DESENVOLVIMENTO DE UM SISTEMA DE CONVERSÃO TEXTO-FALA COM MODELAGEM DE PROSÓDIA

Vagner Luis Latsch

Tese de Doutorado apresentada ao Programa de Pós-graduação em Engenharia Elétrica, COPPE, da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do título de Doutor em Engenharia Elétrica.

Orientador: Sergio Lima Netto

Rio de Janeiro
Junho de 2011

DESENVOLVIMENTO DE UM SISTEMA DE CONVERSÃO TEXTO-FALA COM MODELAGEM DE PROSÓDIA

Vagner Luis Latsch

TESE SUBMETIDA AO CORPO DOCENTE DO INSTITUTO ALBERTO LUIZ COIMBRA DE PÓS-GRADUAÇÃO E PESQUISA DE ENGENHARIA (COPPE) DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE DOUTOR EM CIÊNCIAS EM ENGENHARIA ELÉTRICA.

Examinada por:

Prof. Sergio Lima Netto, Ph.D.

Prof. Luis Pereira Calôba, Dr.Ing.

Prof. João Antonio de Moraes, Ph.D.

Prof. Rui Seara, Ph.D.

Prof. Izabel Christine Seara, Ph.D.

RIO DE JANEIRO, RJ – BRASIL
JUNHO DE 2011

Latsch, Vagner Luis

Desenvolvimento de um Sistema de Conversão Texto-Fala com Modelagem de Prosódia/Vagner Luis Latsch. – Rio de Janeiro: UFRJ/COPPE, 2011.

X, 156 p. 29, 7cm.

Orientador: Sergio Lima Netto

Tese (doutorado) – UFRJ/COPPE/Programa de Engenharia Elétrica, 2011.

Referências Bibliográficas: p. 126 – 132.

1. Conversão-texto-fala. 2. Prosódia. I. Netto, Sergio Lima. II. Universidade Federal do Rio de Janeiro, COPPE, Programa de Engenharia Elétrica. III. Título.

Dedico esta tese à Adriana,
Luisa e Júlia.

Agradecimentos

Esta tese de doutorado é o fechamento de uma trajetória tanto profissional quanto pessoal. Neste espaço, gostaria de agradecer às pessoas que foram importantes neste trajeto e que esta página não me deixará esquecer. Agradeço primeiramente aos meus pais, pela educação e pelas oportunidades que me deram. Agradeço também por terem me ensinado a buscar ser um homem bom, justo e honesto.

Logo após iniciar o doutorado, ingressei no Instituto Nacional da Propriedade Industrial (INPI) como examinador de patentes. Por essa razão, o doutorado foi feito em dedicação parcial, nas horas livres. Porém, essas horas não são exatamente livres, mas sim horas que estaria dedicado à minha família, na companhia de minha esposa, Adriana, e das minhas filhas, Luisa e Júlia. Assim, mais do que agradecer, primeiro devo desculpas a elas por eu ter usado um tempo que nunca foi livre.

Fazer o doutorado em dedicação parcial é trabalhar na solidão. É trabalhar nas horas em que ninguém trabalha. É também ter de conviver, às vezes por semanas inteiras, com problemas não resolvidos. Assim, agradeço novamente à Adriana, Luisa e Júlia, pela paciência e compreensão nos momento de indisposição e mau-humor. Agradeço ainda à Neide, minha sogra, pela ajuda em cuidar das meninas nos momentos em que estive ausente.

Agradeço a todos os colegas do INPI que sempre se mostraram disponíveis a qualquer discussão. Em particular, agradeço a Roberto Ferreira, Telma Lucia e Júlio César pelo apoio ao término deste doutorado. Agradeço também ao próprio INPI, e ao Governo Federal, pela concessão de um ano de dedicação parcial da jornada de trabalho, sem a qual não teria sido possível o término desta Tese.

Agradeço ainda ao prof. Sergio, que tem acompanhado minha formação profissional desde o curso de graduação, pela confiança depositada e pelos momentos como orientador, psicólogo e revisor.

Resumo da Tese apresentada à COPPE/UFRJ como parte dos requisitos necessários para a obtenção do grau de Doutor em Ciências (D.Sc.)

