4.2: *Taxas de erro do módulo de seleção considerando a função distância.*

Além da métrica usada na seleção fonética, outros parâmetros também favorecem uma seleção robusta de classes acústicas em relação ao erro de alinhamento. O número M_c de classes que contribuem de forma ponderada para composição da função de transformação é um outro aspecto a ser considerado. O valor de M_c deve ser escolhido levando em conta que um valor M_c muito grande gera uma transformação bastante natural, mas pouco similar ao sinal destino, devido à generalização da função de transformação causada pela contribuição de muitas classes distintas, ao passo que um valor M_c muito pequeno gera um sinal com alto índice de similaridade mas baixa naturalidade, dada a descontinuidade causada pela introdução de ruído dos erros de classificação do método seletor de classes acústicas.

Segundo a Tabela 4.2, para a maioria das métricas de seleção, se observa o *trade-off* entre os problemas de suavização e ajuste excessivo da transformação, com exceção da distância de Mahalanobis. Os valores da tabela referentes à medida MFCCa foram mostrados na Figura 4.5 usando um escala logarítmica em ambos os eixos a fim de perceber melhor este fenômeno. O valor *default* do sistema é $M_c = 15$, mas este é um parâmetro flexível que pode ser redefinido sem grandes problemas.

O valor médio calculado na tabela anterior toma todos os valores inteiros de 1 a 50, além dos valores mostrados. Dando prosseguimento aos testes, a medida MFCCa é usada por *default* como base para o método de seleção de classe fonética, o qual será utilizado na seção subsequente.

Função de Transformação

Nesta fase de testes, a **função de transformação** toma um par de sentenças paralelas devidamente alinhadas pelo algoritmo DFW. Neste caso, para cada vetor quantizado de entrada $v = \Psi_A$, espera-se que o mesmo possua a menor distorção espectral possível em relação ao vetor de saída ótimo $v^* = \bar{\Psi}_A$ da sentença destino.

Considerando-se que um vetor acústico v está associado a um conjunto de classes C_o do falante origem ponderadas por graus de pertinência definidos pela distorção espectral, a estrutura de mapeamento $@M$ determinada pelo módulo de Mapeamento encontra o conjunto de classes alinhadas

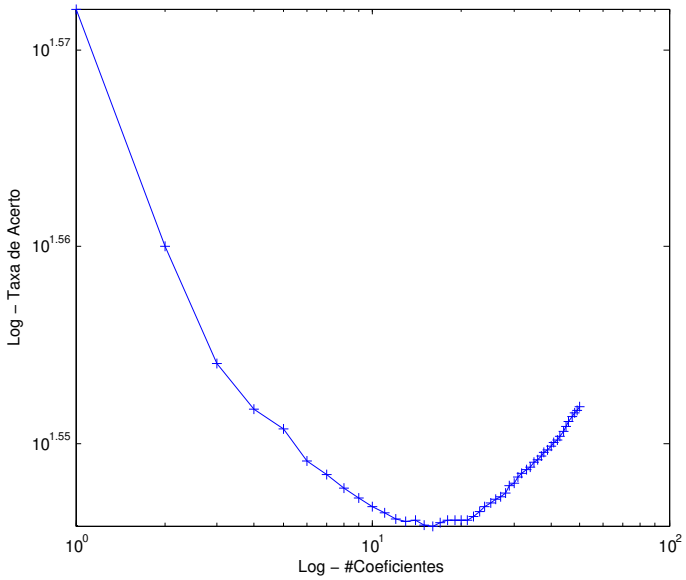

Figura 4.5: *Dados da quarta coluna da Tabela 4.2 exibidos em escala logarítmica.*

C_d correspondentes a cada C_o no corpus destino.

A **função de transformação** toma o vetor v e os momentos estatísticos relativos a C_o e C_d, e conforma v ao vetor ótimo de saída v^* usando os três métodos de transformação descritos na Seção 3.6:

(1) a transformação linear com matriz de covariância completa (LT-Full);

(2) a transformação linear usando somente a diagonal da matriz de covariância, isto é, as variâncias estatísticas (LT-Diag);

(3) a transformação proposta por este trabalho, denominada Deformação em Frequência Normalizada (NFW),

(4) a ressíntese do sinal destino a partir da remontagem ponderada dos centroides de cada classe selecionada no corpus destino, estratégia esta denominada "codebook" neste experimento.

A métrica de seleção fonética utilizada para o experimento foi, conforme dito, a distorção melcepstral $MFCCa$ adaptada aos vetores quantizados.

O arcabouço experimental desta seção segue basicamente os mesmo padrões do experimento realizado para escolher os métodos de seleção de classes da seção anterior. Ou seja, realizam-se buscas por classes do falante origem com maiores graus de pertinência avaliadas em v. Neste caso, o vetor artificial $\bar{v_k}$ relativo ao k-ésimo segmento de voz quantizado por v_k é definido de tal forma que

$$\bar{v_k} = \frac{\sum_{c \in C_o} w(c) \mathcal{T}_{loc}(\Psi_A^k, c, M(c))}{\sum_{c \in C_o} w(c)},$$

onde $w(c) = \mathbf{d}_C(v, c)$ corresponde ao grau de pertinência (MFCCa) do vetor v_k em relação à classe c, e $@\mathcal{T}_{loc}$ é uma das funções de transformação listadas anteriormente na Seção 3.6.

Ao final da transformação, cada vetor transformado $\bar{v_k}$ é comparado com seu respectivo vetor ótimo v^* usando a distância Euclidiana clássica

$$\epsilon_{\mathcal{T}}^{c}(\bar{v_k}, v^*) = |\bar{v_k} - v^*|,$$

onde M_c (*default* 15) corresponde à quantidade de classes usadas na ponderação, isto é, $M_c = |C_o|$ e v^* é o vetor origem ótimo da correspondente sentença paralela alinhada. Observe que o passo final do algoritmo de transformação, o passo de acabamento do sinal de saída conforme apresentado na Seção 3.6, não foi considerado neste experimento, com o efeito de isolar o erro da transformação dos três métodos apresentados deste módulo (opcional) de pós-processamento.

Um teste mais extensivo foi realizado neste caso usando todos os corpora ingleses e espanhóis, os quais foram combinados entre si de modo a cobrir todos os casos de conversão de voz intra-linguística em relação aos gêneros, isto é, $M \rightarrow M$, $M \rightarrow F$, $F \rightarrow M$ e $F \rightarrow F$. A conversão intra-linguística, embora não seja o escopo do trabalho, é usada nesta etapa pela comparação usando corpora paralelos[2]. Em termos de corpus, os pares avaliados foram:

- ES_75 → ES_76: Conversão $ES(F \times F)$ – língua espanhola;
- ES_76 → ES_79: Conversão $ES(F \times M)$ – língua espanhola;
- ES_79 → ES_80: Conversão $ES(M \times M)$ – língua espanhola;
- ES_80 → ES_75: Conversão $ES(M \times F)$ – língua espanhola;
- EN_75 → EN_76: Conversão $EN(F \times F)$ – língua inglesa;
- EN_76 → EN_79: Conversão $EN(F \times M)$ – língua inglesa;
- EN_79 → EN_80: Conversão $EN(M \times M)$ – língua inglesa;
- EN_80 → EN_75: Conversão $EN(M \times F)$ – língua inglesa.

Para facilitar a avaliação sob um ponto de vista global, um conjunto de histogramas é mostrado na Figura 4.6, os quais apresentam os valores médios por segmento de voz das taxas de distorção espectral entre o espectro do sinal modificado reconstruído (com ambas as partes harmônicas e estocásticas embutidas) e o sinal ótimo temporalmente alinhado via DTW. Evidentemente, o erro inerente ao alinhamento com método DTW está adicionado no erro total estimado, o que não afeta a comparação visto que o erro é o mesmo para todos os métodos comparados.

Pela Figura 4.6 se observa uma maior dificuldade de conversão entre gêneros distintos, principalmente quando se trata da conversão (F)eminino para (M)asculino. Tal dificuldade em se recompor uma voz feminina é explicada pela falta de conteúdo harmônico em cada segmento de voz, que seja suficientemente capaz de modelar fielmente o espectro harmônico destes trechos. Observe que existe uma ligeira vantagem das conversões $M \times M$ em relação às outras.

Quanto aos métodos comparados na Figura 4.6, se percebe uma vantagem considerável do método NFW em todos os casos, exceto na conversão crítica $F \times M$. A transformação usando uma abordagem "codebook" também se apresenta como uma proposta bastante promissora, que pode ser estendida para um sistema de síntese a partir dos mapas fonéticos pré-determinados. Para a síntese

[2]Sentenças de uma mesma língua são exigidas neste caso.

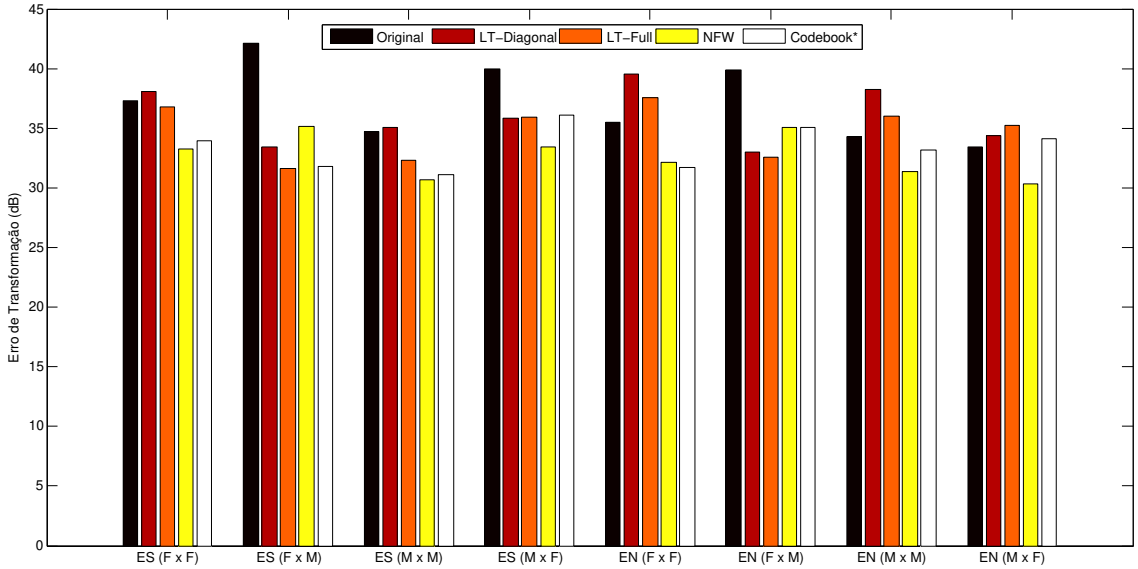

Figura 4.6: *Taxa de distorção espectral média da conversão entre cada par de corpora indicado.*

de voz com uma maior qualidade, espera-se que a quantidade de dados usada na fase de treinamento deva ser consideravelmente maior, para uma maior cobertura fonética no mapa fonético artificial.

Dois destes métodos comparados objetivamente foram escolhidos para serem utilizados na fase de avaliação subjetiva: o método de deformação em frequência proposto por este trabalho (NFW), e a transformação linear usando a diagonal da matriz de covariância. A transformação linear usando a matriz de covariância completa não foi usada na fase subjetiva por apresentar sinais transformados com artefatos ruidosos desagradáveis ao ouvido humano, provavelmente causados pela pouca quantidade de dados usados na fase de treinamento.

Os indicadores de barras mostrados anteriormente inspecionam os métodos de transformação sob uma ótica global, no nível das sentenças. A fim de inspecionar localmente os efeitos da transformação em cada trecho de voz que compõe um segmento longo, duas visualizações são apresentadas: a primeira, no domínio da frequência (Fig. 4.7), mostra uma visualização espectral da sentença original, da sentença destino ótima, e da sentença transformada; já a segunda (Fig. 4.8) exibe a taxa de distorção espectral alinhada temporalmente, de acordo com os segmentos de voz do sinal original no domínio do tempo. Para todas as visualização que seguem, um trecho curto de sentença arbitrária pronunciada pela locutora espanhola '76' (relativo ao corpus 'ES_76') foi convertida para a mesma sentença pronunciada pela locutora '75', também espanhola. Vale relembrar que o sinal de saída ótimo é obtido do alinhamento temporal da sentença pronunciada pela falante '75' e da sentença de entrada da falante '76'. O método usado na transformação local foi o método NFW proposto por este trabalho.

A Figura 4.7 apresenta dois gráficos semelhantes, relativos à mesma transformação. No gráfico de cima é mostrada a relação entre cada espectro origem e destino (ótimo) e seu correspondente espectro médio global, armazenado em seu respectivo corpus. No gráfico de baixo é apresentada a sobreposição do sinal transformado, conforme mostra a legenda. Observe nos gráficos que, fora da faixa aonde se localizam as principais regiões formânticas (F_1 e F_2), estão as regiões que concentram maiores discrepâncias na transformação. Isso se deve ao fato que tais regiões são especializadas em

configurar o timbre dos falantes, deixando a região de concentração formântica (marcada no gráfico superior) livre de interferências espectrais, estabelecendo pontos invariantes bastante importantes para o estabelecimento da comunicação. Embora os locutores sejam de sexos opostos, as regiões formânticas entre ambos os locutores preservam aproximadamente o mesmo formato, uma vez que a mesma sentença foi pronunciada. É desejável que a transformação final não distorça esta região. O papel do reajuste final é transpor as amplitudes harmônicas do sinal final no nível da sentença, para recuperar a energia original do sinal sem distorcer tal "região invariante".

Outro ponto a se observar no mesmo gráfico é a relação que existe entre o sinal original e a média global do corpus origem, relação essa que é transferida de certa forma para o sinal transformado (em cor negra) em relação à média global do corpus destino. Estas médias (desvios) globais correspondem aos valores médios (desvios padrões) de todos os vetores harmônicos quantizados para cada corpus.

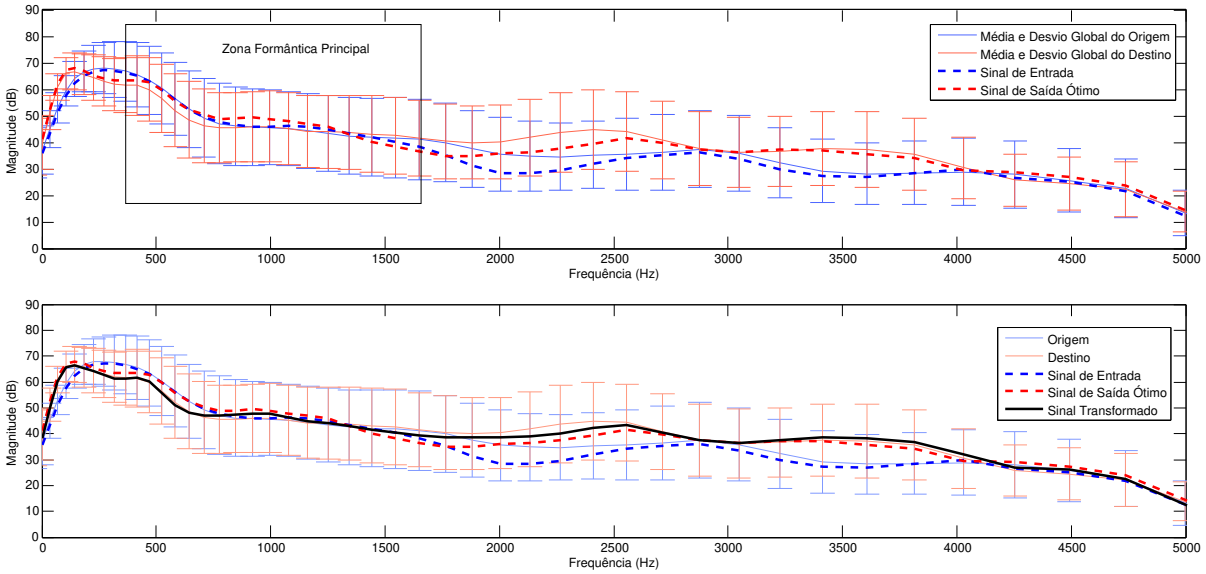

Figura 4.7: *Visualização espectral das sentenças origem, destino e transformada.*

A Figura 4.8, por sua vez, apresenta a evolução temporal da distorção espectral em função do sinal ótimo de saída e do sinal transformado. Tais erros de conversão são exibidos pela curva suavizada do gráfico central, situado entre o sinal original no tempo (gráfico superior) e o sinal transformado (gráfico inferior), verticalmente alinhados. Embora se observe um comportamento imprevisível das distorções espectrais, observa-se uma leve tendência na qual as taxas de distorção sobem à medida em que ocorre um evento de transição de formantes, como está indicado no gráfico central pelas regiões 1, 2 e 3. No entanto, tais oscilações são também (provavelmente) decorrentes da falta de cobertura fonética entre ambos os corpora, visto que uma pequena quantidade de dados ($\approx$ 5 min) foi utilizada. Observe também no gráfico central, que os diversos valores de M_c não alteram significativamente o valor da distorção espectral (SD) apresentada.

A taxa de distorção espectral foi calculada neste caso usando cada sinal de voz reconstruído de modo segmentado. Ou seja, foram tomados o sinal transformado e o sinal de saída ótimo já alinhados no tempo, e realizou-se uma segmentação pareada em pequenos trechos de 256 amostras. Para cada par de segmentos foi tomada a norma euclidiana da diferença entre o log da amplitude

das respectivas FFT, ou seja,

$$SD(x, y) = 10\,|\log_{10}(|\mathrm{fft}(x)|) - \log_{10}(|\mathrm{fft}(y)|)|\,,$$

onde x e y correspondem aos segmentos pareados. A distorção espectral foi escolhida neste caso por se tratar de uma medida bastante intuitiva, onde se deseja estimar o erro espectral entre ambas as sentenças alinhadas localmente. O valor médio de todos os segmentos corresponde aos valores plotados no gráfico central.