DESENVOLVIMENTO DE UM SISTEMA DE CONVERSÃO TEXTO-FALA COM MODELAGEM DE PROSÓDIA

Vagner Luis Latsch

Junho/2011

Orientador: Sergio Lima Netto

Programa: Engenharia Elétrica

Este trabalho aborda o desenvolvimento de um sistema de conversão texto-fala, incluindo as etapas de obtenção da transcrição fonética automática a partir do texto e da modelagem das variáveis prosódicas de duração, *pitch* e intensidade. Em uma etapa de processamento de texto, a transcrição automática é obtida por meio da construção de uma estrutura hierárquica de dados, usada tanto para a síntese da fala quanto ao longo da modelagem das variáveis prosódicas. É proposto um método de parametrizar as variáveis prosódicas no domínio da sílaba, para o qual um *corpus* de fala foi construído para obtenção de dados de análise. Ao longo do estudo da modelagem prosódica, foi dado enfoque ao estudo da prosódia relacionada à expressividade do falante, mais particularmente às atitudes. Primeiramente, foi viabilizado o transplante da prosódia de atitudes para sinais produzidos sem expressividade, através do alinhamento temporal dos sinais. Com a modelagem das variáveis prosódicas, foi possível executar o trasplante na forma paramétrica, entre sentenças de conteúdos diferentes. Para implementação de todas as técnicas aqui discutidas, foi realizado um sistema de desenvolvimento para a conversao texto-fala e manipulação prosódica, denominado SASPRO. Esse sistema inclui as funcionalidades de gravação e edição do sinais, processamento do texto, manipulação prosódica, transplante de prosódia e conversão texto-fala, descritas ao longo desta tese. A descrição completa desse sistema é apresentada no Capítulo 6 desta tese.

Abstract of Thesis presented to COPPE/UFRJ as a partial fulfillment of the requirements for the degree of Doctor of Science (D.Sc.)

DEVELOPMENT OF A TEXT-TO-SPEECH SYSTEMS WITH PROSODY MODELING

Vagner Luis Latsch

June/2011

Advisor: Sergio Lima Netto

Department: Electrical Engineering

This thesis describes the development of a text-to-speech system, including the automatic phonetic transcription from the text and the modeling of prosody variables: duration, pitch and intensity. The automatic transcription is achieved in a text processing module by using a hierarchical data structure suitable for both stages of speech synthesis and prosody modeling. A syllable-domain coding method is proposed to the prosodic variables in which a speech corpus was constructed for data analysis. The expressiveness of the speaker, more particularly speech acts, was the focus of the prosody modeling. The prosody transplantation from speech acts onto signals with neutral acts has been achived by temporal alignment of signals in a first step. After prosodic modeling the transplantation could be done in the parametric form and between sentences with different contents. A complete platform was developed integrating all speech processing techniques discussed previously. The so-called SASPRO platform includes capabilities such as signal recording and edition, text processing, prosodic manipulation, prosodic transplantation and text-to-speech, presented throughtout the thesis. A complete description of this platform is presented in Chapter 6 of this thesis.

Sumário

Capítulo 1

Introdução

Os sistemas de conversão texto-fala (TTS - *text-to-speech*) e de reconhecimento da fala (ASR - *automatic speech recognition*), logo que começaram a apresentar maturidade nos laboratótios de pesquisa, levaram muitas empresas, tais como IBM e MICROSOFT, a colocarem no mercado diferentes aplicações comercias desses sistemas. No entanto, algumas aplicações que supostamente se apresentavam como promissoras não foram bem aceitas pelos usuários em geral.

Atualmente, as aplicações mais populares dos sistemas TTS são os navegadores por GPS, os sistemas de atendimento eletrônico e os sistemas de acessibilidade para deficientes visuais. Ainda hoje, algumas aplicações, como a leitura de textos literários, ainda encontram resistência por parte do usuário pela falta de naturalidade [1]. RABINER [2] considera que a maior tarefa de um sistema TTS é produzir um sinal de fala com alta inteligibilidade, para que a síntese seja útil como meio de comunicação entre homem e máquina, e a naturalidade seja tão perto da real quanto possível.

Neste contexto, os termos "qualidade", "inteligibilidade" e "naturalidade" são utilizados com os seguintes sentidos:

- Qualidade refere-se à qualidade segmental, ou seja, à capacidade do sistema em reproduzir corretamente o sinal correspondente aos sons dos segmentos que caracterizam a linguagem falada. Com boa qualidade, um ouvinte é hábil em reconhecer facilmente os sons característicos da língua.

- Inteligibilidade refere-se à capacidade de compreensão do conteúdo linguístico do sinal de fala sintetizado, de modo que tenha significado para um falante natural da língua.

- Naturalidade refere-se à capacidade do sistema em produzir a fala sintetizada o mais próximo possível da fala humana, incorporando nuances provenientes da expressividade do falante.

A definição de naturalidade é bastante ampla e subjetiva [1], mas podemos supor que a percepção de naturalidade seja função da riqueza do contorno melódico e do padrão rítmico aplicados ao sinal [3].

A prosódia no estilo de leitura, quando caracterizada pela informação linguística extraída do texto, pode permitir uma naturalidade razoável para um sistema TTS [4]. Ultrapassar esse limite razoável de naturalidade é que se apresenta como um problema atual para sistemas TTS e se tornou o foco desta tese de doutorado.

1.1 Objetivos

O objetivo principal desta tese é apresentar um sistema de apoio à pesquisa e desenvolvimento de um sistema de conversão texto-fala, abordando as etapas essenciais da conversão texto-fala incluindo a modelagem da prosódia.

As etapas da conversão texto-fala são realizadas em um sistema TTS protótipo e para a modelagem das variáveis prosódicas de duração, pitch e intensidade, tomou-se como meta modelar a prosódia de atitudes, inspiradas no artigo de MORAES [5].

Devido à multidisciplinaridade envolvida nesta tese, em alguns tópicos é necessária a incursão em áreas da Linguística que, a princípio, não são comuns com a Engenharia Elétrica. Isto torna os objetivos desta tese desafiadores e motivantes. Além disso, a escassez de dados de análise e ferramentas de acesso livre, com enfoque para o português do Brasil, exigiu que muito do esforço dedicado a esta tese fosse direcionado à construção dos próprios recursos.

Assim, alguns dos assuntos tratados nesta tese são objetos de uma discussão teórica mais rigorosa e mais aprofundada no campo da linguística experimental e teórica, ou no campo do processamento de linguagens naturais. Cabe neste ponto a observação de CAMPBELL [6] quando propõe o estudo das durações: "o procedimento utilizado neste trabalho não segue uma base teórica *per se*, mas simplesmente decisões convenientes, e coerentes na medida do possível, para se atingir um objetivo".

1.2 Considerações iniciais

1.2.1 O sistema protótipo TTS

A arquitetura de um sistema TTS pode ser modelada por uma etapa de processamento do texto que codifica o texto para uma forma de representação intermediária e uma etapa de síntese que codifica esta forma intermediária para o sinal de fala [1].

Em [7], os autores descrevem que a informação processada por um sistema TTS é linguisticamente heterogênea, ou seja, compreendem informações sintáticas, morfológicas, fonológicas, fonéticas, prosódicas e acústicas e portanto, é desejável que o sistema guarde as informações linguísticas de diferentes tipos em uma estrutura de dados formal.

Assim, nesta tese, propõe-se implementar uma etapa de processamento de texto que resulte na transcrição fonética do texto, mas que também organize as informações linguísticas obtidas ao longo do processo em uma estrutura de dados hierárquica que represente a forma comum intermediária.

A etapa de síntese é vista como um interpretador desta estrutura de dados intermediária. Conforme TAYLOR [1], as técnicas de síntese da fala podem ser classificadas em três gerações:

- A primeira geração é baseada em regras e na modelagem do trato vocal, tal como os sintetizadores por formantes ou por predição linear.

- A segunda geração é tipicamente caracterizada pela concatenação de um número reduzido de unidades e por um mecanismo de manipulação prosódica dos segmentos, tal como TD-PSOLA (*time domain - pitch synchronous overlap and add*).