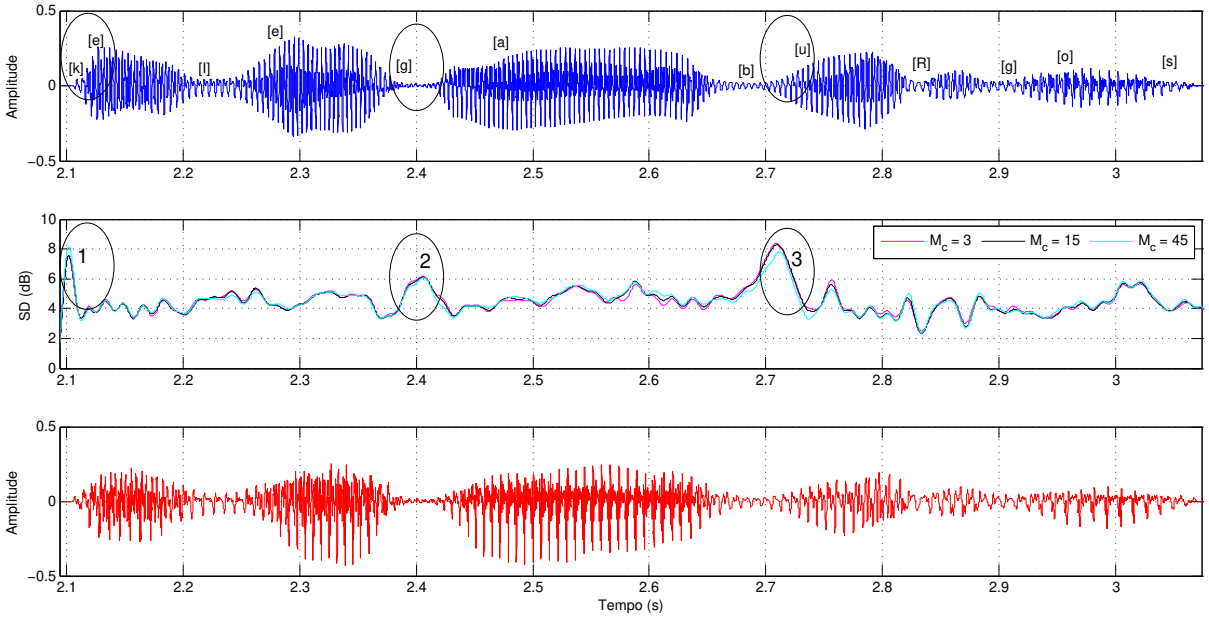

Figura 4.8: *Visualização das taxas de distorção espectral relacionando as sentenças origem, destino e transformada alinhadas no tempo.*

Apesar das taxas obtidas, observa-se experimentalmente que tais medidas objetivas não estão necessariamente correlacionadas com a percepção ou preferências humanas [238]. De fato, alguns trabalhos indicam que distâncias espectrais relativamente grandes podem fornecer uma boa transformação da voz [275]. A realização de uma avaliação subjetiva dos sinais convertidos é de grande importância para a validação do sistema proposto em termos não só numéricos, mas principalmente num sentido perceptual. Devido à dificuldade de se definir medidas de distância objetiva confiáveis, isto é, que sejam sensíveis aos requisitos perceptuais humanos, um teste perceptual é realizado na seção seguinte.

4.3 Avaliação Subjetiva

Normalmente, a validação de sistemas de processamento de áudio, e neste caso particular, o processamento de voz, requer o uso de métricas subjetivas associadas à percepção humana, a fim de analisar os resultados finais do sistema proposto. As ferramentas de teste subjetivas conhecidas como *MOS* e *ABX* serão utilizadas nesta seção.

4.3.1 Teste Perceptual

O teste de Opinião Média (do inglês *Mean Opinion Score – MOS*) é um processo de avaliação através de um questionário, que é muito apropriado quando se deseja avaliar a qualidade de uma voz convertida e sua similaridade em relação à voz do falante destino. Basicamente, o teste de opinião média realiza uma pesquisa em que os ouvintes avaliam a voz convertida utilizando uma escala de 5 valores (**5**: excelente, **4**: bom, **3**: regular, **2**: ruim, **1**: péssimo). Esta ferramenta de avaliação foi adotada como padrão dentro do projeto TC-STAR [23; 242], o qual propõe uma bateria de testes perceptuais comparativos, utilizando o *MOS* como métrica de similaridade e qualidade da conversão.

Outros autores [60; 301; 312] utilizam o teste *ABX*, um teste de múltiplas escolhas muito usado em verificação de similaridade da fala convertida em relação à fala do destino. Neste teste, os ouvintes devem decidir se uma dada sentença de voz processada X é mais próxima em termos perceptuais de uma das sentenças A ou B pronunciadas pelos falantes origem e destino, não necessariamente nesta ordem. O sucesso é medido em termos percentuais, cujas respostas são do tipo $X \approx T$, onde $T \in \{A, B\}$ é uma sentença de entrada qualquer.

Existe ainda outra forma de se utilizar a ferramenta *ABX* aplicada ao mesmo problema. Neste caso, a sentença X é uma sentença do destino original, e A e B são convertidas usando duas técnicas distintas: a técnica central da investigação e uma técnica de referência (porém nem sempre na mesma ordem). Os ouvintes então respondem qual das sentenças A ou B mais se parece com a sentença do destino original, sendo que neste caso, o ouvinte pode responder "nenhuma delas". As taxas de sucesso são calculadas para cada técnica como a porcentagem de sentenças convertidas que foram escolhidas como mais próximas da sentença pronunciada pelo destino.

Para a realização dos testes perceptuais desta tese, um sistema online foi desenvolvido, a fim de realizar os testes perceptuais necessários para a avaliação do sistema de conversão de voz inter-linguística apresentado de maneira interativa. A ferramenta MOS foi utilizada neste trabalho como padrão para avaliar a naturalidade e a similidade dos sons convertidos, assim como foi feito no projeto TC-STAR. No sistema de avaliação, foi solicitado para que os usuários utilizassem fones de ouvido de alta fidelidade e com volume confortavelmente ajustado. A entrevista perceptual foi composta por uma série de 16 grupos de perguntas relativas a cada conjunto de sentenças emparelhadas. A fim de evitar viéses causados pela ordem de exposição, os exemplos de áudio foram embaralhados. A fim de evitar o cansaço ou a saturação auditiva, o experimento permitiu que o teste pudesse ser realizado em várias seções, isto é, era possível responder quantas perguntas fosse confortável ao entrevistado, e em outra ocasião mais oportuna o teste poderia ser retomado de onde havia sido interrompido da última vez. Em cada rodada do teste duas questões foram respondidas:

- A primeira solicitou que o usuário do sistema escutasse cada sinal de voz transformado, e individualmente, apontasse um valor que representasse o quanto cada um destes sinais soava natural aos seus ouvidos, numa escala de 5 valores (Teste MOS). Neste contexto, foi explicado que a naturalidade estava associada tanto à qualidade do áudio escutado (em termos de artefatos ruidosos), quanto às qualidades vocais propriamente, ou seja, o quão "humanos" pareciam os sons pronunciados.

- Na segunda pergunta, o entrevistado deveria comparar cada sinal transformado com os sinais

relativos aos falantes origem e destino, discriminados como Falante A e Falante B, não necessariamente nesta ordem (Teste ABX). Foi explicado também que a similaridade do áudio era o único requisito que deveria ser levado em consideração, em oposição à qualidade ou à naturalidade do áudio.

Os métodos utilizados para comparação ABX foram o método proposto por este trabalho (Normalized Frequency Warping – NFW) e o método de transformação linear usando matriz diagonal (LT-DIAG); além destes o experimento também coletou medidas MOS relativas a um método embrionário de síntese usando a técnica *codebook* (COD), isto é, a sentença de saída é ressintetizada usando somente os centroides interpolados das classes fonéticas selecionadas a partir do módulo de seleção na fase de transformação, conforme visto no fim da Seção 3.6.

O teste proposto apresentou as sentenças transformadas sempre em uma língua distinta da dos sinais de origem e destino. Por exemplo, caso as sentenças do origem e destino estivessem na língua espanhola, os sinais transformados estariam na língua inglesa, e vice-versa. Deste modo, o fator de dependência quanto às diferenças de língua foi eliminado do experimento, ou seja, o sujeito não escolheria A ou B meramente em função desta sentença estar na mesma língua da sentença alvo. Neste caso experimental subjetivo, todos os métodos comparados passaram pelo estágio de pós-processamento proposto no fim do módulo de transformação (ver Seção 3.6), tais como suavização espectral aplicada na transição de segmentos harmônicos consecutivos, o rebalanceamento energético pela variância global ao nível da sentença e o reajuste energético entre as partes harmônica e estocástica dos sinais de voz.

A cada rodada do experimento (das 16 requisitadas), um conjunto de 5 sentenças de voz estava disponível para análise: as duas sentenças de entrada (do origem e do destino), e três sentenças de saída relativas aos métodos NFW, LT-DIAG e COD. Para evitar as respostas enviesadas dos entrevistados, as duas sentenças de entrada, bem como as sentenças a serem comparadas eram embaralhadas independentemente a cada rodada. Os pares de sentenças origem e destino (e consequentemente as convertidas) foram escolhidos sistematicamente de modo a cobrir todos os casos de conversão envolvendo as sentenças que compõem os corpora TC-STAR. Os pares organizados são exibidos na Tabela 4.3, bem como suas respectivas taxas de pontuação média obtidas para os quesitos de naturalidade e similaridade.

	***MOS* – Naturalidade**			***MOS* – Similaridade**		
Sentenças Pareadas	LT-DIAG	NFW	COD	LT-DIAG	NFW	COD
EN_F1 → ES_F2	**4.0727**	3.9636	2.0364	3.3091	**3.5455**	3.4425
ES_F1 → EN_F2	**4.1200**	4.0800	1.9400	**3.1200**	3.0200	2.6600
EN_F2 → ES_F1	**3.5909**	3.5606	1.8333	2.4091	2.8030	**2.8788**
ES_F2 → EN_F1	**3.4400**	3.3600	1.7800	3.0200	2.9600	**3.0800**
EN_F1 → ES_M1	**2.1800**	2.0600	1.4400	3.3600	3.4400	**3.6200**
ES_F1 → EN_M1	**2.8750**	2.6563	1.6094	3.6094	**3.7500**	3.4219
EN_F2 → ES_M2	**2.0800**	1.9800	1.3200	3.7600	**3.8600**	3.6400
ES_F2 → EN_M2	**2.4000**	2.2364	1.5091	3.5636	3.6364	**3.6909**
EN_M1 → ES_M2	3.2742	**3.4677**	1.9839	3.3065	**3.5484**	3.4516
ES_M1 → EN_M2	3.4510	**3.6863**	2.2745	3.5098	**3.9020**	3.7255
EN_M2 → ES_M1	**3.4151**	**3.4151**	1.6604	3.0000	**3.4528**	3.1509
ES_M2 → EN_M1	3.3800	**3.4800**	1.7600	3.2364	**3.5600**	3.4200
EN_M1 → ES_F2	2.7600	**2.7800**	1.4600	3.7000	3.7200	**3.8000**
ES_M1 → EN_F2	3.2838	**3.3243**	2.0676	3.5946	**3.7162**	3.5000
EN_M2 → ES_F1	2.5800	**2.7000**	1.4800	**3.7200**	3.7000	3.5600
ES_M2 → EN_F1	3.2241	**3.4138**	1.6897	3.6552	**3.7931**	3.5517

Tabela 4.3: *Resultado completo da entrevista perceptual (MOS).*

Evidentemente, as conversões foram realizadas de uma língua para outra, utilizando corpora

não-paralelos na fase de treinamento, e não as bases paralelas como nos sistemas de conversão de voz intra-linguística. Ou seja, supondo que se deseja converter uma sentença em espanhol de um falante masculino, digamos ES_M2, para uma falante inglesa EN_F1. Nesta caso, o mapeamento de classes acústicas bem como todo o processo que define a transformação utiliza os corpora de cada um destes falantes em suas respectivas línguas ES_M2 e EN_F1, e não a versão do corpus em Espanhol ES_F1 da locutora destino feminina. Tal restrição preserva a hipótese inicial do sistema de conversão inter-linguística de corpora não-paralelos.

Para que seja possível ter uma visão geral das transformações, os resultados médios dos experimentos foram tomados para cada par de corpora agrupados por gêneros: $F \rightarrow F$, $F \rightarrow M$, $M \rightarrow M$ e $M \rightarrow F$, de acordo com a Tabela 4.4. O experimento contou com a colaboração de **50** participantes.

Sentenças Pareadas	***MOS* – Naturalidade**			***MOS* – Similaridade**		
	LT-DIAG	NFW	COD	LT-DIAG	NFW	COD
F → F	**3.8059**	3.7411	1.8974	2.9645	**3.0821**	3.0411
F → M	**2.3838**	2.2332	1.4696	3.5733	**3.6716**	3.5932
M → M	3.3801	**3.5123**	1.9197	3.3441	**3.6158**	3.4370
M → F	2.9620	**3.0545**	1.6743	3.6674	**3.7323**	3.6029

Tabela 4.4: *Resultado médio da entrevista perceptual (MOS).*

A partir dos resultados experimentais, uma discussão a respeito dos aspectos acústicos é tratada na seção seguinte.

4.3.2 Discussão

Alguns pontos relevantes podem ser extraídos das informações obtidas nas tabelas do experimento anterior. A principal desvantagem dos indicadores subjetivos é a falta de um padrão referencial, um patamar capaz de permitir a comparação de tais resultados com outros da literatura.

Com relação aos resultados por método, observa-se que houve uma forte correlação em relação à avaliação objetiva, onde se confirmou a preferência do método NFW em relação ao método LT de transformação linear, exceto no caso de conversão de falantes femininos para outros falantes, conforme é mostrado na Figura 4.6. Quanto aos resultados obtidos, se observa a necessidade de algumas melhorias que possibilitem a implementação de uma versão comercialmente aceitável. Entre estas melhorias, uma maior quantidade de dados utilizada na fase de treinamento, uma maior taxa de quantização dos vetores acústicos e dos mapas fonéticos artificiais, assim como a estimação adequada das fases iniciais das componentes harmônicas [11] poderiam ser considerados a fim de aumentar a qualidade da conversão.

Embora saibamos que valores deste tipo de teste não sejam facilmente comparáveis com aqueles obtidos por diferentes autores, é interessante inspecionar alguns eventos que são incidentes na maioria dos resultados relatados e que podem enriquecer esta discussão.

A Tabela 4.5 exibe dois conjuntos de resultados de *MOS* tanto para naturalidade quanto para similaridade da conversão para falantes de uma mesma língua, respectivamente de cada sistema representado por seu autor. Alguns autores não realizaram os testes de similaridade separadamente dos de naturalidade, e por esta razão, ambos os requisitos estão representados por um único valor que mescla naturalidade e similaridade da conversão. Tipicamente, na maioria dos testes experimentais, considera-se dois tipos de conversão de voz: a conversão *intra-gênero* ($M \rightarrow M$ e $F \rightarrow F$) e a conversão *inter-gênero* ($M \rightarrow F$ e $F \rightarrow M$).

Ano	Autor	*MOS* para Naturalidade	*MOS* para Similaridade
1997	Kim [118]	3.42	
1998	Kain [108]	4.20 ($M \rightarrow M$) 2.70 ($M \rightarrow F$)	$\approx$ 2.63 ($M \rightarrow M$) $\approx$ 4.88 ($M \rightarrow F$)
1998	Stylianou [257]	$\approx$ 2.70	
2001	Toda [270]	$\approx$ 4.20 ($M \rightarrow M$) $\approx$ 2.70 ($F \rightarrow F$)	
2003	Rentzos [206]	3.65	
2004	Pfitzinger [189]	$\approx$ 1.50	
2005	Toda [268]	$\approx$ 3.10 ($F \rightarrow M$) $\approx$ 3.30 ($M \rightarrow F$)	
2006	Nurminen [175]	2.09	3.10 1.77 ($M \rightarrow M$) 2.20 ($M \rightarrow F$) 3.05 ($F \rightarrow M$) ($F \rightarrow F$)
2006	Duxans [48]	2.37	3.18
2006	Sündermann [244]	2.70 (Texto-Dependente) 2.60 (Texto-Independente)	
2006	Rao [203]	4.56 ($M \rightarrow F$) 4.71 ($F \rightarrow M$)	2.92 ($M \rightarrow F$) 3.23 ($F \rightarrow M$)
2006	Shuang [233]	4.09 (UK English) 3.68 (CN Mandarin)	1.87 (UK English) 2.77 (CN Mandarin)
2007	Dutoit [45]	2.56	2.77
2007	Erro [50]	3.27 ($M \rightarrow M$) 3.00 ($M \rightarrow F$) 3.60 ($F \rightarrow M$) 4.20 ($F \rightarrow F$)	2.93 ($M \rightarrow M$) 3.27 ($M \rightarrow F$) 2.53 ($F \rightarrow M$) 3.00 ($F \rightarrow F$)
2007	Fujii [60]	3.03 ($F \rightarrow F$) 2.75 ($M \rightarrow F$)	
2008	Shuang [234]	3.48	2.20
2008	Zhang [313]	3.00 ($M \rightarrow M$) 2.70 ($M \rightarrow F$) 3.10 ($F \rightarrow M$) 2.80 ($F \rightarrow F$)	
2008	Desai [42]	$\approx$ 2.70	
2009	Zhang [312]	2.70 ($F \rightarrow M$) 2.50 ($F \rightarrow F$)	2.20 ($M \rightarrow M$) 2.30 ($M \rightarrow F$) 2.50 ($F \rightarrow M$) 2.10 ($F \rightarrow F$)
2013	**Machado**	**3.51** ($M \rightarrow M$) **3.05** ($M \rightarrow F$) **2.38** ($F \rightarrow M$) **3.81** ($F \rightarrow F$)	**3.62** ($M \rightarrow M$) **3.73** ($M \rightarrow F$) **3.67** ($F \rightarrow M$) **3.08** ($F \rightarrow F$)

Tabela 4.5: *Relação de resultados experimentais de Opinião Média de Acerto – MOS.*

Todos estes testes foram realizados usando falantes origem e destino de uma mesma língua. Alguns resultados *MOS* para conversão inter-linguística entre Inglês e Espanhol foram feitas por Duxans [48] em 2006. Neste estudo, a pontuação *MOS* para qualidade foi de **2.33** do Espanhol para o Inglês, e a pontuação *MOS* para similaridade foi de **2.79**, também do Espanhol para o Inglês.

A Tabela 4.6 exibe um conjunto de resultados *ABX* de alguns sistemas de conversão de voz, como propostos por seus autores. O principal problema com a interpretação dos índices *ABX* é o fato de que não é permitido responder o quanto uma sentença X é similar a A ou a B. Disto, se pode inferir que um método que obteve alto índice de sucesso de acordo com o teste *ABX* poderia obter baixos valores de similaridade de acordo com o teste *MOS*.