- A terceira geração é baseada em grande quantidade de *corpus*, tais como os sintetizadores por seleção de unidades e os sintetizadores por modelos estocásticos baseados em HMMs (*hidden Markov models*).

Atualmente, a técnica de síntese da fala por seleção automática de unidades tem sido dominante [1] em sistemas comerciais. Utilizando extensos *corpora* de fala, o sistema busca nesse *corpus*, em tempo de execução, por uma unidade que melhor se ajuste ao contexto fonético e prosódico que se deseja sintetizar. Isto permite pouca ou quase nenhuma manipulação prosódica da unidade e alta qualidade segmental [8]. Porém, um fator limitante dessa técnica é que o número de contextos prosódicos e fonéticos que podem ocorrer é extremamente alto [9].

A síntese da prosódia pode ser classificada pelo mesmo ponto de vista, em que as técnicas de primeira e segunda geração utilizam um modelo explícito da prosódia, enquanto os modelos de terceira geração buscam estimar a prosódia por métodos de aprendizagem automática, a partir de uma grande quantidade de *corpus* [1].

Nesta tese, a síntese do sinal é feita pela concatenação de unidades temporais, no qual o inventário de unidades utilizado vem sendo aprimorado [10]. Para a manipulação das variáveis prosódicas do sinal concatenado é usado o algoritmo TD-PSOLA.

1.2.2 A modelagem da prosódia

TAYLOR [1] divide o processo de comunicação falada em dois componentes: o verbal, equivalente a um sistema simbólico de palavras que se organizam para criar sentenças; e o componente prosódico, com a função de ajudar o ouvinte a encontrar a forma verbal a partir da fala e também a expressar emoções primárias, tais como raiva e alegria, e expressões secundárias, tais como atitudes (*speech acts*).

Em [11] a autora considera que existe uma distinção importante entre a expressão de emoções e expressão de atitudes. A mesma autora cita em [12] que "a intonação é um meio essencial para sinalizar como nos sentimos a respeito do que dizemos (atitudes) ou como nos sentimos quando dizemos (emoções)".

Sob um ponto de vista particular, a componente da prosódia que ultrapassa o conteúdo verbal, como o caso das atitudes, carrega por si só uma mensagem. Portanto, o falante codifica esta mensagem por meio das variáveis de duração, pitch e intensidade, e o ouvinte é capaz de identificar e interpretar. Porém, quais são os códigos envolvidos nessa codificação que o falante realiza e que o ouvinte é capaz de decodificar? Esta é uma das motivações que nos levou a enfocar a pesquisa em modelagem prosódica no estudo da prosódia nas atitudes.

Ao longo do estudo da modelagem prosódica, na primeira etapa desta tese, foi viabilizado o transplante de prosódia por alinhamento dos sinais, tendo sido proposto um método de transplante baseado na conjugação do algoritmos DTW (*dynamic timer warping*) e TD-PSOLA.

Em uma segunda etapa, parte-se da hipótese de que as variáveis prosódicas podem ser modeladas em um nível mais abstrato do que o domínio do tempo e vinculadas à estrutura linguística do texto. Para esta modelagem, foi constuído um *corpus* de fala para a obtenção das distribuições estatísticas das variáveis prosódicas. Deste modo, é proposto um procedimento de parametrização das variáveis prosódicas no domínio da sílaba, permitindo o transplante de prosódia na forma paramétrica e o estudo dos padrões prosódicos das atitudes.

1.3 Organização da tese

Esta tese está organizada da seguinte forma:

No Capítulo 2, é apresentada a etapa de processamento do texto, onde serão propostos algoritmos alternativos para a obtenção da transcrição fonética automática, viabilizados por meio da construção de uma estrutura organizada da informação extraída do texto.

No Capítulo 3, é descrito o procedimento de transplante por alinhamento temporal do sinal de fala, onde será proposto um método de transplante que busca otimizar

a associação do método de alinhamento por DTW e o algoritmo TD-PSOLA.

O Capítulo 4 é dedicado ao procedimento adotado para codificação das variáveis prosódicas. Neste capítulo, destaca-se a construção de um *corpus* de análise e adaptações particulares para a codificação das atitudes.