No entanto, ABX é uma ferramenta poderosa, prática e extremamente simples para discriminar, dentre um par de métodos, aquele que realiza uma melhor conversão de voz em termos perceptuais. Dentre os autores que realizaram este tipo de teste estão Pozo [194] e Desai [42]. Pozo comparou seu método *Joint Estimation Analysis Synthesis (JEAS)* em relação ao método *Pitch-Synchronous Harmonic Model (PSHM)*, obtendo os seguintes resultados de sucesso: 41% ($M \rightarrow M$), 37% ($M \rightarrow F$), 33% ($F \rightarrow M$) e 36.5% ($F \rightarrow F$). Desai [42] comparou seu método de conversão usando ANN

Ano	Autor	Índice ABX
1998	Stylianou [257]	97%
1999	Arslan [6]	78% ($M \rightarrow M$) 100% ($M \rightarrow F$)
2001	Toda [270]	$\approx$ 77% ($M \rightarrow M$) $\approx$ 83% ($F \rightarrow F$)
2004	Orphanidou [183]	79.5% ($M \rightarrow M$) 86.3% ($M \rightarrow F$) 88.6% ($F \rightarrow M$) 77.3% ($F \rightarrow F$)
2005	Toda [268]	$\approx$ 84% ($M \leftrightarrow F$)
2005	Zhang [310]	87.5%
2006	Ye e Young [301]	91.8%
2007	Fujii [60]	100% ($M \rightarrow M$) 100% ($M \rightarrow F$) 100% ($F \rightarrow M$) 98.0% ($F \rightarrow F$)
2007	Hanzlicek [83]	87.2% ($F \rightarrow M$) 70.8% ($F \rightarrow F$)
2008	Yue [305]	92.0%
2008	Zhang [313]	$\approx$ 62% ($M \rightarrow M$) $\approx$ 80.5% ($M \rightarrow F$) $\approx$ 78.5% ($F \rightarrow M$) $\approx$ 55% ($F \rightarrow F$)
2009	Zhang [312]	68% ($M \rightarrow F$) 84% ($F \rightarrow F$)

Tabela 4.6: *Relação de resultados experimentais utilizando método ABX.*

com métodos tradicionais que usam GMM, obtendo taxa *ABX* de sucesso para similaridade de 65.0%. De igual modo, Türk [275] comparou sua proposta (conversão *Subband*) com os métodos clássicos (conversão *Full-band*) e obteve um índice *ABX* 92.9% para similaridade.

O objetivo de se analisar os resultados subjetivos neste caso não é discriminar qual é o melhor dentre todos os métodos, porque não é razoável comparar os resultados empíricos tais como os descritos nesta seção sem considerar os detalhes da experimentação, já que existem muitos fatores que podem influenciar significativamente o resultado do experimento, tais como o número de sentenças, a quantidade de participantes, a sensibilidade ou acuidade auditiva dos ouvintes, a qualidade do áudio original, a (não) ambiguidade das questões, entre outros fatores. Por esta razão, os experimentos devem ser cuidadosamente definidos com questões muito bem elaboradas a fim de se obter uma consistência experimental. A descrição do experimento deve ser suficientemente detalhada de modo a permitir uma replicação independente dos resultados experimentais. Apesar de tudo, alguns eventos estatísticos reincidentes sobre os resultados anteriormente apresentados podem ser utilizados como detectores de "entraves", isto é, de problemas inerentes a um sistema de conversão de voz com difíceis soluções, independentemente das condições experimentais.

Como exemplo disso, a relação existente entre a qualidade e a similaridade da conversão de voz, de acordo com alguns dos resultados apresentados, é uma relação competitiva. O sucesso de um destes quesitos está inversamente ligado ao sucesso do outro, ou seja, é bastante improvável que ambos os indicadores sejam altamente pontuados ao mesmo tempo. Tal conflito é em parte justificável diante da dualidade entre o ajuste excessivo (*overfitting*) de cada sentença em nível local e a suavização excessiva (*over-smoothing*) da transformação ao nível global.

O ajuste excessivo é proveniente da transformação segmento a segmento usando como suporte classes acústicas independentes no tempo. Assim, a continuidade temporal é comprometida, o que degrada substancialmente o sinal de voz na proporção em que se especializa a transformação.

Por outro lado, a suavização excessiva da transformação (não do sinal) ocorre quando um grande

número de classes são utilizadas na mistura ponderada. Neste caso, com a contribuição de muitas classes, a transforação do sinal passa a ser realizada num escopo mais global, mantendo assim, grande parte das estruturas locais do sinal original invariantes. Como resultado, um alto índice de correlação entre o sinal transformado e o original é estabelecido, diminuindo drasticamente a taxa de similaridade com o sinal destino.

Outro ponto interessante a se considerar é a dificuldade de se converter vozes femininas para masculinas, conforme pode ser observado em muitos dos métodos apresentados. Não se conhece com total segurança o motivo pelo qual tal conversão não é bem sucedida. Alguns autores [50] acreditam que este fenômeno está associado à dificuldade de se obter os envelopes espectrais de vozes femininas, dado um maior espaçamento dos harmônicos espectrais. O presente trabalho não é exceção à dificuldade observada na literatura em relação ao problema da conversão de voz entre homens e mulheres, embora alguma solução melhor tenha sido buscada por diversos caminhos possíveis. Nesta empreitada, variados experimentos preliminares (informais) foram feitos a fim de inspecionar o fenômeno da conversão $F \rightarrow M$, sem no entanto obter qualquer tipo de resultado contundente ou de validação estatística completa.

Pela experiência do autor advinda desta série de experimentos, os resultados obtidos sugerem que os problema de conversão de voz carrega consigo outro problema de difícil resolução: o problema da conversão das componentes de configuração fásica (φ) das partes harmônicas dos segmentos vozeados. Tal problema foi observado em praticamente todos os casos de conversão intra-gênero e inter-gênero, sendo mais evidente quando a conversão era feita de uma voz mais aguda para uma mais grave. Neste caso, a conversão $F \rightarrow M$ é a que mais sofre prejuízos, provavelmente devido a problemas com a reconstrução das fases iniciais das componentes harmônicas, como em outros trabalhos anteriores [137]. Embora tais observações não sejam conclusivas, o autor acredita na conjectura de que a consideração das componentes de fases harmônicas na conversão poderia aumentar significativamente a qualidade da conversão, seja a partir de uma modelagem das componentes de fase a serem armazenadas nos respectivos corpora, ou mesmo a partir de métodos de reconstrução destas componentes buscando sinais com alto grau de naturalidade. Se por um lado a conversão $F \rightarrow M$ é prejudicada pela falta de uma reconstrução adequada de configurações fásicas com alto grau de naturalidade, por outro se acredita que a conversão $M \rightarrow F$ carece de informações suficientes para reconstruir as regiões formânticas dos espectros reais do destino (feminino), uma vez que os mesmos são amostrados usando poucos harmônicos [50].

Outro ponto crítico que poderia aumentar a qualidade do som convertido compreende modelar mais cuidadosamente a parte estocástica dos corpora, estendendo a conversão a um nível local (de segmento) destas componentes. Neste caso, um espaço de características acústicas da parte estocástica deve ser considerado. A estimação da parte estocástica sem a contaminação de resíduos harmônicos é outro aspecto a ser considerado para uma melhor conversão de voz. Tal estimação depende essencialmente de um detector de pitch robusto.

Segue o capítulo de conclusão, onde são apresentadas as considerações finais deste trabalho.

Capítulo 5

Conclusões

Esta tese apresentou um sistema completo de conversão de voz e num contexto mais específico, a conversão de voz inter-linguística, onde foram discutidas algumas aplicações, tais como em personalização de sistemas TTS [46; 107; 108] e intérpretes virtuais [48; 50; 234].

O desafio da conversão de voz inter-linguística tem despertado interesse em estudos sobre similaridades e diferenças fonéticas entre línguas, bem como no desenvolvimento de transformações fonéticas automáticas entre línguas. Neste sentido, o trabalho apresentado propôs um modelo de treinamento, o qual independe de pré-requisitos específicos tais como o uso de pares de corpora paralelos, ou mesmo de dados rotulados. Além disso, ferramentas para análise e manipulação espectral foram introduzidas no Capítulo 3.

O trabalho apresentou um conjunto de técnicas úteis para clusterização e classificação dos segmentos de voz em termos de classes fonéticas artificiais. Ademais, o sistema também apresentou uma série de outros métodos para transformação espectral, em especial num contexto segmental, os quais assumem um papel crucial na fase de conversão das sentenças pronunciadas.

O texto traz também resultados experimentais que indicam que os métodos propostos podem ser considerados boas alternativas para conversão de voz entre falantes de línguas diferentes, ainda que o problema de obtenção de uma perfeita conversão de voz em termos perceptuais continua a ser um desafio em aberto, ajudando a indicar caminhos para extensão e aperfeiçoamento dos métodos propostos em trabalhos futuros.

5.1 Considerações Finais

Como pode ser constatado no texto, a conversão de voz é um problema complexo por tratar da manipulação de parâmetros acústicos que correspondem aos aspectos ligados à identidade sonora do indivíduo. Até o momento não se conhece um quadro completo e exaustivo dos elementos acústicos responsáveis pela identidade sonora, apesar da identificação de vários elementos que a compõem, tais como o ritmo de pronúncia, contorno melódico e o conteúdo harmônico, todos estes de difícil modelagem. Desde uma perspectiva de pesquisa e exploração de aplicações, o problema da conversão de voz inter-linguística se revela um nicho onde diversos tópicos de processamento de fala e computação musical são entrecruzados. Por exemplo, os problemas de reconhecimento tanto de fala quanto de voz são explorados indiretamente na conversão de voz, a fim de extrair informações acústicas e fonéticas de um falante, ao passo que na etapa de transformação são utilizados conceitos fundamentais de síntese aditiva, além de mapas acústicos, agrupamentos fonéticos k-verossímeis,

entre outros. Neste nível de abstração, é possível presumir que as ferramentas e conceitos desenvolvidos nesta tese possam contribuir para o desenvolvimento de novas ferramentas em tarefas gerais de processamento de fala.

Quanto ao problema de conversão de voz inter-linguística apresentado por este trabalho, alguns temas ainda não foram solucionados, tais como o problemas das fases de componentes harmônicas no modelo HSM. Segundo experimentos preliminares (não conclusivos), o problema de estimação das relações fásicas de segmento harmônicos parece estar relacionado com a diminuição significativa da taxa de naturalidade do som de voz convertido. Evidências experimentais do autor [137] apontam que a resolução do problema de estimação da configuração fásica (apresentado na Seção 3.2) tem potencial para tornar o sinal reconstruído mais natural do que a ressíntese utilizando as fases do falante origem.

Um outro problema que pode ter influenciado a transformação do sinal é o problema da cobertura fonética. Neste caso, a conversão sofre por não haver dados suficientes para representar um número considerável de classes fonéticas artificiais em cada corpus, para posterior alinhamento e conversão. Vale salientar que o uso de uma quantidade maior de dados de treinamento combinado a uma especificação mais densa do mapa fonético artificial são caminhos alternativos para melhorar a qualidade da conversão final. Pode-se considerar que, quanto maior o banco de voz, maior a riqueza do mapa fonético artificial, e portanto serão melhores tanto o alinhamento quanto a definição da função de transformação final em termos de momentos estatísticos. Soma-se a isso o fato de que a fase de treinamento na conversão inter-linguística pode ser realizada usando grandes bancos de voz sem qualquer pré-requisito fonético, isto é, sem a necessidade de sentenças rotuladas.

5.2 Trabalhos Aceitos em Anais de Congressos

Como resultado do doutorado realizado pelo autor, cinco artigos foram aceitos em congressos importantes dentro da área de processamento de sinais, tecnologias de fala e computação musical.

1. O primeiro deles foi aceito na conferência *Sound and Music Computing Conference 2010* em Barcelona na forma de um Survey Crítico, cujo conteúdo aborda tanto os problemas levantados, quanto aponta para direções futuras destacadas por recentes contribuições da literatura, como por exemplo, o desenvolvimento de um padrões de testes perceptuais para Sistemas de Conversão de Voz [139].

2. Nesse mesmo ano, um artigo foi apresentado em Taiwan no *IEEE International Symposium on Multimedia* o qual apresentou uma visão geral de um sistema de conversão inter-linguístico, com enfoque em discussões a respeito de técnicas mais significativas usadas no processo de conversão [140].

3. Ao final de 2012, um outro artigo foi aceito na conferência *IberSPEECH 2012 - VII Jornadas en Tecnología del Habla and III Iberian SLTech Workshop*, realizada na Espanha, o qual descreve uma nova abordagem em representação espectral usando soma de distribuições Gaussianas. A representação de envelopes espectrais usando parâmetros Gaussianos é bastante útil em tarefas de manipulação de sinais de voz, principalmente na manipulação de regiões formânticas e na deformação em frequência. [137]

4. A partir da representação espectral usando Gaussianas, no início de 2013 uma nova proposta foi apresentada no Canadá no congresso da IEEE intitulado *International Conference on Acoustics, Speech and Signal Processing*. Esse artigo apresenta uma técnica de representação espectral usando soma de funções paramétricas de base radial, as quais incluem uma vasta classe de janelas representadas por sua frequência central, amplitude e largura de banda. Resultados experimentais mostram que o uso de outras bases, como a janela de Hann, são mais apropriadas para a representação do envelope espectral do que o uso do modelo puro de soma de Gaussianas [138].

5.3 Sugestões para Pesquisas Futuras

O sistema apresentado pela tese foi testado em um cenário ótimo, sem a presença de ruído de fundo, e de prosódia com pouca expressividade. A ausência de emoção, bem como a ausência de ruído de fundo é um agente facilitador no processo de conversão. Sendo assim, como proposta de extensão de trabalho, um sistema de conversão robusto deve ser capaz de se adaptar a situações adversas como as mencionadas.

Um dos maiores desafios futuros é tornar o sistema de conversão de voz executável em tempo real. Grande parte dos módulos de conversão de voz são paralelizáveis, como por exemplo, o módulo de parametrização, onde cada base radial pode ser otimizada independentemente, o que facilita a implementação de uma versão paralela a ser executada por unidades de processamento específicas (como por exemplo, GPGPU [136]).

Com o avanço acelerado da tecnologia de fala, em pouco tempo será possível encontrar dispositivos de tradução simultânea de voz falada (intérpretes virtuais). Seria muito interessante que os sistemas de conversão de voz estivessem disponíveis como plugins acopláveis a tais dispositivos, personalizando vozes com timbres específicos, como os dos dois interlocutores de línguas diferentes. Neste exemplo, a exigência de funcionamento em tempo real está diretamente ligada à fase de conversão, e em especial, a parametrização acústica.

Outra sugestão de pesquisa futura está na criação de um banco de dados em outras línguas, e em especial, na língua portuguesa, a fim de atender os requisitos de cobertura fonética da língua e habilitar o uso de tal língua dentro do sistema. Na prática, é suficiente dispor de uma base de dados com cerca de uma hora de gravação de sentenças arbitrárias pronunciadas por um mesmo locutor, preferencialmente sem ruído de fundo.

Atualmente, o processo de estimação das bases paramétricas consome um tempo computacional inviável para o processamento em tempo real. Como sugestão de projetos futuros, seria muito interessante que existisse um método de estimação direta dos parâmetros associados a cada base, sem a necessidade de soluções exatas para problemas difíceis de otimização. Nesta direção, pode-se mencionar a estimação de bases paramétricas (vide Seção 3.3) a partir do método de mínimos quadrados aplicados sobre funções de bases linearizáveis, como a função Gaussiana por exemplo. Tal linearização aumenta significativamente a velocidade do cálculo dos coeficientes paramétricos que representam cada segmento de voz.

A conversão de voz também propicia uma série de ferramentas que podem ser difundidas em diversas áreas do processamento de voz e fala humana. Cada uma das ferramentas propostas por este trabalho pode ser adaptada para outros contextos de processamento digital de voz:

1. **Mapas Fonéticos Artificiais**: O mapa fonético artificial (Seção 3.4) proposto neste trabalho carrega consigo informações fonéticas relacionadas às posições dos formantes principais de cada língua. Se tais mapas acústicos forem devidamente rotulados, utilizando-se por exemplo um banco de voz com rótulos fonéticos precisos, é possível determinar para cada segmento de voz de uma sentença arbitrária, a classe fonética artificial que melhor representa tal segmento, conforme visto na Seção 4.2.2. Como cada classe é previamente rotulada, a tarefa de reconhecimento de voz é facilitada em grande parte com o uso destas ferramentas. A fim de expandir o potencial do mapa formântico na discriminação fonética, uma versão n-dimensional do mapa pode ser usada, de modo a representar até o n-ésimo formante de um sinal qualquer. Dentro do reconhecimento de fala, uma tarefa associada é a estimação de envelopes espectrais que sejam mais fiéis ao envelope verdadeiro. É possível utilizar os segmentos de voz de cada classe artificial a fim de estimar tais envelopes, usando as estratégias de *Análise Multi-Frame* apresentadas na Seção 3.3. Este é outro tema de continuação deste trabalho.

2. **Agrupamentos Fonéticos k-verossímeis**: Além do uso dos agrupamentos fonéticos k-verossímeis em conversão de voz, podemos usar tais agrupamentos em outras tarefas, tais como síntese de voz a partir de texto (TTS), ou mesmo reconhecimento de fala, uma vez que se espera que cada agrupamento pertença ao mesmo fonema. Neste contexto, uma sentença de fala é sintetizada a medida que são dadas as sequências de fonemas, bem como outros parâmetros de controle prosódico, como os contornos de pitch e energia, por exemplo. A taxa de articulação é controlada pelo deslocamento de um ponto que se desloca sobre o mapa acústico, e seleciona os vetores acústicos (ou centroides de classes k-verossímeis internas às classes fonéticas) mais próximos, a fim de compor o segmento de saída. Num contexto de computação musical, estes agrupamentos fonéticos também seriam uma boa base de dados a ser utilizada por algoritmos de síntese granular [43].

3. **Modelos de Representação Espectral usando Bases Radiais**: Um modelo de representação espectral flexível foi apresentado, o qual poderia ser aperfeiçoado a fim de representar propriamente as regiões formânticas com o menor número de parâmetros possível. Isso seria de grande interesse em tarefas de compactação de áudio e processamento de voz, e permanece aberto na literatura. Neste caso, a combinação de segmentos vizinhos no sinal de voz, bem como de outros parâmetros acústicos (derivadas e elementos de prosódia, por exemplo) é aconselhável na quantização paramétrica de cada segmento, a fim de obter uma melhor representação espectral do segmento de áudio.