No Capítulo 5, é apresentado o transplante de prosódia na forma paramétrica, onde ressalta-se o estudo dos padrões prosódicos de um conjunto de atitudes, sob o ponto de vista dos parâmetros das variáveis prosódicas codificadas.

No Capítulo 6, são apresentadas as ferramentas de *software* implementadas neste trabalho.

Por fim, este trabalho é concluido no Capítulo 7 onde serão apresentadas as conclusões e propostas de trabalhos futuros.

Nos Apêndices A, B, C e D, são apresentados exemplos, regras, tabelas e diagramas citados ao longo desta tese.

Capítulo 2

Processamento do texto

2.1 Introdução

Neste trabalho, o processamento do texto foi implementado como um meio de extrair do texto informações linguísticas suficientes para o estudo da prosódia, em particular para o estudo da prosódia das atitudes.

O objetivo final do processamento do texto é obter a transcrição automática do texto. Além disso, um objetivo não menos importante é organizar as informações linguísticas extraídas do texto em um estrutura de dados hierárquica, desde o nível das palavras até o nível da transcrição fonética.

No Capítulo 4, será descrita a construção de um *corpus* de análise para a codificação das variáveis prosódicas, onde a etiquetagem deste *corpus* foi feita adotando-se a transcrição automática como referência. Assim, a transcrição automática representa o elo de ligação entre a estrutura do texto e o sinal de fala equivalente. É desse modo que será possível obter o sincronismo de todos os níveis da estrutura do texto com o sinal de fala, sem a necessidade de etiquetar o sinal em diferentes níveis, ou seja, de indicar as fronteiras de sílabas ou palavras.

Encontram-se na literatura trabalhos anteriores com objetivos similares aos propostos neste capítulo. Porém, devido à dificuldade de acesso a um sistema que pudesse ser usado livremente e a necessidade de adaptação para o português, optou-se implementar todo o procedimento de transcrição automática e de montagem da estrutura de dados.

No âmbito de um sistema TTS, quando a entrada de texto é irrestrita, o processamento do texto geralmente requer uma etapa de normalização, isto é, uma transformação do texto para que contenha somente palavras. É nessa etapa que elementos no texto, tais como abreviaturas, números, siglas e marcações são identificados e transformados em texto por extenso. Essa identificação é considerada geralmente o trabalho mais difícil [1]. Neste trabalho, considera-se *a priori* que o

texto de entrada contém somente palavras e símbolos de pontuação, dispensando portanto a etapa inicial de normalização do texto.

Diferentes autores [13], [14] consideram, em maior ou menor grau, que o agrupamento de palavras está associado à informação sintática da sentença ou, no mínimo, à categoria morfológica ou POS (*parts of speech*) das palavras. Assim, foi agregado um classificador morfológico ou POS *tagger* que será brevemente descrito na Seção 2.2.

Após a classificação morfológica das palavras, a sentença é submetida à transcrição automática, que será descrita na Seção 2.3. Ao longo desta seção, serão propostos algoritmos baseados em regras, para a conversão grafema-fonema, separação silábica e determinação da tonicidade.

Na Seção 2.4, serão apresentados conceitos e ilustrações do procedimento de montagem da estrutura hierárquica ao longo das etapas de processamento do texto.

O resultado parcial de todo o processamento descrito neste capítulo será apresentado na Seção 2.5, seguido das conclusões, na Seção 2.6.

2.2 Classificador morfossintático

A classificação morfológica automática consiste em um processo de atribuição de etiquetas, convencionalmente chamadas *tags*, a cada palavra ou símbolo do texto, chamados *tokens*. Essas etiquetas indicam a categoria morfológica ou POS (*parts of speech*) das palavras.

Classificadores morfológicos, ou POS *taggers*, são objetos de pesquisa no âmbito do processamento da linguagem natural. Nesta tese, não se pretende aprofundar neste tópico, mas sim viabilizar a implementação de uma ferramenta de apoio. Assim, foi feita uma pesquisa a respeito dos classificadores POS considerando três aspectos: disponibilidade, aplicação ao português e simplicidade de uso e adaptação.