4. **Módulo de Mapeamento de Classes Fonéticas**: Tal módulo desenvolvido usando conceitos da Teoria de Grafos se apresentou bastante eficaz no alinhamento de classes fonéticas artificiais, além de ser útil também no problema de alinhamento entre segmentos de voz paralelos (vide fim da Seção 3.5).

5. **Função de Frequência Normalizada**: Ao final da Seção 3.6, o trabalho apresentou um conjunto de conceitos ligados à representação espectral por meio de funções de distribuições energéticas. Todavia, as mesmas ainda carecem de exploração experimental, a fim de as validar e extrair propriedades úteis para as tarefas de processamento de fala. A deformação em frequência, por exemplo, mostrou-se bastante apropriada na conversão de voz inter-linguística.

No entanto, outros conceitos como a distribuição energética espectral, ou mesmo a distribuição em frequência normalizada, foram ferramentas pouco exploradas (individualmente) no contexto deste trabalho, ficando em aberto tal tema para investigação futura.

Um dos problemas mais difíceis e relevantes para esta área corresponde à elaboração de um modelo de avaliação objetiva que reflita de forma um pouco mais fiel os resultados obtidos em pesquisas de opinião. Tal modelo depende de investigações mais profundas de outros aspectos temporais e espectrais da fala (além dos abordados na tese) que influenciam diretamente na qualidade e similaridade da conversão. A similaridade e a qualidade da conversão, o contorno prosódico, bem como modelos de representação de um sinal de voz poderiam ser abordados separadamente, a fim de se analisar e categorizar mais precisamente a contribuição de cada um destes tópicos numa avaliação subjetiva. Além disto, a definição de um *benchmark* para comparação subjetiva de sistemas de conversão de voz para avaliação tanto de qualidade quanto de similaridade parece ser um dos desafios futuros mais urgentes, para que seja possível se obter uma versão comercial de alta fidelidade. Por fim, a utilização de sistemas de conversão de voz em aplicações musicais/artísticas se apresenta como um campo inexplorado de ideias a ser investigado num futuro próximo.

Apêndice A

Experimentos Preliminares

Os experimentos a seguir não pretendem definir um conjunto de escolhas ótimas para todos os casos apresentados nos quais se exige a escolha de um valor default. Simplesmente, tal seção tem por objetivo justificar a escolha das configurações iniciais utilizadas no sistema, tais como as bases do módulo de parametrização, a experiência preliminar a respeito da importância das fases iniciais num sistema HSM, assim como outros tipos de métrica usadas para fins específicos dentro do trabalho.

A.1 Experimento I

Um primeiro experimento se propõe a comparar, ainda que informalmente, três propostas clássicas de estimação de envelope espectral: usando LPC, usando Cepstrum e os interpoladores definidos na Seção 3.3, a fim de definir qual o envelopador será usado por *default* para estimar as versões suavizadas dos espectros de magnitude e fase harmônicas, bem como o espectro estocástico desta seção. Para isso, foi definido um experimento prático aplicado a um sinal de voz real,

Tabela A.1: *A média do EMQ entre o envelope estimado e cada vetor de amplitude harmônica do conjunto fonético.*

Envelopador	EMQ (10^3)
Interp. Radial – Blackman-Harris	4.2419
Interp. Radial – Blackman-Nutthall	4.2485
Interp. Radial – Gaussiano	**4.2298**
Interp. Radial – Hann	4.2636
Interp. Radial – Nuttall	4.2401
LPC (16)	6.8844
LPC (24)	6.1623
CEPS (16)	7.8842
CEPS (24)	7.0893
SPLINES^3	4.3985
LINEAR	4.4116
MFA	4.3699

que corresponde a uma vogal /*a*/ sustentada por 3 segundos onde o locutor variou a altura musical (*em glissando*) entre 148 Hz e 316 Hz, aproximadamente. Foram tomados os envelopes

harmônicos amostrados entre as frequências 0 e 5000 Hz, e a média de todos os envelopes foi calculada para cada envelopador. Finalmente, o erro médio quadrático (EMQ) entre cada amostra harmônica e o envelopador é obtido, conforme mostra a Tabela A.1. O MFA corresponde ao método de estimação espectral baseado na estratégia Multi-Frame Analysis de Shiga [227].

Embora este experimento tenha sido bastante informal, sendo que uma avaliação mais robusta dependeria de um conjunto de dados maior assim como de um ferramental estatístico mais abrangente, os resultados preliminares sugerem que vários dos métodos propostos são alternativas razoáveis para a representação espectral segmento-a-segmento, e em especial, a interpolação radial com funções de base *Gaussianas* obteve o melhor desempenho neste caso específico. Sendo assim, o interpolador Gaussiano foi usado como interpolador radial padrão dos envelopes espectrais harmônicos e estocásticos. Uma avaliação mais rigorosa será realizada na seção de experimentação, a fim de reconfirmar a escolha deste método envelopador.

A.2 Experimento II

Este experimento perceptual preliminar tem por objetivo destacar levantar evidências que reforçam a conjectura de que nosso sistema perceptual é de fato sensível à configuração das componentes de fases em sinais quase-periódicos, principalmente em sinais harmônicos de baixa frequência.

No experimento perceptual proposto, que contou com a colaboração de um grupo de 43 avaliadores, foi gerada uma sequência de 60 fragmentos sintéticos de áudio, com 2 segundos cada. Esses fragmentos foram particionados em duas categorias: sons harmônicos e sons inarmônicos. A primeira categoria é composta por sinais constituídos pela soma de pulsos linearmente espaçados dentro de intervalo $[0, 5]$ kHz. Assim, a n-ésima amostra deste conjunto possui n harmônicos de amplitude $1/n$, espaçados pela frequência fundamental $f_0 = 5000/(n + 1)$ Hz.

O outro conjunto de sinais (inarmônicos) tem seu espectro composto por um conjunto aleatório de pulsos de magnitude $1/n$, todos dentro do mesmo intervalo de 0 a 5 kHz, porém sem o espaçamento linear (harmônico). Então cada uma destas sequências foi sintetizada usando três conjuntos de configurações fásicas: (1) com fases iniciais linearmente espaçadas entre $-\pi$ e π, (2) com fase zero e (3) com fases aleatórias. Finalmente, um conjunto embaralhado destes sinais harmônicos e inarmônicos devidamente emparelhados, de modo a conter os três casos de teste, foram submetidos a uma pesquisa de opinião, para averiguar se os pares eram ou não perceptualmente idênticos. Para evitar o desgaste auditivo do voluntário, o número de provas auditivas requisitadas em cada rodada de teste variou entre 5 e 30 questões. Até então, cerca de 830 questões foram submetidas.

A premissa básica do experimento é detectar se o ouvinte é capaz de diferenciar as três versões do sinal (com mesmo espectro de magnitude), reconstruído com diferentes configurações de fase.

A Figura A.1 apresenta o resultado do experimento, e revela aspectos do sistema de percepção humana de certa forma inusitados. De acordo com este resultado, é clara a distinção entre os dois conjuntos testados. O ouvido humano se demonstrou pouco sensível à detecção de configuração de fases em sons complexos. Por outro lado, sinais mais simples, como os trechos vozeados de sinais de voz, compostos por um número controlado de frequências harmônicas, são diferenciáveis quanto à configuração destas fases, principalmente em se tratando de sinais sonoros mais graves. Talvez esta seja um fato que justifique resultados tão contraditórios entre os trabalhos de tecnologia musical [196] e de processamento de fala [192].

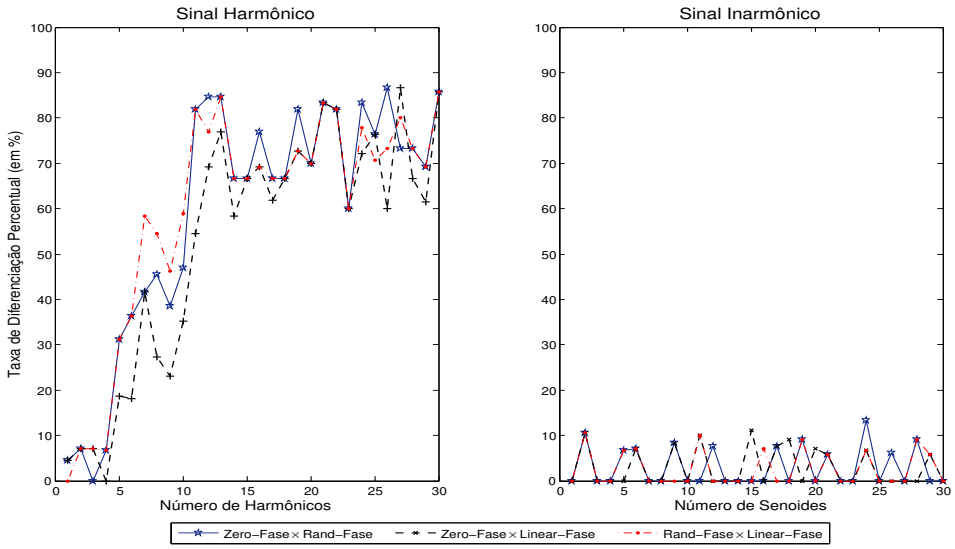

Figura A.1: *Resultado perceptual dos testes de configuração fásica.*

A.3 Experimento III

Uma simples investigação dentro do módulo de seleção de classes justifica a escolha função de normalização quártica que define a Equação 4.6 na Seção 4.2.3. A formulação foi selecionada após ter sido realizado um pequeno teste preliminar usando a distorção mel-cepstral ($\mathbf{d}_C(v,c)$) como base. O experimento também utilizou o corpus espanhol 'ES_75' como referência, assim como as seguintes formulações alternativas listadas:

F_1 – Normalização Linear: $\mathbf{w}(v,c) = N(v,c)$.

F_2 – Normalização Quadrática: $\mathbf{w}(v,c) = N(v,c)^2$.

F_3 – Normalização Cúbica: $\mathbf{w}(v,c) = N(v,c)^3$.

F_4 – Normalização Quártica: $\mathbf{w}(v,c) = N(v,c)^4$.

F_R – Normalização Raiz Quadrática: $\mathbf{w}(v,c) = \sqrt{N(v,c)}$.

F_e – Normalização Exponencial: $\mathbf{w}(v,c) = e^{N(v,c)}$.

onde

$$N(v,c) = \left(1 - \frac{\mathbf{d}_C(v,c) - \min_{\forall k}\{\mathbf{d}_C(v,C_k)\}}{\max_{\forall k}\{\mathbf{d}_C(v,C_k)\} - \min_{\forall k}\{\mathbf{d}_C(v,C_k)\}}\right).$$

O arcabouço experimental básico desta seção consiste em realizar buscas por classes c_i, $i \in [M_c]$ no corpus origem com maiores graus de pertinência $\mathbf{w}(v_k, c_i)$, para todos os vetores $v_k = \Psi_{\cdot A}^{k}$ do sinal de entrada. A partir das classes c_i, um vetor artificial $\bar{v_k}$ relativo à v_k é gerado de tal forma que

$$\bar{v_k} = \frac{\sum_{i=1}^{M_c} \mathbf{w}(v_k, c_i) \cdot \mu_{c_i}}{\sum_{i=1}^{M_c} \mathbf{w}(v_k, c_i)}.$$

M_c	$\Gamma(M_c)$ segundo o tipo de normalização					
	F_1	F_2	F_3	F_4	F_R	F_e
1	33.1470	33.1470	33.1470	33.1470	33.1470	33.1470
2	32.2986	32.2764	32.2560	32.2366	32.3097	32.2995
3	31.9165	31.8753	31.8369	31.7994	31.9376	31.9183
4	31.7873	31.7331	31.6807	31.6302	31.8144	31.7891
5	31.7375	31.6765	31.6164	31.5571	31.7680	31.7411
6	31.6325	31.5684	31.5042	31.4409	31.6655	31.6361
7	31.5905	31.5232	31.4559	31.3877	31.6242	31.5941
8	31.5676	31.4921	31.4166	31.3402	31.6044	31.5720
9	31.5462	31.4660	31.3858	31.3039	31.5854	31.5514
10	31.5267	31.4426	31.3569	31.2702	31.5674	31.5319
11	31.5161	31.4282	31.3386	31.2473	31.5591	31.5221
12	31.5002	31.4110	31.3193	31.2258	31.5448	31.5071
13	31.5061	31.4122	31.3165	31.2192	31.5513	31.5129
14	31.5188	31.4226	31.3240	31.2227	31.5664	31.5264
15	31.5145	31.4138	31.3097	31.2039	31.5644	31.5230
16	31.5214	31.4176	31.3105	31.2017	31.5720	31.5298
17	31.5455	31.4380	31.3279	31.2135	31.5985	31.5548
18	31.5714	31.4591	31.3427	31.2238	31.6258	31.5814
19	31.5817	31.4649	31.3447	31.2212	31.6393	31.5926
20	31.5978	31.4765	31.3518	31.2230	31.6577	31.6095
21	31.6089	31.4838	31.3538	31.2212	31.6702	31.6213
22	31.6353	31.5056	31.3718	31.2339	31.6990	31.6486
23	31.6649	31.5314	31.3937	31.2519	31.7300	31.6781
24	31.6944	31.5578	31.4164	31.2708	31.7610	31.7083
25	31.7185	31.5781	31.4337	31.2851	31.7866	31.7324
26	31.7479	31.6030	31.4532	31.2993	31.8191	31.7634
27	31.7669	31.6174	31.4631	31.3056	31.8395	31.7832
28	31.7979	31.6446	31.4857	31.3235	31.8728	31.8150
29	31.8359	31.6773	31.5138	31.3462	31.9140	31.8538
30	31.8614	31.6982	31.5302	31.3581	31.9410	31.8801
31	31.8972	31.7287	31.5553	31.3779	31.9790	31.9166
32	31.9270	31.7539	31.5760	31.3940	32.0119	31.9480
33	31.9554	31.7777	31.5952	31.4087	32.0426	31.9772
34	31.9797	31.7968	31.6098	31.4180	32.0692	32.0023
35	32.0108	31.8233	31.6302	31.4339	32.1025	32.0341
36	32.0375	31.8441	31.6458	31.4435	32.1322	32.0624
37	32.0698	31.8703	31.6659	31.4584	32.1667	32.0954
38	32.0960	31.8920	31.6824	31.4696	32.1960	32.1232
39	32.1253	31.9153	31.7005	31.4817	32.2283	32.1541
40	32.1526	31.9373	31.7173	31.4933	32.2586	32.1820
41	32.1858	31.9645	31.7385	31.5085	32.2948	32.2168
42	32.2106	31.9842	31.7522	31.5178	32.3227	32.2424
43	32.2394	32.0070	31.7698	31.5295	32.3536	32.2719
44	32.2719	32.0342	31.7927	31.5471	32.3891	32.3059
45	32.3076	32.0648	31.8164	31.5656	32.4271	32.3432
46	32.3428	32.0940	31.8404	31.5837	32.4645	32.3792
47	32.3828	32.1272	31.8669	31.6035	32.5082	32.4206
48	32.4122	32.1515	31.8853	31.6167	32.5406	32.4516
49	32.4382	32.1723	31.9010	31.6273	32.5687	32.4783
50	32.4675	32.1957	31.9200	31.6404	32.6010	32.5091
MÉDIA	31.8993	31.7515	31.6004	**31.4466**	31.9719	31.9168

Tabela A.2: *Taxas de erro do módulo de seleção considerando o número de coeficientes M_c e o tipo de normalização.*

Como resultado do experimento, a Tabela A.2 foi criada, exibindo as taxas de erro $\Gamma(M_c)$ do módulo de seleção, conforme apresentadas anteriormente no arcabouço experimental básico. Observe que a normalização quártica é de fato uma boa escolha para **função de normalização**, que embora com vantagem ter sido desprezível, obteve menor taxa de erro de classificação.

Apêndice B

Conceitos Básicos Complementares

Esta parte do trabalho reúne conceitos básicos de outras disciplinas relacionadas ao trabalho, não necessariamente dentro do escopo de processamento de voz. Por exemplo, a Seção B.1 aborda alguma noção básica, quiza útil, para a fundamentação dos conceitos primordiais para a definição do mapa formântico apresentado na Seção 3.4. A Seção B.2 traz uma breve revisão a respeito da Teoria de Grafos, aplicável no módulo de mapeamento de classes fonéticas artificiais visto na Seção 3.5. Este módulo faz referência ao uso dos operadores morfológicos tidos básicos da Morfologia Matemática, a qual é reportada na Seção B.3.

B.1 Linguística Aplicada à Localização dos Formantes

Ferdinand de Saussure [94], o pai da linguística, estabeleceu algumas distinções fundamentais entre a fala e a língua: língua é um sistema de comunicação compatível com um código (a linguagem natural) e fala é a utilização desse sistema de comunicação. No código da língua, as palavras são compostas por concatenações de fonemas, como se fossem imagens ideais dos diversos sons. Um fonema é a menor unidade sonora (fonética) de uma língua que estabelece contraste de significado para diferenciar palavras.

Na fala cotidiana, a pronúncia de cada fonema varia de acordo com as pessoas, aspectos culturais, contextos das palavras ou momentos em que tais palavras são emitidas. É interessante perceber que uma palavra nunca é pronunciada do mesmo modo por todos os falantes, o que torna necessário definir as unidades fonéticas capazes de modificar o significados das palavras, sem desprezar o conteúdo que define a identidade sonora do indivíduo. Tais unidades fonéticas são de grande importância no processo de conversão de voz, uma vez que introduzem a noção de inteligibilidade das sentenças pronunciadas. É evidente que na fase de transformação estas unidades fonéticas devem ser conservadas ao longo da conversão.

Sabe-se que a escrita da pronúncia real dos sons pode ser feita por meio de um alfabeto especial, o **alfabeto fonético**, uma vez que não existe um alfabeto comum capaz de reproduzir com precisão todos sons de qualquer linguagem. Por exemplo, a letra *c* possui diferentes sonoridades em *caso* e *cedo*; esses sons são reproduzidos, no alfabeto fonético, por [k] e [s]. Analogamente, a letra [z] pode corresponder aos fonemas s, z. A fim de evitar confusões, os fonemas são sempre transcritos entre barras oblíquas (/*/), enquanto que os símbolos fonéticos são denotados entre colchetes ([*]).