Encontram-se na literatura classificadores baseados na aplicação de regras previamente formuladas [15], e outros baseados em aprendizagem automática a partir de um *corpus* de texto previamente etiquetado [16]. Definir o conjunto de *tags* a ser inserido, estabelecer as regras de classificação e executar a etiquetagem manual são tarefas extremamente trabalhosas, que fugiriam ao escopo deste trabalho, e que requerem o conhecimento de um profissional experiente. Por esses motivos, foi feito uma pesquisa por classificadores que já haviam sido aplicados para o português ou que disponibilizassem o *corpus* utilizado para treino. Além disso, buscou-se por classificadores desenvolvidos na linguagem C ou C++, cuja linguagem vem sendo utilizada no *software* desenvolvido.

Conforme os aspectos desejados, destaca-se a tese de mestrado de AIRES [16]

que comparou na época diferentes classificadores, tendo especificado um conjunto de *tags* e etiquetado um *corpus* de 100.000 palavras para o português. Dentre os classificadores avaliados pela autora, um deles, chamado de TBL (*Transformation Based Learning*) desenvolvido em [17], destaca-se pelo fato de ser disponibilizado com o código fonte aberto, na linguagem C, com todo o procedimento de execução e treino devidamente documentados, tornando-o simples de ser usado e adaptado.

2.2.1 Etiquetador TBL

O classificador POS proposto por BRILL [17] é nomeado por ele como um classificador treinável baseado em regras. Esse classificador por vezes é enquadrado como híbrido, por fazer uso tanto de um procedimento baseado em regras quanto um procedimento estocástico. O autor descreve que classificadores estocásticos possuem vantagens sobre classificadores baseados em regras, dispensando a necessidade do trabalho manual de construção das regras e capturando informação que não foram observadas na construção das regras. No entanto, os classificadores estocásticos apresentam a desvantagem de que a informação linguística é capturada somente de forma indireta, em grandes tabelas de probabilidades. Então, BRILL [17] propõe um classificador treinável, porém diferente dos classificadores estocásticos, no qual a informação linguística é codificada diretamente em um conjunto de regras simples, mas com uma eficiência comparável aos classificadores estocásticos.

O classificador foi então treinado com o *corpus* [1] disponível no trabalho de AIRES [16], totalizando cerca de 82.000 palavras. Após o treino, o mesmo *corpus* foi submetido ao classificador resultando em uma taxa de acerto de 92%. Em [16], a autora conclui que os fatores principais que afetam a taxa de acerto é o conjunto de etiquetas adotado e o tamanho reduzido do *corpus*, sendo necessário um *corpus* pelo menos 10 vezes maior para um aumento na taxa. O maior problema do etiquetador TBL, tal como observado pela autora, é o tempo de treino. No treino do *corpus* usado aqui nesta tese foram gastos 2 dias em um processador *Intel quad core i5*.

O conjunto de *tags* retornadas pelo classificador é mostrado na tabela 2.1, onde além das 36 *tags* indicadas, ocorrem ainda *tags* compostas provenientes de contrações e ênclises. Por exemplo, a contração de uma preposição com artigo, no caso de "em o = no", é classificada como PREP+ART.

Após a submissão de uma frase, o classificador retorna a frase original e acrescenta, junto a cada palavra, as *tags* correspondentes à classificação. Na figura 2.1 é mostrado um exemplo de classificação da frase "O sinal emitido é captado por receptores".

[1] *corpus* obtido de http://www.nilc.icmc.usp.br/nilc/tools/nilctaggers.html

Tabela 2.1: Conjunto de *tags* consideradas pelo classificador [16].

Classe morfológica	*tag*
Adjetivo	ADJ
Advérbio	ADV
Artigo	ART
Número Cardinal	NC
Número Ordinal	ORD
Outros Números	NO
Substantivo Comum	N
Nome Próprio	NP
Conj. Coordenativa	CONJCOORD
Conj. Subordinativa	CONJSUB
Pronome Demonstrativo	PD
Pronome Indefinido	PIND
Pronome Oblíquo Átono	PPOA
Pronome Pessoal Caso Reto	PPR
Pronome Possessivo	PPS
Pronome Relativo	PR
Pronome Oblíquo Tônico	PPOT
Pronome Interrogativo	PINT
Pronome Apassivador	PAPASS
Pronome de Realce	PREAL
Pronome Tratamento	PTRA
Preposição	PREP
Verbo Auxiliar	VAUX
Verbo de Ligação	VLIG
Verbo Intransitivo	VINT
Verbo Transitivo Direto	VTD
Verbo Transitivo Indireto	VTI
Verbo Bitransitivo	VBI
Interjeição	I
Locução Adverbial	LADV
Locução Conjuncional	LCONJ
Locução Prepositiva	LPREP
Locução Pronominal	LP
Palavra Denotativa	PDEN
Locução Denotativa	LDEN
Palavras ou Símbolos Residuais	RES
Pontuação	PON