Os fonemas são classificados em **vogais**, **semivogais** e **consoantes**. Uma vogal é um tipo de fonema produzido pela excitação das cordas vocais, a qual não encontra nenhum obstáculo ao

atravessar o aparelho fonador. A mais importante classificação vocálica é feita quanto ao modo de articulação. Existem basicamente as *vogais orais*, pronunciadas completamente através da cavidade oral (boca), e as *vogais nasais*, que permitem que uma parte do ar usada na pronúncia escape pela cavidade nasal.

O Português utiliza basicamente 33 fonemas, sendo 12 vogais, 2 semivogais e 19 consoantes. Destas 12 vogais, sete são orais: [a], [ê], [é], [i], [ô], [ó] e [u]; e cinco são nasais: [ã], [em], [im], [õ] e [um] As semivogais são fonemas que não ocupam a posição tônica da sílaba, sendo associadas a uma vogal para formarem uma sílaba: [I], [w]. As *consoantes* por outro lado, são fonemas produzidos pela corrente de ar ao ultrapassar um obstáculo do aparelho fonador. Estes obstáculos incluem os dentes, a língua, lábios e o palato, entre outros. Podemos classificar as consoantes quanto ao *papel das cordas vocais*, quanto ao *modo de articulação* e quanto ao *ponto de articulação*. Diferentemente das vogais, existem consoantes não-vozeadas, como as consoantes [k], [p], [s], [t] e [f], e existem as consoantes vozeadas, como por exemplo, as consoantes [j], [l], [m], [n], [r], [v] e [z], entre outras.

Existe um padrão universal de representação de fonemas que constitui o Alfabeto Fonético Internacional (AFI) [128], criado em 1888 pela Associação Fonética Internacional, que utiliza letras latinas e algumas letras gregas. Entretanto, para facilitar nossa assimilação quanto aos fonemas, adotaremos os fonemas adaptados à língua portuguesa, conforme mostra a Tabela B.1. Tal tabela foi extraída de http://pt.wikipedia.org/wiki/Português_brasileiro e exibe os principais fonemas do Português Brasileiro.

B.2 Bases da Teoria de Grafos

A Teoria de Grafos fornece um conjunto de ferramentas teóricas que modelam um grande conjunto de problemas práticos computacionais. Um dos problemas é o emparelhamento entre grafos bipartidos. Assim, esta seção se dedica a definir alguns conceitos que devem ser compreendidos a fim de descrever o algoritmo emparelhador.

Define-se um **emparelhamento** como um subconjunto $E \subset A$, no qual para todo $a_{i,j}, a_{m,n} \in E$ com $a_{i,j} \neq a_{m,n}$ vale que $m \neq i$ e $n \neq j$. Um emparelhamento E é dito **máximo** quando não existe outro emparelhamento E^* tal que $|E^*| > |E|$. O **custo de um emparelhamento** E é definido como $c(E) = \sum_{\forall a_m \in E} c(a)$, onde $c(a)$ é o custo de a. Seja $\mathcal{E} = \{E_1, E_2, \ldots, E_M\}$ o conjunto de todos os emparelhamentos máximos de $\mathbb{G}$. Um **emparelhamento máximo de custo mínimo** $E_m \subset \mathcal{E}$, é um emparelhamento no qual $c(E_m) \leq c(E_k),\ \forall E_k \in \mathcal{E},\ k \neq m$.

Os algoritmos de busca por emparelhamento normalmente utilizam conceitos como caminhos alternantes ou aumentativos, entre outros. Um caminho é uma sequência $\{v_1, v_2, \ldots, v_k\}$ de vértices distintos, tal que para cada par de vértices (v_i, v_{i+1}), existe uma aresta $a_{i,i+1} \in A$. Um **caminho alternante** é um caminho no qual suas arestas pertencem, alternadamente, a um emparelhamento E de $\mathbb{G}$. Quando os extremos de um caminho alternante não são ligados a nenhuma aresta de E, dizemos que tal caminho é um **caminho aumentativo**. Tais vértices extremos são conhecidos como vértices livres.

Somente com estas informações já é possível implementar um algoritmo de emparelhamento, dentre os clássicos que existem na literatura.

	Fonema	Característica fonética	Exemplos
Vogais	/a/	Aberta, frontal, oral, não arredondada	**á**tomo, **a**rte
	/ɐ/	Aberta, central, oral, arredondada	cam**a**, canj**a**
	/ã/	Semi-aberta, central, nasal, não arredondada	**an**tes, **am**plo, maç**ã**
	/ɛ/	Semi-aberta, frontal, oral, não arredondada	m**é**trica, p**e**ça
	/e/	Semi-fechada, frontal, oral, não arredondada	m**e**do, p**ê**ssego
	/ẽ/	Semi-fechada, frontal, nasal, não arredondada	s**em**pre, **ê**mbolo, c**en**tro
	/ɔ/	Semi-aberta, posterior, oral, arredondada	**ó**tima, **o**va
	/o/	Semi-fechada, posterior, oral, arredondada	r**o**lha, av**ô**
	/õ/	Semi-fechada, posterior, nasal, arredondada	**om**bro, **on**tem, c**ôn**sul
	/i/	Fechada, frontal, oral, não arredondada	**i**tem, silv**í**cola
	/ĩ/	Fechada, frontal, nasal, não arredondada	s**im**ples, s**ím**bolo, t**in**ta
	/u/	Fechada, posterior, oral, arredondada	**u**va, **ú**tero
	/ũ/	Fechada, posterior, nasal, arredondada	alg**um**, n**un**ca, m**ui**to
Semivogais	/y/	Oral, palatal, sonora	**ui**vo, mã**e**, á**re**a
	/w/	Oral, velar, sonora	a**u**tomático, móv**el**, freq**ue**nte
Consoantes	/m/	Nasal, sonora, bilabial	**m**arca
	/n/	Nasal, sonora, alveolar	**n**ervo
	/ɲ/	Nasal, sonora, palatal	arra**nh**ado
	/b/	Oral, oclusiva, bilabial, sonora	**b**arco
	/p/	Oral, oclusiva, bilabial, surda	**p**ato
	/d/	Oral, oclusiva, linguodental, sonora	**d**ata
	/t/	Oral, oclusiva, linguodental, surda	**t**elha
	/g/	Oral, oclusiva, velar, sonora	**g**ato
	/k/	Oral, oclusiva, velar, surda	**c**arro, **qu**anto
	/v/	Oral, fricativa, labiodental, sonora	**v**ento
	/f/	Oral, fricativa, labiodental, surda	**f**arelo
	/z/	Oral, fricativa, alveolar, sonora	**z**ero, ca**s**a, e**x**alar
	/s/	Oral, fricativa, alveolar, surda	**s**eta, **c**ebola, e**x**cesso
	/ʒ/	Oral, fricativa, pós-alveolar, sonora	**g**elo, **j**arro
	/ʃ/	Oral, fricativa, pós-alveolar, surda	**x**arope, **ch**uva
	/ʀ/	Oral, vibrante, sonora, uvular	**r**ato, ca**rr**oça
	/r/	Oral, vibrante, sonora, alveolar	va**r**iação
	/ʎ/	Oral, lateral aproximante, sonora, palatal	caval**lh**eiro
	/l/	Oral, lateral aproximante, sonora, alveolar	**l**uz

Tabela B.1: *Fonemas adaptados ao contexto da língua portuguesa, espanhola e inglesa.*

B.3 Bases da Morfologia Matemática

A *Morfologia Matemática* (MM), fundada por J. Serra e G. Matheron ([223; 224]) entre os anos de 1960 e 1969 e baseada na teoria de reticulados [16], é uma disciplina muito popular e importante na área de Processamento de Imagens, criada para processar estruturas geométricas, como neste contexto para encontrar a componente conexa mais à esquerda de cada mapa fonético.

Matematicamente, um **reticulado completo** $(A, \leq)$ é um conjunto de dados A com N elementos, parcialmente ordenado pela relação $\leq$, no qual todos os subconjuntos $A_i \subseteq A$ possuem um valor mínimo, bem como um valor máximo. Um conjunto A é dito ser parcialmente ordenado em relação a $\leq$ quando as seguintes propriedades são satisfeitas para todos $a, b, c \in A$, como segue:

1. Reflexibilidade: quando $a \leq a$;
2. Anti-simetria: se $a \leq b$ e $b \leq a$ então $a = b$;
3. Transitividade: se $a \leq b$ e $b \leq c$ então $a \leq b$.

A MM fornece um compêndio de ferramentas construídas a partir da composição de operações básicas da teoria dos conjuntos aplicadas a reticulados completos. O conjunto destas ferramentas são os chamados **operadores morfológicos**. Além dos propósitos de extração de componentes

conexas abordados neste trabalho, a MM serve também para deformar (ou empenar) o espaço acústico para propósitos gerais, tais como transformações espectrais, transposição e filtragem de mapas formânticos, entre outras.

Dizemos que um operador morfológico $\Pi : \mathcal{P}(A) \to \mathcal{P}(A)$ é **invariante por translação** (i.t.) se para todo conjunto $S \in \mathcal{P}(A)$ e $p \in \mathbb{Z}^d$ é verdade que

$$\Pi(S_x) = \Pi(S)_x,$$

onde a notação Ω_x corresponde à translação do conjunto Ω em relação a x, ou seja, $\Omega_x = \{p+x \mid p \in \Omega\}$.

Outra classe importante de operadores morfológicos são os **operadores localmente definidos** (l.d.) por uma janela centrada em um ponto. Um operador $\Pi(\cdot)$ é l.d. em W centrada em c se

$$x \in \Pi(S) \iff x \in \Pi\left(S \cap W_c\right).$$

A classe dos operadores morfológico l.d. e i.t., os chamados **W-operadores**, são muito importantes em MM, uma vez que possibilitam realizar transformações globais avaliando-se regiões locais da estrutura geométrica em questão. Tais operadores utilizam um tipo especial de estrutura morfológica contínua $B_o : \mathbb{Z}^d \to \mathbb{Z}$ conhecido como **elemento estruturante** (EE) centrado na origem o do espaço $\mathbb{Z}^d$.

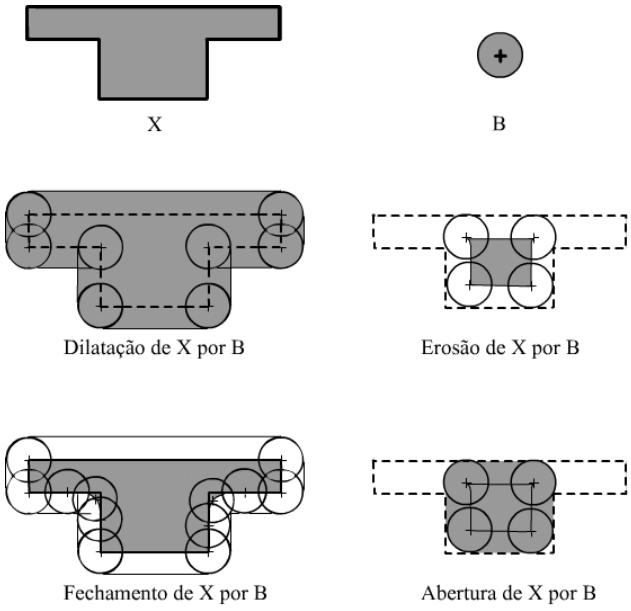

Figura B.1: *Principais operadores morfológicos.*

Os W-operadores morfológicos tidos como bases da morfologia matemática clássica são os operadores clássicos de erosão e dilatação morfológica, os quais são largamente usadas para definir uma série de outros operadores. A **erosão** morfológica $\varepsilon : \mathbb{Z}^d \to \mathbb{Z}^d$ de um conjunto discreto $A \in \mathcal{P}(\mathbb{Z}^d)$ pelo elemento estruturante $B \subseteq (\mathbb{Z}^d)$, denotada por $\varepsilon_B(A)$, é definida como

$$\varepsilon_B(A) = \bigcap_{b \in B} A_{-b} = \{x \in \mathbb{Z} \mid \forall b \in B \text{ tal que } x + b \in A\} \tag{B.1}$$

O operador dual da erosão, a **dilatação** morfológica $\delta : \mathbb{Z}^d \to \mathbb{Z}^d$ de um conjunto A pelo elemento estruturante B, denotada por $\delta_B(A)$, é definida como

$$\delta_B(A) = \bigcup_{b \in B} A_b = \{x \in \mathbb{Z} \mid \exists b \in B \text{ tal que } x - b \in A\} \tag{B.2}$$

Outros dois operadores muito usados na morfologia matemática são os operadores de **abertura**

$$\gamma_B(A) = \delta_B(\varepsilon_B(A)) = \bigcup_{h \in \mathbb{Z}^d} \{B_h \mid B_h \subseteq A\}$$

e **fechamento** morfológico

$$\phi_B(A) = \varepsilon_B(\delta_B(A)).$$

A Figura B.1 ilustra uma interpretação geométrica de como funcionam os operadores morfológicos descritos até aqui.

Referências Bibliográficas

[1] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara. Voice conversion through vector quantization. In *International Conference on Acoustics, Speech and Signal Processing, ICASSP-88*, pages 655–658, 1988. 7, 11, 58

[2] M. Abe, K. Shikano, and H. Kuwabara. Cross-language voice conversion. In *International Conference on Acoustics, Speech, and Signal Processing, ICASSP-90*, pages 345–348, 1990. 7, 11, 39, 122

[3] S. Ahmadi, A.S. Spanias, N.M.P. Inc, and C.A. San Diego. Cepstrum-based pitch detection using a new statistical V/UV classification algorithm. *IEEE Transactions on Speech and Audio Processing*, 7(3):333–338, 1999. 40

[4] L.D. Alsteris and K.K. Paliwal. Evaluation of the modified group delay feature for isolated word recognition. In *International Symposium on Signal Processing and its Applications (ISSPA-2005), Sydney, Australia*, pages 715–718, 2005. 76

[5] T.W. Anderson. *An introduction to multivariate statistical analysis*. Wiley New York, 1958. 59

[6] L.M. Arslan. Speaker transformation algorithm using segmental codebooks (STASC). *Speech Communication*, 28(3):211–226, 1999. 7, 11, 58, 164

[7] L.M. Arslan and D. Talkin. Voice conversion by codebook mapping of line spectral frequencies and excitation spectrum. In *Fifth European Conference on Speech Communication and Technology*. Citeseer, 1997. 10

[8] L.M. Arslan and D. Talkin. Speaker transformation using sentence HMM based alignments and detailed prosody modification. In *International Conference on Acoustics, Speech and Signal Processing, ICASSP-98*, volume 1. IEEE, 1998. 7, 11, 55

[9] B. Atal. Efficient coding of LPC parameters by temporal decomposition. In *International Conference on Acoustics, Speech and Signal Processing, ICASSP-83*, volume 8, 1983. 25

[10] B. Atal and L. Rabiner. A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 24(3):201–212, 1976. 40

[11] R. Balan, P. Casazza, and D. Edidin. On signal reconstruction without phase. *Applied and Computational Harmonic Analysis*, 20(3):345–356, 2006. 77, 79, 162

[12] E.R. Banga, C.G. Mateo, and X.F. Salgado. Concatenative text-to-speech synthesis based on sinusoidal modelling. *Improvements in Speech Synthesis: COST 258: The Naturalness of Synthetic Speech*, pages 52–63, 2002. 36

[13] M. Basseville. Distance measures for signal processing and pattern recognition. *Signal processing*, 18(4):349–369, 1989. 132

[14] G. Baudoin and Y. Stylianou. On the transformation of the speech spectrum for voice conversion. In *Fourth International Conference on Spoken Language Processing*, 1996. 7

[15] D. Berndt and J. Clifford. Using dynamic time warping to find patterns in time series. *AAAI-94 Workshop on Knowledge Discovery in Databases (KDD-94)*, 1994. 2, 58, 151

[16] G. Birkhoff. *Lattice Theory*. American Mathematical Society Colloquium Publications, Rhode Island, 1967. 179

[17] C.M. Bishop. *Neural networks for pattern recognition*. Oxford University Press, USA, 1995. 56

[18] C.M. Bishop, M. Svensén, and C.K.I. Williams. GTM: The generative topographic mapping. *Neural computation*, 10(1):215–234, 1998. 57

[19] A. Black and P. Taylor. The festival speech synthesis system: System documentation. *Tech. Rep. HCRC/TR-83*, 1997. 5

[20] A. Black, P. Taylor, and R. Caley. The Festival speech synthesis system, 1999. 5, 16

[21] P. Boersma. Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. In *Proceedings of the institute of phonetic sciences*, volume 17, pages 97–110. Amsterdam, 1993. 40, 41, 80

[22] B.P. Bogert, M.J.R. Healy, and J.W. Tukey. The quefrency alanysis of time series for echoes: Cepstrum, pseudo-autocovariance, cross-cepstrum and saphe cracking. In *Proceedings of the Symposium on Time Series Analysis*, pages 209–243, 1963. 29

[23] A. Bonafonte, H. Höge, I. Kiss, A. Moreno, U. Ziegenhain, H. van den Heuvel, H.U. Hain, X.S. Wang, and M.N. Garcia. TC-STAR: Specifications of language resources and evaluation for speech synthesis. In *LREC*, 2006. 5, 142, 160

[24] B. Boyanov, T. Ivanov, S. Hadjitodorov, and G. Chollet. Robust hybrid pitch detector. *Electronics Letters*, 29:1924, 1993. 34

[25] J.W. Brown and R.V. Churchill. *Complex variables and applications*. McGraw-Hill New York, 1996. 63

[26] F. Buschmann, K. Henney, and D.C. Schmidt. *Pattern Oriented Software Architecture: On Patterns and Pattern Languages*, volume 6. Wiley, 2007. 69

[27] T. Ceyssens, W. Verhelst, and P. Wambacq. A strategy for pitch conversion and its evaluation. *3rd SPS-2002, Leuven, Belgium*, 2002. 11

[28] T. Ceyssens, W. Verhelst, and P. Wambacq. On the construction of a pitch conversion system. In *EUSIPCO*, pages 1301–1304, 2002. 11

[29] C.F. Chan and W.K. Hui. Wideband re-synthesis of narrowband CELP-coded speech using multiband excitation model. In *Fourth International Conference on Spoken Language Processing*. Citeseer, 1996. 33

[30] D. Chazan, R. Hoory, G. Cohen, and M. Zibulski. Speech reconstruction from mel frequency cepstral coefficients and pitch frequency. In *International Conference on Acoustics, Speech and Signal Processing, ICASSP-00*, 2000. 31

[31] A. Cheveigné and H. Kawahara. Yin, a fundamental frequency estimator for speech and music. *The Journal of the Acoustical Society of America*, 111:1917, 2002. 40

[32] D.G. Childers. Glottal source modeling for voice conversion. *Speech Communication*, 16(2):127–138, 1995. 7, 11, 39, 44

[33] D.G. Childers, B. Yegnanarayana, and Ke Wu. Voice conversion: Factors responsible for quality. *International Conference on Acoustics, Speech and Signal Processing, ICASSP-85*, pages 748–751, 1985. 2, 7

[34] R.P. Cohn. Robust voiced/unvoiced speech classification using a neural net. In *International Conference on Acoustics, Speech and Signal Processing, ICASSP-91*, pages 437–440, 1991. 40

[35] M.M.M.O. Cunha. Variação acústica das vogais orais de crianças no português europeu. 2011. 107

[36] J.R. Deller Jr., J.H.L. Hansen, and J.G. Proakis. *Discrete-Time Processing of Speech Signals*. Wiley-IEEE, 1999. 26, 30, 31, 39, 46, 79

[37] P. Delsarte and Y. Genin. The split levinson algorithm. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 34(3):470–478, 1986. 43

[38] A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–38, 1977. 24

[39] Z. Deng, M. Bulut, U. Neumann, and S. Narayanan. Automatic dynamic expression synthesis for speech animation. *Computer Animation and Social Agents 2004*, pages 267–274, 2004. 63

[40] P. Depalle and T. Helie. Extraction of spectral peak parameters using a short-time fourier transform modeling and no sidelobe windows. In *Applications of Signal Processing to Audio and Acoustics, 1997. 1997 IEEE ASSP Workshop on*, pages 4–pp. IEEE, 1997. 41

[41] S. Desai, A. Black, B. Yegnanarayana, and K. Prahallad. Spectral mapping using artificial neural networks for voice conversion. *IEEE Transactions on Audio, Speech and Language Processing*, 18(5):954–964, 2010. 11

[42] S. Desai, E.V. Raghavendra, B. Yegnanarayana, A. Black, and K. Prahallad. Voice conversion using artificial neural networks. In *submitted at IEEE workshop on Spoken Language Technologies*, 2008. 11, 56, 60, 163

[43] J. Donaldson, I. Knopke, and C. Raphael. Chroma palette: chromatic maps of sound as granular synthesis interface. In *Proceedings of the 7th International Conference on New Interfaces for Musical Expression*, pages 213–218. ACM, 2007. 170

[44] R.O. Duda, P.E. Hart, and D.G. Stork. Pattern classification. *New York: John Wiley, Section*, 1:654, 2001. 145

[45] T. Dutoit, A. Holzapfel, M. Jottrand, A. Moinet, J. Perez, and Y. Stylianou. Towards a voice conversion system based on frame selection. In *International Conference on Acoustics, Speech and Signal Processing, ICASSP-07*, pages 513–516. IEEE, 2007. 2, 11, 121, 163

[46] H. Duxans. Voice conversion applied to text-to-speech systems. *Unpublished Ph. D. thesis. Universitat Politecnica. Barcelona, Spain*, 2006. 4, 167

[47] H. Duxans, A. Bonafonte, A. Kain, and J. Santen. Including Dynamic and Phonetic Information in Voice Conversion Systems. In *Eighth International Conference on Spoken Language Processing*. ISCA, 2004. 5, 11, 57

[48] H. Duxans, D. Erro, J. Pérez, F. Diego, A. Bonafonte, and A. Moreno. Voice conversion of non-aligned data using unit selection. In *TC-STAR Workshop on Speech to Speech Translation*. Citeseer, 2006. 2, 5, 11, 54, 57, 121, 163, 167

[49] D. Erro and A. Moreno. Frame alignment method for cross-lingual voice conversion. *Proceedings of Interspeech 2007*, pages 1969–1972, 2007. 3, 5, 11, 44, 54, 119

[50] D. Erro and A. Moreno. Weighted frequency warping for voice conversion. In *Interspeech*, 2007. 5, 11, 38, 65, 163, 165, 167

[51] D. Erro, A. Moreno, and A. Bonafonte. Flexible harmonic/stochastic speech synthesis. In *6th ISCA Workshop on Speech Synthesis*, 2007. 40, 42, 72, 74, 80

[52] B.S. Everitt and D.J. Hand. *Finite mixture distributions*. Chapman and Hall, 1981. 9

[53] G. Fant. The LF-model revisited. Transformations and frequency domain analysis. *Speech Trans. Lab. Q. Rep., Royal Inst. of Tech. Stockholm*, 2:3, 1995. 21, 32, 76

[54] G. Fant. The voice source in connected speech. *Speech communication*, 22(2-3):125–139, 1997. 32, 74

[55] G. Fant, J. Liljencrants, and Q. Lin. A four-parameter model of glottal flow. *STL-QPSR*, 4(1985):1–13, 1985. 32

[56] G. Fant and Q. Lin. Frequency domain interpretation and derivation of glottal flow parameters. *STL-QPSR*, 29(2-3):1–21, 1988. 32

[57] A. Faria and D. Gelbart. Efficient pitch-based estimation of VTLN warp factors. In *Ninth European Conference on Speech Communication and Technology*. Citeseer, 2005. 28

[58] J.L. Flanagan, D.I.S. Meinhart, R.M. Golden, and M.M. Sondhi. Phase vocoder. *The Journal of the Acoustical Society of America*, 38:939, 1965. 75

[59] A.B. Fontes. *Desenvolvimento e Avaliação de Controladores Preditivos Baseados em Modelos Bilineares*. PhD thesis, Tese de Doutorado, PPgEE/UFRN, 2002. 63

[60] K. Fujii, J. Okawa, and K. Suigetsu. High-Individuality Voice Conversion Based on Concatenative Speech Synthesis. *International Journal of Computer Science and Engineering*, 2:1, 2007. 2, 11, 12, 160, 163, 164

[61] H. Fujisaki. Dynamic Characteristics of Voice Fundamental Frequency in Speech and Singing. Acoustical analysis and physiological interpretations. In *Proceedings of the 4th FASE Symposium on Acoustics and Speech*, volume 2, pages 57–70, 1981. 49

[62] H. Fujisaki. *Dynamic Characteristics of Voice Fundamental Frequency in Speech and Singing. The Production of Speech*. P.F. MacNeilage, 1983. 49

[63] H. Fujisaki, C. Wang, S. Ohno, and W. Gu. Analysis and synthesis of fundamental frequency contours of Standard Chinese using the command-response model. *Speech communication*, 47(1-2):59–70, 2005. 50

[64] T.F. Furtună. Dynamic Programming Algorithms in Speech Recognition. *Revista Informatica Economică nr*, 2:46–94, 2008. 58

[65] M.J.F. Gales. Maximum likelihood linear transformations for HMM-based speech recognition. *Computer Speech and Language*, 12:75–98, 1998. 55

[66] A. Ganapathiraju. *Support vector machines for speech recognition*. PhD thesis, Citeseer, 2000. 56

[67] A. Ganapathiraju, J. Hamaker, and J. Picone. Applications of support vector machines to speech recognition. *IEEE Transactions on Signal Processing*, 52(8):2348–2355, 2004. 56

[68] G. Garau, S. Renals, and T. Hain. Applying vocal tract length normalization to meeting recordings. In *Ninth European Conference on Speech Communication and Technology*. Citeseer, 2005. 28

[69] E. George and M. Smith. A new speech coding model based on a least-squares sinusoidal representation. In *International Conference on Acoustics, Speech and Signal Processing, ICASSP-87*, volume 12, pages 1641–1644. IEEE, 1987. 36

[70] E. George and M. Smith. Speech analysis/synthesis and modification using an analysis-by-synthesis/overlap-add sinusoidal model. *Speech and Audio Processing, IEEE Transactions on*, 5(5):389–406, 1997. 35

[71] A. Gersho. Advances in speech and audio compression. *Readings in multimedia computing and networking*, page 23, 2002. 33

[72] E. Godoy, O. Rosec, and T. Chonavel. Speech spectral envelope estimation through explicit control of peak evolution in time. In *10th International Conference on Information Sciences Signal Processing and their Applications, ISSPA-10*, pages 209–212. IEEE, 2010. 11, 24

[73] E. Godoy, O. Rosec, and T. Chonovel. Voice conversion using dynamic frequency warping with amplitude scaling, for parallel or nonparallel corpora. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(4):1313–1323, 2012. 32

[74] B. Gold and L. R. Rabiner. Parallel Processing Techniques for Estimating Pitch periods of Speech in the Time Domain. *Journal of the Acoustical Society of America*, 46(2):442–448, 1969. 51

[75] A. Goshtasby and W.D. O'Neill. Curve fitting by a sum of gaussians. *CVGIP: Graphical Model and Image Processing*, 56(4):281–288, 1994. 24, 92

[76] F.A. Graybill. Theory and Application of the Linear Model. Wadsworth and Brooks/Cole, 1976. 61

[77] D.W. Griffin, J.S. Lim, et al. Multiband excitation vocoder. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 36(8):1223–1235, 1988. 36

[78] W. Gu, K. Hirose, and H. Fujisaki. A method for automatic extraction of F 0 contour generation process model parameters for Mandarin. In *2003 IEEE Workshop on Automatic Speech Recognition and Understanding, 2003. ASRU'03*, pages 682–687. 50

[79] R.C. Guido, L. Sasso Vieira, S. Barbon Júnior, F.L. Sanchez, C. Dias Maciel, E. Silva Fonseca, and J. Carlos Pereira. A neural-wavelet architecture for voice conversion. *Neurocomputing*, 71(1-3):174–180, 2007. 11, 56

[80] J.M. Gutierrez-Arriola, Y.S. Hsiao, J.M. Montero, J.M. Pardo, and D.G. Childers. Voice conversion based on parameter transformation. In *Fifth International Conference on Spoken Language Processing*, 1998. 11, 39, 72

[81] M. Hagmüller and G. Kubin. Poincaré pitch marks. *Speech Communication*, 48(12):1650–1665, 2006. 76

[82] Z. Hanzlicek and J. Matousek. F0 transformation within the voice conversion framework. In *International Conference on Spoken Language Processing*, pages 1961–1964, 2007. 11, 12, 51

[83] Z. Hanzlicek and J. Matousek. Voice Conversion based on Probabilistic Parameter Transformation and Extended Inter-Speaker Residual Prediction. *Lecture Notes in Artificial Intelligence*, 4629:480–487, 2007. 11, 164

[84] Z. Hanzlicek and J. Matousek. On using warping function for LSFs transformation in a voice conversion system. In *9th International Conference on Signal Processing, ICSP-2008*, pages 2725–2728, 2008. 11

[85] J.C. Hardwick and J.S. Lim. Voiced/unvoiced estimation of an acoustic signal, June 1 1993. US Patent 5,216,747. 21, 40

[86] B. Hayes. Linguistics 103: Introduction to General Phonetics. 107

[87] S. Haykin. *Adaptive Filter Theory*. Prentice Hall, 1996. 25

[88] P. Hedelin. A tone oriented voice excited vocoder. In *International Conference on Acoustics, Speech and Signal Processing, ICASSP-81*, volume 6, pages 205–208. IEEE, 1981. 34

[89] E. Helander, T. Virtanen, J. Nurminen, and M. Gabbouj. Voice conversion using partial least squares regression. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(5):912–921, 2010. 11

[90] E.E. Helander and J. Nurminen. A novel method for prosody prediction in voice conversion. In *International Conference on Acoustics, Speech and Signal Processing, ICASSP-07*, volume 4, pages 509–512, 2007. 11, 12

[91] O. Helene. *Método dos Mínimos Quadrados com Formalismo Matricial*. Editora Livraria da Física, 2006. 36, 45

[92] H. Hermansky. Perceptual linear predictive (plp) analysis of speech. *The Journal of the Acoustical Society of America*, 87:1738, 1990. 24

[93] J. Högberg. Prediction of formant frequencies from linear combinations of filterbank and cepstral coefficients. *KTH-STL Quarterly Progress Rep, Royal Inst. Technol. Stockholm, Sweden*, pages 41–49, 1997. 85

[94] D. Holdcroft. *Saussure: Signs, System, and Arbitrariness*. Cambridge University Press, 1991. 177

[95] J.E. Hopcroft and R.M. Karp. An $n^{5/2}$ algorithm for maximum matchings in bipartite graphs. *SIAM Journal on ComSputing*, 2(4):225–231, 1973. 125

[96] X. Huang, A. Acero, and Hon H. *Spoken Language Processing: A guide to Theory, Algorithms, and System Development*. Prentice Hall, 2001. 31, 119

[97] K.M. Indrebo, R.J. Povinelli, and M.T. Johnson. Sub-banded reconstructed phase spaces for speech recognition. *Speech Communication*, 48(7):760–774, 2006. 21

[98] M.R. Iseli and A. Alwan. Inter-and Intra-speaker Variability of Glottal Flow Derivative using the LF Model. In *Sixth International Conference on Spoken Language Processing*, 2000. 32, 74

[99] A.K. Jain and R.C. Dubes. Algorithms for clustering data. 1988. 101

[100] A.K. Jain, M.N. Murty, and P.J. Flynn. Data clustering: a review. *ACM computing surveys (CSUR)*, 31(3):264–323, 1999. 54, 101

[101] L.L. Janer, J.J. e Bonet, and E. Lleida-Solano. Pitch detection and voiced/unvoiced decision algorithm based on wavelet transforms. In *Fourth International Conference on Spoken Language Processing*, volume 2, page 12091212, 1996. 40

[102] M.T. Johnson, R.J. Povinelli, A.C. Lindgren, J. Ye, X. Liu, and K.M. Indrebo. Time-domain isolated phoneme classification using reconstructed phase spaces. *IEEE Transactions on Speech and Audio Processing*, 13(4):458–466, 2005. 21

[103] N.L. Johnson, S. Kotz, and N. Balakrishnan. *Continuous Multivariate Distributions, volume 1, Models and Applications*. New York: John Wiley & Sons,, 2002. 59

[104] R.A. Johnson and D.W. Wichern. *Applied multivariate statistical analysis*. Prentice Hall Englewood Cliffs, NJ, 1998. 62

[105] P. Kabal and R.P. Ramachandran. The computation of line spectral frequencies using Chebyshev polynomials. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 34(6):1419–1426, 1986. 28

[106] S.G. Kafka, F.S. Pacheco, C. Seara, Izabel, R Seara, and S. Klein. Utilização de segmentos transicionais homorgânicos em síntese de fala concatenativa. *Laboratório de Circuitos e Processamento de Sinais UFSC http://linse. ufsc. br*, 2002. 22

[107] A. Kain and M.W. Macon. Spectral voice conversion for text-to-speech synthesis. In *International Conference on Acoustics, Speech and Signal Processing, ICASSP-98*, volume 1. IEEE, 1998. 4, 7, 11, 12, 59, 167

[108] A. Kain and M.W. Macon. Text-to-Speech voice adaptation from sparse training data. In *Fifth International Conference on Spoken Language Processing*. Citeseer, 1998. 4, 7, 11, 12, 163, 167

[109] A. Kain and M.W. Macon. Design and evaluation of a voice conversion algorithm based on spectral envelope mapping and residual prediction. In *International Conference on Acoustics, Speech and Signal Processing, ICASSP-01*, volume 2. IEEE, 2001. 11, 12, 38

[110] T. Kamm, G. Andreou, and J. Cohen. Vocal Tract Normalization in Speech Recognition: Compensating for Systematic Speaker Variability. In *15th Annual Speech Research Symposium*, 1995. 8, 28

[111] Y. Kang, Z. Shuang, J. Tao, W. Zhang, and B. Xu. A hybrid gmm and codebook mapping method for spectral conversion. *Lecture notes in computer science*, 3784:303, 2005. 11, 60

[112] H. Kawahara. Speech representation and transformation using adaptive interpolation of weighted spectrum: Vocoder revisited. In *International Conference on Acoustics, Speech and Signal Processing, ICASSP-97*, volume 2, pages 1303–1306, 1997. 24, 29

[113] H. Kawanami, Y. Iwami, T. Toda, H. Saruwatari, and K. Shikano. GMM-based voice conversion applied to emotional speech synthesis. In *Eighth European Conference on Speech Communication and Technology*. Citeseer, 2003. 11

[114] M. Kay, J.M. Gawron, and P. Norvig. Verbmobil: A translation system for face-to-face dialog. *CSLI Lecture Notes*, (33):748–751, 1994. 5

[115] S.A. Kibey, J.P. Kulkarni, and P.D. Sarode. A fast LSF search algorithm based on interframe correlation in G. 723.1. *EURASIP Journal on Applied Signal Processing*, pages 1107–1112, 2004. 28

[116] D.S. Kim. Perceptual phase redundancy in speech. In *International Conference on Acoustics, Speech and Signal Processing, ICASSP-00*, volume 3, pages 1383–1386. IEEE, 2000. 76

[117] D.S. Kim. On the perceptually irrelevant phase information in sinusoidal representation of speech. *IEEE Transactions on Speech and Audio Processing*, 9(8):900–905, 2001. 76

[118] E.K. Kim, S. Lee, and Y.H. Oh. Hidden Markov model based voice conversion using dynamic characteristics of speaker. In *Fifth European Conference on Speech Communication and Technology*, 1997. 11, 55, 163

[119] T. Kohonen. Self-organized formation of topologically correct feature maps. *Biological cybernetics*, 43(1):59–69, 1982. 103

[120] T. Kohonen. Self-organization and associative memory. 1988. 57

[121] T. Kohonen. *Self-organizing maps*, volume 30. Springer Verlag, 2001. 57

[122] B. Kosko and J.C. Burgess. Neural networks and fuzzy systems. *The Journal of the Acoustical Society of America*, 103:3131, 1998. 56

[123] R. Kubichek. Mel-cepstral distance measure for objective speech quality assessment. In *Communications, Computers and Signal Processing, 1993., IEEE Pacific Rim Conference on*, volume 1, pages 125–128. IEEE, 1993. 132

[124] H.W. Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2:83–97, 1955. 125

[125] A. Kumar and A. Verma. Using phone and diphone based acoustic models for voice conversion: a step towards creating voice fonts. In *International Conference on Acoustics, Speech and Signal Processing, ICASSP-03*, pages 720–723, 2003. 11, 38

[126] B.R. Kumar, J.S. Bhat, and N. Prasad. Cepstral Analysis of Voice in Persons With Vocal Nodules. *Journal of Voice*, 2010. 31

[127] P. Lanchantin and X. Rodet. Dynamic model selection for spectral voice conversion. In *Interspeech*, 2010. 11

[128] J. Laver. *Principles of Phonetics*. 1994. 21, 178

[129] K.F. Lee and H.W. Hon. Speaker-independent phone recognition using hidden Markov models. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 37(11):1641–1648, 1989. 55

[130] K.S. Lee, D.H. Youn, and I.W. Cha. A new voice transformation method based on both linear and nonlinear prediction analysis. In *Fourth International Conference on Spoken Language, ICSLP-96*, volume 3, 1996. 11, 29, 64

[131] L. Lee and R.C. Rose. Speaker normalization using efficient frequency warping procedures. In *International Conference on Acoustics, Speech and Signal Processing, ICASSP-96*, volume 1, 1996. 64

[132] A. Lepschy and G. Mian. A note on line spectral frequencies. *International Conference on Acoustics, Speech and Signal Processing, ICASSP-88*, 36(8):1355–1357, 1988. 26

[133] R.P. Lippmann. Review of neural networks for speech recognition. *Neural computation*, 1(1):1–38, 1989. 56

[134] L. Liu, J. He, and G. Palm. Effects of phase on the perception of intervocalic stop consonants1. *Speech Communication*, 22(4):403–417, 1997. 76

[135] A. Lopez, N. Aoyong, and M. Co. Voice Transformation Method using Vector Quantization. 2001. 11, 58, 70

[136] David Luebke, Mark Harris, Naga Govindaraju, Aaron Lefohn, Mike Houston, John Owens, Mark Segal, Matthew Papakipos, and Ian Buck. Gpgpu: general-purpose computation on graphics hardware. In *IEEE Conference on Supercomputing*, page 208. ACM, 2006. 169

[137] A.F. Machado, A. Bonafonte, and M Queiroz. Spectral envelope representation using sums of gaussians. In *IberSPEECH 2012 - VII Jornadas en Tecnología del Habla and III Iberian SLTech Workshop*, 2012. 24, 76, 99, 165, 168

[138] A.F. Machado, A. Bonafonte, and M. Queiroz. Parametric decomposition of the spectral envelope. In *International Conference on Acoustics, Speech and Signal Processing, ICASSP-13*, 2013. 88, 92, 99, 169

[139] A.F. Machado and M. Queiroz. Voice conversion: A critical survey. *Proceedings of the 7th Sound and Music Computing Conference - SMC*, pages 291–298. 8, 9, 168

[140] A.F. Machado and M. Queiroz. Techniques for crosslingual voice conversion. In *Proceedings of the IEEE International Symposium on Multimedia, ISM-2010*, pages 365–370. IEEE, 2010. 8, 9, 168

[141] M.W. Macon. Speech synthesis based on sinusoidal modeling. 1996. 35

[142] R. Maesschalck, D. Jouan-Rimbaud, and DL Massart. The mahalanobis distance. *Chemometrics and Intelligent Laboratory Systems*, 50(1):1–18, 2000. 132, 152

[143] J. Makhoul. Linear prediction: A tutorial review. *Proceedings of the IEEE*, 63(4):561–580, 1975. 25

[144] J. Makhoul. Stable and efficient lattice methods for linear prediction. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 25(5):423–428, 1977. 25

[145] M. Marcus and H. Minc. *Elementary linear algebra*. New York: The Macmillan Company; London: Collier-Macmillan Ltd, 1968. 63

[146] J. Markel. The SIFT algorithm for fundamental frequency estimation. *IEEE Transactions on Audio and Electroacoustics*, 20(5):367–377, 1972. 40

[147] D.W. Marquardt. An algorithm for least-squares estimation of nonlinear parameters. *Journal of the society for Industrial and Applied Mathematics*, 11(2):431–441, 1963. 96

[148] M. Mashimo, T. Toda, H. Kawanami, H. Kashioka, K. Shikano, and N. Campbell. Evaluation of cross-language voice conversion using bilingual e non-bilingual databases. *Interspeech*, pages 293–296, 2002. 8, 10, 122

[149] M. Mashimo, T. Toda, K. Shikano, and N. Campbell. Evaluation of cross-language voice conversion based on GMM and STRAIGHT. In *Seventh European Conference on Speech Communication and Technology*, 2001. 11, 12

[150] J.J. Matras. O som. *São Paulo: Ed. Martins Fontes*, 1991. 76

[151] R. McAulay and T. Quatieri. Magnitude-only reconstruction using a sinusoidal speech model. In *International Conference on Acoustics, Speech and Signal Processing, ICASSP-84*, volume 9, pages 441–444. IEEE, 1984. 76

[152] R. McAulay and T. Quatieri. Speech analysis/synthesis based on a sinusoidal representation. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 34(4):744–754, 1986. 35, 41, 45

[153] G.J. McLachlan and K.E. Basford. *Mixture Models: Inference and Applications to Clustering*. Marcel Dekker, 1988. 9

[154] G.J. McLachlan and T. Krishnan. *The EM algorithm and extensions*. Wiley New York, 1997. 59

[155] L. Mesbahi, V. Barreaud, and O. Boeffard. Comparing GMM-based speech transformation systems. In *Proceedings of Interspeech*, pages 1989–1992, 2007. 12, 60, 103

[156] H. Mizuno and M. Abe. Voice conversion algorithm based on piecewise linear conversion rules of formant frequency and spectrum tilt. *Speech Communication*, 16(2):153–164, 1995. 28

[157] S. Molau, M. Pitz, R. Schluter, and H. Ney. Computing Mel-frequency cepstral coefficients on the power spectrum. In *International Conference on Acoustics, Speech and Signal Processing, ICASSP-01*, volume 1. Citeseer, 2001. 64

[158] M.C. Monteiro. *Uma análise espectrográfica das formantes das vogais orais do português brasileiro falado em São Paulo*. PhD thesis, Escola Paulista de Medicina – São Paulo, 1995. 107

[159] T.K. Moon. The expectation-maximization algorithm. *Signal Processing Magazine, IEEE*, 13(6):47–60, 1996. 119

[160] M. Morf, B. Dickinson, T. Kailath, and A. Vieira. Efficient solution of covariance equations for linear prediction. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 25(5):429–433, 1977. 25

[161] D.P. Morgan and C.L. Scofield. *Neural networks and speech processing*. Kluwer Academic Publishers Norwell, MA, USA, 1991. 60

[162] R. Mori. *Computer models of speech using fuzzy algorithms*. Perseus Publishing, 1983. 56

[163] R.W. Morris and M.A. Clements. Modification of formants in the line spectrum domain. *IEEE Signal Processing Letters*, 9(1):19–21, 2002. 26

[164] E. Moulines and F. Charpentier. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication*, 9(5-6):453–467, 1990. 16, 38

[165] E. Moulines and W. Verhelst. *Time-Domain and Frequency-Domain Techniques for Prosodic Modification of Speech*. Elsevier Science B.V., 1995. 34, 48, 51, 75

[166] N. Murata, S. Ikeda, and A. Ziehe. An approach to blind source separation based on temporal structure of speech signals. *Neurocomputing*, 41(1-4):1–24, 2001. 39, 72

[167] K.S.R. Murty and B. Yegnanarayana. Combining evidence from residual phase and MFCC features for speaker recognition. *IEEE signal processing letters*, 13(1):52–55, 2006. 31, 79, 133, 153

[168] B.R. Musicus. *Levinson and fast Choleski algorithms for Toeplitz and almost Toeplitz matrices*. Citeseer, 1988. 144

[169] M. Narendranath, H.A. Murthy, S. Rajendran, and B. Yegnanarayana. Transformation of formants for voice conversion using artificial neural networks. *Speech Communication*, 16(2):207–216, 1995. 7, 11, 12, 28, 44, 60

[170] S. Narusawa, N. Minematsu, K. Hirose, and H. Fujisaki. Automatic extraction of model parameters from fundamental frequency contours of English utterances. In *Seventh International Conference on Spoken Language Processing*, 2002. 50

[171] E.P. Neuburg. Dynamic frequency warping, the dual of dynamic time warping. *The Journal of the Acoustical Society of America*, 81(S1):S94–S94, 1987. 9, 65

[172] B.P. Nguyen and M. Akagi. Spectral modification for voice gender conversion using temporal decomposition. *Journal of Signal Processing*, 2007. 32

[173] M. Nishiguchi, J. Matsumoto, and S. Ono. Voiced/unvoiced decision based on frequency band ratio, September 28 1999. US Patent 5,960,388. 40

[174] A.M. Noll. Clipstrum pitch determination. *The journal of the acoustical society of America*, 44:1585, 1968. 29

[175] J. Nurminen, V. Popa, J. Tian, Y. Tang, and I. Kiss. A parametric approach for voice conversion. In *TC-STAR Workshop on Speech-to-Speech Translation*, pages 225–229, 2006. 5, 11, 163

[176] J. Nurminen, J. Tian, and V. Popa. Novel Method for Data Clustering and Mode Selection with Application in Voice Conversion. In *Ninth International Conference on Spoken Language Processing*. Citeseer, 2006. 5, 11, 54, 120

[177] D. O'Brien and A.I.C. Monaghan. Concatenative synthesis based on a harmonic model. *Speech and Audio Processing, IEEE Transactions on*, 9(1):11–20, 2001. 36, 45

[178] M.K. Omar, M. Hasegawa-Johnson, and S.E. Levinson. Gaussian mixture models of phonetic boundaries for speech recognition. In *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. Citeseer, 2001. 59

[179] A. V. Oppenheim and R. W. Schafer. *Discrete-Time Signal Processing*. Prentice-Hall, 1989. 25, 29, 31, 43

[180] A.V. Oppenheim. A speech analysis-synthesis system based on homomorphic filtering. *Journal of the Acoustical Society of America*, 45(2):293–309, 1969. 40

[181] A.V. Oppenheim and R.W. Schafer. From frequency to quefrency: A history of the cepstrum. *IEEE signal processing Magazine*, 21(5):95–106, 2004. 29, 41

[182] C. Orphanidou, I.M. Moroz, and S.J. Roberts. Voice morphing using the generative topographic mapping. *Proceedings of CCCT 03*, 1:222–225, 2003. 11, 57

[183] C. Orphanidou, I.M. Moroz, and S.J. Roberts. Wavelet-based voice morphing. *WSEAS Transactions on Systems*, 3(10):3297–3302, 2004. 11, 164

[184] R. Oxford. Language learning strategies: An update. *Eric Digest*, pages 95–02, 1994. 5

[185] K.K. Paliwal. A study of LSF parameters for speaker recognition. *The Journal of the Acoustical Society of America*, 87:S108, 1990. 26

[186] K.K. Paliwal. On the use of line spectral frequency parameters for speech recognition. *Digital Signal Processing*, 2(2):80–87, 1992. 26

[187] S. Panchapagesan and A. Alwan. Frequency warping for VTLN and speaker adaptation by linear transformation of standard MFCC. *Computer Speech and Language*, 23(1):42–64, 2009. 28, 29, 31, 64

[188] I. Patel and Y.S. Rao. Speech Recognition Using Hidden Markov Model with MFCC-Subband Technique. In *International Conference on Recent Trends in Information, Telecommunication and Computing*, pages 168–172. IEEE, 2010. 31, 133

[189] H.R. Pfitzinger. Unsupervised speech morphing between utterances of any speakers. In *10th Australian Int. Conf. on Speech Science and Technology (SST 2004)*, pages 545–550. Citeseer, 2004. 11, 62, 163

[190] M. Pitz and H. Ney. Vocal tract normalization as linear transformation of MFCC. In *Eighth European Conference on Speech Communication and Technology*. Eurospeech, 2003. 29, 31, 153

[191] M. Pitz and H. Ney. Vocal tract normalization equals linear transformation in cepstral space. *IEEE Trans. Speech and Audio Processing*, 13(5):930–944, 2005. 29

[192] H. Pobloth and W.B. Kleijn. On phase perception in speech. In *International Conference on Acoustics, Speech, and Signal Processing, ICASSP-99*, volume 1, pages 29–32. IEEE, 1999. 76, 174

[193] V. Popa, J. Nurminen, and M. Gabbouj. A Novel Technique for Voice Conversion Based on Style and Content Decomposition with Bilinear Models. 2009. 8, 11, 63, 64

[194] A. Pozo and S. Young. The linear transformation of LF glottal waveforms for voice conversion. In *Interspeech*, pages 1457–1460, 2008. 8, 11, 163

[195] A. Přibilová and R. Vích. Non-linear frequency scale mapping for voice conversion. In *14th Int. Czech-Slovak Scientific Conf. Radioelektronika, Bratislava, Slovak Republic*, pages 100–103, 2004. 8, 11

[196] M.S. Puckette. Phase bashing for sample-based formant synthesis. In *Proceedings of the International Computer Music Conference*, pages 733–736, 2005. 76, 174

[197] T.F. Quatieri and R.J. McAulay. Shape invariant time-scale and pitch modification of speech. *IEEE Transactions on Signal Processing*, 40(3):497–510, 1992. 76

[198] A. Quilis. *Bibliografía de fonética y fonología españolas*, volume 9. CSIC-Dpto. de Publicaciones, 1984. 107

[199] L. Rabiner and M. Sambur. Application of an LPC distance measure to the voiced-unvoiced-silence detection problem. *IEEE Transactions on Acoustics Speech and Signal Processing*, 25(4):338–343, 1977. 40

[200] L. Rabiner and M. Sambur. Voiced-unvoiced-silence detection using the Itakura LPC distance measure. In *International Conference on Acoustics, Speech and Signal Processing, ICASSP-77*, volume 2, 1977. 40, 132

[201] L.R. Rabiner. A tutorial on hidden markov model and selected applications in speech recognition. *Proceedings of the IEEE*, (77), 1989. 55

[202] B. Rainer, M. Dell'Amico, and S. Martello. Assignment problems. *Philadelphia, Pennsylvania: SIAM*, page 382, 2009. 125

[203] K.S. Rao and B. Yegnanarayana. Voice conversion by prosody and vocal tract modification. In *Proceedings of the 9th International Conference on Information Technology*, pages 111–116. IEEE Computer Society, 2006. 8, 11, 61, 163

[204] C.H. Reinsch. Smoothing by spline functions. *Numerische Mathematik*, 10(3):177–183, 1967. 86

[205] S. Renals, N. Morgan, H. Bourlard, M. Cohen, and H. Franco. Connectionist probability estimators in HMM speech recognition. *IEEE Transactions on Speech and Audio Processing*, 2(1 Part 2):161–174, 1994. 55

[206] D. Rentzos, S. Vaseghi, E. Turajlic, Q. Yan, and C.H. Ho. Transformation of speaker characteristics for voice conversion. In *IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 706–711, 2003. 11, 163

[207] D. Rentzos, S. Vaseghi, Q. Yan, and C.H. Ho. Voice conversion through transformation of spectral and intonation features. In *International Conference on Acoustics, Speech and Signal Processing, ICASSP-04*, volume 1, 2004. 8, 11

[208] D.A. Reynolds, T.F. Quatieri, and R.B. Dunn. Speaker verification using adapted Gaussian mixture models. *Digital signal processing*, 10(1-3):19–41, 2000. 59

[209] A. Rinscheid. Voice conversion based on topological feature maps and time-variant filtering. In *Fourth International Conference on Spoken Language Processing*. Citeseer, 1996. 11, 56

[210] A.E. Rosenberg. Effect of glottal pulse shape on the quality of natural vowels. *J. Acoust. Soc. Am*, 49(2):583–590, 1971. 32

[211] P.S. Rossi, P. Falco, A. Budillon, D. Mattera, and F. Palieri. Prosody modification and Fujisaki's model: Preserving natural soundness. In *EUSIPCO. Conference*, 2004. 49

[212] J. Rothweiler. On polynomial reduction in the computation of LSP frequencies. *IEEE Transactions on Speech and Audio Processing*, 7(5):592–594, 1999. 28

[213] K. Russel. Formants, 2005. [Online; accessed 17-12-2012]. 107

[214] T. Saitou, M. Goto, M. Unoki, and M. Akagi. Vocal Conversion from Speaking Voice to Singing Voice Using STRAIGHT. *Interspeech-07, TuC. SS-2*, 2007. 29

[215] F.L. Sanchez, B. Júnior, et al. Wavelet-based cepstrum calculation. *Journal of Computational and Applied Mathematics*, 227(2):288–293, 2009. 34

[216] S. Saoudi, B. Alain, and J. Marc. A new efficient algorithm to compute the LSP parameters for speech coding. *Signal Processing*, 28(2):201–212, 1992. 28

[217] I. Saratxaga, I. Hernaez, D. Erro, E. Navas, and J. Sanchez. Simple representation of signal phase for harmonic speech models. *Electronics letters*, 45(7):381–383, 2009. 79

[218] I. Saratxaga, I. Hernaez, M. Pucher, E. Navas, and I. Sainz. Perceptual importance of the phase related information in speech. *a a*, 2:2. 79

[219] Y. Sato. Voice Conversion using interactive evolution of prosodic control. In *Proceedings of the 2002 Genetic and Evolutionary Computation Conference (GECCO-2002), Morgan Kaufmann Publishers, San Francisco, CA*, pages 1204–1211, 2002. 13, 38

[220] F. Saussure. *Curso de lingüística geral*. Editora Cultrix, 2008. 16

[221] M.I. Savic, S.H. Tan, and I.H. Nam. Speech transformation system, July 5 1994. US Patent 5,327,521. 12

[222] M. Schuster, T. Haderlein, E. Nöth, J. Lohscheller, U. Eysholdt, and F. Rosanowski. Intelligibility of laryngectomees substitute speech: automatic speech recognition and subjective rating. *European Archives of Oto-Rhino-Laryngology*, 263(2):188–193, 2006. 5, 16

[223] J. Serra. *Image Analysis and Mathematical Morphology*. Academic Press, New York, 1982. 179

[224] J. Serra. *Image Analysis and Mathematical Morphology. Volume 2: Theoretical Advances*. Academic Press, 1988. 179

[225] X. Serra. Musical sound modeling with sinusoids plus noise. *Musical signal processing*, pages 91–122, 1997. 36, 76

[226] Xavier Serra. *A system for sound analysis/transformation/synthesis based on a deterministic plus stochastic decomposition*. Number 58. Department of Music, Stanford University, 1989. 36, 76

[227] Y. Shiga and S. King. Estimating the spectral envelope of voiced speech using multi-frame analysis. 2003. 89, 174

[228] Y. Shiga and S. King. Estimation of voice source and vocal tract characteristics based on multi-frame analysis. 2003. 89

[229] Y. Shiga and S. King. Accurate spectral envelope estimation for articulation-to-speech synthesis. In *Fifth ISCA Workshop on Speech Synthesis*, 2004. 89

[230] C. Shih and R. Sproat. Issues in text-to-speech conversion for Mandarin. *Computational Linguistics and Chinese Language Processing*, 1(1):37–86, 1996. 10

[231] K. Shikano, K.F. Lee, and R. Reddy. Speaker adaptation through vector quantization. In *International Conference on Acoustics, Speech and Signal Processing, ICASSP-86*, volume 11, 1986. 7, 11, 58, 70

[232] A.N. Shiryaev. *Probability*. Number 2. Springer 1996. 39

[233] Z. Shuang, R. Bakis, and Y. Qin. Voice Conversion Based On Mapping Formants. In *TC-STAR Workshop on Speech-to-Speech Translation. Barcelona, Spain*, pages 219–223, 2006. 5, 11, 163

[234] Z. Shuang, R. Bakis, and Y. Qin. IBM Voice Conversion Systems for 2007 TC-STAR Evaluation. *Tsinghua Science & Technology*, 13(4):510–514, 2008. 5, 11, 163, 167

[235] L. Siegel and A. Bessey. Voiced/unvoiced/mixed excitation classification of speech. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 30(3):451–460, 1982. 40

[236] E. Singer and R.P. Lippman. A speech recognizer using radial basis function neural networks in an HMM framework. In *International Conference on Acoustics, Speech and Signal Processing, ICASSP-92*, volume 1, 1992. 56, 86

[237] J.O. Smith and J.S. Abel. Bark and ERB bilinear transforms. *IEEE Transactions on Speech and Audio Processing*, 7(6):697–708, 1999. 63

[238] D. Sündermann. Voice Conversion: State-of-the-Art and Future Work. *Fortschritte der Akustik*, 31(2):735, 2005. 12, 159

[239] D. Sündermann, A. Bonafonte, H. Höge, AG Siemens, H. Ney, and R. Aachen. Voice conversion using exclusively unaligned training data. In *Spanish Society for Natural Language Processing Conference, Barcelona, Spain*, pages 41–48, 2004. 8

[240] D. Sündermann, A. Bonafonte, and H. Ney. A first step towards text-independent voice conversion. *Proceedings of Interspeech*, 2004. 8

[241] D. Sündermann, A. Bonafonte, H. Ney, and H. Höge. A first step towards text-independent voice conversion. In *ICSLP-04*, 2004. 3, 135

[242] D. Sündermann, H. Höge, A. Bonafonte, H. Ney, and J. Hirschberg. Tc-star: Cross-language voice conversion revisited. *Proceedings of the TC-STAR Workshop on Speech-to-Speech Translation*, pages 231–236, 2006. 5, 160

[243] D. Sündermann, H. Höge, A. Bonafonte, H. Ney, and J. Hirschberg. Text-independent cross-language voice conversion. In *Ninth International Conference on Spoken Language Processing*, 2006. 4, 54

[244] D. Sündermann, H. Hoge, A. Bonafonte, H. Ney, A. Black, and S. Narayanan. Text-independent voice conversion based on unit selection. In *International Conference on Acoustics, Speech and Signal Processing, ICASSP-06*. Citeseer, 2006. 4, 5, 11, 54, 163

[245] D. Sündermann and H. Ney. An automatic segmentation and mapping approach for voice conversion parameter training. *AST-03, Maribor, Slovenia*, 2003. 8, 11

[246] D. Sündermann and H. Ney. VTLN-based voice conversion. *power*, 8(7):8α, 2003. 8, 11, 28, 38

[247] D. Sündermann, H. Ney, and H. Hoge. VTLN-based cross-language voice conversion. In *ASRU*. Citeseer, 2003. 3, 5, 8, 11, 28, 38

[248] D. Sündermann, G. Strecha, A. Bonafonte, H. Höge, and H. Ney. Evaluation of VTLN-Based Voice Conversion for Embedded Speech Synthesis. In *Interspeech*, 2005. 12, 28

[249] P. Song, Y. Jin, L. Zhao, and C. Zou. Voice conversion based on hybrid svr and gmm. *Archives of Acoustics*, 37(2):143–149, 2012. 11

[250] F. Soong and B. Juang. Line spectrum pair (LSP) and speech data compression. In *International Conference on Acoustics, Speech and Signal Processing, ICASSP-84*, volume 9, 1984. 27

[251] A.S. Spanias. Speech coding: A tutorial review. *Proceedings of the IEEE*, 82(10):1541–1582, 1994. 33

[252] S.S. Stevens, J. Volkman, and E. Newman. A scale for the measurement of the psychological magnitude pitch. *Journal of the Acoustical Society of America*, 8(3):185–190, 1937. 18

[253] T.G. Stockham Jr, T.M. Cannon, and R.B. Ingebretsen. Blind deconvolution through digital signal processing. *Proceedings of the IEEE*, 63(4):678–692, 1975. 29

[254] A. Stolcke, L. Ferrer, S. Kajarekar, E. Shriberg, and A. Venkataraman. MLLR transforms as features in speaker recognition. In *Ninth European Conference on Speech Communication and Technology*. ISCA, 2005. 16, 62

[255] Y. Stylianou. Harmonic plus noise models for speech, combined with statistical methods, for speech and speaker modification. 1996. 34, 37, 41, 74

[256] Y. Stylianou. Concatenative speech synthesis using a harmonic plus noise model. In *The Third ESCA/COCOSDA Workshop (ETRW) on Speech Synthesis*. Citeseer, 1998. 37

[257] Y. Stylianou. Continuous probabilistic transform for voice conversion. *IEEE Transactions on Speech and Audio Processing*, (6):131–142, 1998. 7, 11, 12, 59, 163, 164

[258] Y. Stylianou. On the implementation of the harmonic plus noise model for concatenative speech synthesis. In *International Conference on Acoustics, Speech and Signal Processing, ICASSP-00*, volume 2. IEEE; 1999, 2000. 16, 37, 40, 74, 80

[259] Y. Stylianou. Removing linear phase mismatches in concatenative speech synthesis. *IEEE Transactions on Speech and Audio Processing*, 9(3):232–239, 2001. 76, 78

[260] Y. Stylianou. Voice transformation: A survey. *International Conference on Acoustics, Speech and Signal Processing, ICASSP-09*, pages 3585–3588, 2009. 11, 12

[261] Y. Stylianou, O. Cappe, and E. Moulines. Statistical methods for voice quality transformation. In *Fourth European Conference on Speech Communication and Technology*. ISCA, 1995. 59

[262] D. Talkin. A robust algorithm for pitch tracking (RAPT). *Speech coding and synthesis*, 495:518, 1995. 34, 40

[263] K. Tanaka and M. Abe. A new fundamental frequency modification algorithm with transformation of spectrum envelope according to F0. In *International Conference on Acoustics, Speech and Signal Processing, ICASSP-97*, volume 2, 1997. 12

[264] P. Taylor, A. Black, and R. Caley. The architecture of the Festival speech synthesis system. In *The Third ESCA/COCOSDA Workshop (ETRW) on Speech Synthesis*, 1998. 5, 16

[265] J. Tebelskis. *Speech recognition using neural networks*. PhD thesis, Citeseer, 1995. 60

[266] J.B. Tenenbaum and W.T. Freeman. Separating style and content with bilinear models. *Neural Computation*, 12(6):1247–1283, 2000. 63

[267] D. Titterington, A. Smith, and U. Makov. *Statistical Analysis of Finite Mixture Distributions*. John Wiley and Sons, 1985. 9, 103

[268] T. Toda, A. Black, and K. Tokuda. Spectral conversion based on maximum likelihood estimation considering global variance of converted parameter. In *International Conference on Acoustics, Speech and Signal Processing, ICASSP-05*, volume 1, pages 9–12. Citeseer, 2005. 11, 12, 103, 163, 164

[269] T. Toda, A. Black, and K. Tokuda. Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory. *IEEE Transactions on Audio Speech and Language Processing*, 15(8):2222, 2007. 12

[270] T. Toda, H. Saruwatari, and K. Shikano. Voice conversion algorithm based on Gaussian mixture model with dynamic frequency warping of STRAIGHT spectrum. *Power [dB]*, 30:40, 2001. 7, 11, 12, 64, 135, 163, 164

[271] K. Tokuda, T. Masuko, N. Miyazaki, and T. Kobayashi. Hidden Markov models based on multi-space probability distribution for pitch pattern modeling. In *International Conference on Acoustics, Speech and Signal Processing, ICASSP-99*, volume 1, pages 229–232, 1999. 55

[272] K. Tokuda, T. Masuko, N. Miyazaki, and T. Kobayashi. Multi-space probability distribution HMM. *IEICE Transaction on Informatiom and Systems E Series D*, 85(3):455–464, 2002. 55

[273] P.A. Torres-Carrasquillo, D.A. Reynolds, and J.R. Deller. Language identification using Gaussian mixture model tokenization. In *International Conference on Acoustics, Speech and Signal Processing, ICASSP-02*, volume 1. IEEE, 2002. 59

[274] O. Türk. *Cross-Lingual Voice Conversion*. Bogaziçi University, 2007. 3

[275] O. Türk and L.M. Arslan. Subband based voice conversion. In *Seventh International Conference on Spoken Language Processing*. Citeseer, 2002. 8, 11, 12, 38, 159, 164

[276] O. Türk and L.M. Arslan. Voice conversion methods for vocal tract and pitch contour modification. In *Eighth European Conference on Speech Communication and Technology*, 2003. 11

[277] O. Türk and L.M. Arslan. Robust processing techniques for voice conversion. *Computer Speech & Language*, 20(4):441–467, 2006. 4

[278] O. Türk and L.M. Arslan. Donor Selection for Voice Conversion. 2007. 13

[279] O. Türk and M. Schröder. A Comparison of Voice Conversion Methods for Transforming Voice Quality in Emotional Speech Synthesis. *Interspeech, 2008*, pages 2282–2285. 8, 13

[280] O. Türk, M Schröder, B Bozkurt, and L.M. Arslan. Voice Quality Interpolation for Emotional Text-To-Speech Synthesis. pages 797–800, 2005. 13

[281] B. Tuller and JAS Kelso. The production and perception of syllable structure. *Journal of Speech and Hearing Research*, 34(3):501, 1991. 39, 72

[282] A.J. Uriz, P.D. Aguero, A. Bonafonte, and J.C. Tulli. Voice conversion using k-histograms and frame selection. *Proceeding of Interspeech 2009*, 2009. 8, 11, 12, 54, 55, 121

[283] A.J. Uriz, P.D. Aguero, J.C. Tulli, E. Gonzalez, and A. Bonafonte. Voice conversion using frame selection and warping functions. *Proceedings of RPIC 2009. ISBN: 950-665-340*, 2:417–422, 2009. 8, 11, 12, 55, 64

[284] H. Valbret, E. Moulines, and J.P. Tubach. Voice Tranformation Using PSOLA Technique. In *Second European Conference on Speech Communication and Technology*, 1991. 7, 11, 62

[285] J. Van den Berg. Myoelastic-aerodynamic theory of voice production. *Journal of Speech, Language and Hearing Research*, 1(3):227, 1958. 19

[286] R. Veldhuis. A computationally efficient alternative for the Liljencrants–Fant model and its perceptual evaluation. *The Journal of the Acoustical Society of America*, 103:566, 1998. 33

[287] W. Verhelst and J. Mertens. Voice conversion using partitions of spectral feature space. In *International Conference on Acoustics, Speech and Signal Processing, ICASSP-96*, volume 1, 1996. 11

[288] L.S. Vieira. *Conversão de voz baseada na transformada wavelet*. PhD thesis, Depto. Eng. Elétrica/Universidade de São Carlos - SP - Brasil, 2007. 4, 5

[289] M.N. Vieira. Uma introdução à acústica da voz cantada. *Seminário Música Ciência Tecnologia*, 1(1), 2004. 22

[290] S. Wagner. Intralingual speech-to-text-conversion in real-time: Challenges and Opportunities. In *EU High Level Scientific Conference Series*, page 210. 3

[291] D. Wang and J. Lim. The unimportance of phase in speech enhancement. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 30(4):679–681, 1982. 76

[292] T. Watanabe, T. Murakami, M. Namba, T. Hoya, and Y. Ishida. Transformation of spectral envelope for voice conversion based on radial basis function networks. In *Seventh International Conference on Spoken Language Processing*. ISCA, 2002. 11, 86

[293] S. Wegmann, D. McAllaster, J. Orloff, and B. Peskin. Speaker normalization on conversational telephone speech. In *International Conference on Acoustics, Speech and Signal Processing, ICASSP-96*, volume 1, 1996. 28

[294] M.M. Wilde and A.B. Martinez. Probabilistic principal component analysis applied to voice conversion. In *IEEE Conference Record of the Thirty-Eighth Asilomar Conference on Signals, Systems and Computers. California, USA*, pages 2255–2259. Citeseer, 2004. 11, 60

[295] J. Wouters and M.W. Macon. A perceptual evaluation of distance measures for concatenative speech synthesis. In *ICSLP*, 1998. 12

[296] W. Yang. *Enhanced modified bark spectral distortion (EMBSD): An objective speech quality measure based on audible distortion and cognition model*. PhD thesis, Temple University, 1999. 59, 132

[297] W. Yang, M. Benbouchta, and R. Yantorno. Performance of the modified bark spectral distortion as an objective speech quality measure. In *International Conference on Acoustics, Speech and Signal Processing, ICASSP-98*, volume 1, pages 541–544. IEEE, 1998. 132

[298] H. Ye and S. Young. Perceptually weighted linear transformations for voice conversion. In *Eighth European Conference on Speech Communication and Technology*, 2003. 11, 12, 63, 133

[299] H. Ye and S. Young. High quality voice morphing. In *International Conference on Acoustics, Speech and Signal Processing, ICASSP-04*. Citeseer, 2004. 8, 12

[300] H. Ye and S. Young. Voice conversion for unknown speakers. In *Eighth International Conference on Spoken Language Processing*, 2004. 8, 11, 12

[301] H. Ye and S. Young. Quality-enhanced voice morphing using maximum likelihood transformations. *IEEE Transactions on Audio, Speech and Language Processing*, 14(4):1301–1312, 2006. 11, 32, 60, 160, 164

[302] S.S. Yedlapalli. Transforming Real Linear Prediction Coefficients to Line Spectral Representations With a Real FFT. *IEEE Transactions on Speech and Audio Processing*, 13(5):733–740, 2005. 28

[303] C.A. Ynoguti and F. Violaro. On the use of principal component analysis over mel cepstral coefficients. *Revista Científica*, 1516:2338, 2002. 31

[304] T. Yonezawa, N. Suzuki, S. Abe, K. Mase, and K. Kogure. Perceptual continuity and naturalness of expressive strength in singing voices based on speech morphing. *EURASIP Journal on Audio, Speech, and Music Processing*, 2007(3):2, 2007. 4

[305] Z. Yue, X. Zou, Y. Jia, and H. Wang. Voice Conversion Using HMM combined with GMM. In *Image and Signal Processing, 2008. CISP'08. Congress on*, volume 5, 2008. 8, 11, 44, 55, 164

[306] K. Yutani, Y. Uto, Y. Nankaku, A. Lee, and K. Tokuda. Voice conversion based on simultaneous modelling of spectrum and F0. In *International Conference on Acoustics, Speech and Signal Processing, ICASSP-09*, pages 3897–3900. IEEE, 2009. 11, 55

[307] S.A. Zahorian and H. Hu. A spectral/temporal method for robust fundamental frequency tracking. *The Journal of the Acoustical Society of America*, 123:4559, 2008. 41

[308] P. Zhan and A. Waibel. Vocal tract length normalization for large vocabulary continuous speech recognition. *CMU Computer Science Technical Reports*, 1997. 16, 28

[309] P. Zhan and M. Westphal. Speaker normalization based on frequency warping. In *International Conference on Acoustics, Speech and Signal Processing, ICASSP-97*, volume 2, 1997. 64

[310] J. Zhang, J. Sun, and B. Dai. Voice conversion based on weighted least squares estimation criterion and residual prediction from pitch contour. *Lecture notes in computer science*, 3784:326, 2005. 11, 38, 164

[311] M. Zhang, J. Tao, J. Nurminen, J. Tian, and X. Wang. Phonetic anchor based state mapping for text-independent voice conversion. In *9th International Conference on Signal Processing, ICSP-2008*, pages 723–727, 2008. 11

[312] M. Zhang, J. Tao, J. Nurminen, J. Tian, and X. Wang. Phoneme cluster based state mapping for text-independent voice conversion. In *International Conference on Acoustics, Speech and Signal Processing, ICASSP-09*, pages 4281–4284. IEEE, 2009. 11, 54, 160, 163, 164

[313] M. Zhang, J. Tao, J. Tian, and X. Wang. Text-independent voice conversion based on state mapped codebook. In *International Conference on Acoustics, Speech and Signal Processing, ICASSP-08*, 2008. 8, 11, 163, 164

[314] W. Zhang, LQ Shen, and D. Tang. Voice conversion based on acoustic feature transformation. In *6th National Conference on Man-Machine Speech Communications*, 2001. 11, 62

[315] F. Zheng, G. Zhang, and Z. Song. Comparison of different implementations of MFCC. *Journal of Computer Science and Technology*, 16(6):582–589, 2001. 31

[316] P. Zolfaghari and T. Robinson. Formant analysis using mixtures of gaussians. In *Fourth International Conference on Spoken Language, ICSLP-96*, volume 2, pages 1229–1232. IEEE, 1996. 24, 31

[317] P. Zolfaghari, S. Watanabe, A. Nakamura, and S. Katagiri. Bayesian modelling of the speech spectrum using mixture of gaussians. In *International Conference on Acoustics, Speech and Signal Processing, ICASSP-04*, volume 1, pages I–553. IEEE, 2004. 24

[318] T.C. Zorilă, D. Erro, and I. Hernaez. Improving the quality of standard gmm-based voice conversion systems by considering physically motivated linear transformations. *Advances in Speech and Language Technologies for Iberian Languages*, pages 30–39, 2012. 11

[319] E. Zwicker. Subdivision of the audible frequency range into critical bands. *The Journal of the Acoustical Society of America*, 33:248, 1961. 18