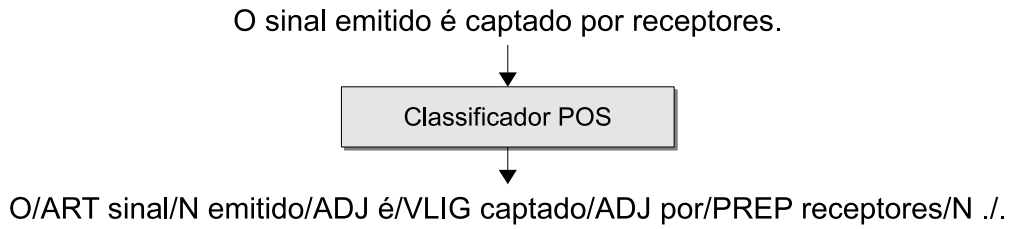


Figura 2.1: Exemplo do processo de classificação de uma frase usando o classificador TBL [17].

2.3 Transcrição automática

A transcrição automática a partir do texto pode ser vista como um processo de estimação, *a priori*, da transcrição dos sons produzidos pela leitura do texto. BARRY e FOURCIN [18] (apud [19]) formalizam esse processo de transcrição em diferentes níveis, dos quais podem-se citar:

- Nível fonêmico: equivale a uma sequência de fonemas obtida da forma ortográfica por meio de regras de conversão ou dicionários léxicos. Nesse nível, **não** são considerados fatores que ocorrem em função de fala espontânea.

- Transcrição fonética ampla: nesse nível são considerados processos que inserem, apagam, ou transformam um fonema em outro, e que podem ser determinados a partir de regras fonológicas aplicadas no nível fonêmico. Nesse estágio, os símbolos usados ainda possuem o *status* de fonemas, pois alofones não são representados.

- Transcrição fonética restrita: é o primeiro nível no qual é representado o que o falante realmente disse no tempo da gravação com base na reprodução do sinal e na inspeção visual do espectrograma. Nesse nível diferentes alofones são representados. Um segmento nesse nível pode conter mais do que um segmento no nível acústico-fonético.

- Acústico-fonético: nesse nível distingue-se cada porção que é reconhecida como um segmento, distinto a partir da forma de onda e do espectrograma. Por exemplo, as oclusivas quando identificadas por duas fases: uma fase de oclusão e a liberação da oclusão.

Geralmente a transcrição é realizada no nível fonêmico ou no fonético, dificilmente chegando nos níveis de detalhe de uma transcrição acústico-fonética [19]. Neste trabalho, a transcrição automática busca estimar o nível de transcrição fonética ampla.

A transcrição automática a partir do texto é reconhecida como um processo de conversão das letras do texto, ou "grafemas", para uma representação de fonemas ou

fones, sendo nomeada geralmente conversão grafema-fonema ou grafema-fone (G2P - *grapheme to phoneme/phone*) ou ainda *letter-to-sound*. Além de obter a sequência de fones, a transcrição automática também é responsável por indicar a separação silábica e a tonicidade da sílabas.

O português é considerado por vários autores como uma língua cuja conversão por meio de regras é bem sucedida [20]. A formulação do inventário de regras para o português, assim como o procedimento de aplicação destas regras, são apresentadas por diferentes autores. Para o português europeu (PE), por exemplo, em [21] [22], os autores descrevem a transformação de regras obtidas manualmente para transdutores de estados finitos. Ainda para o PE, em prosseguimento do projeto ANTÍGONA, iniciado na Universidade do Porto, TEIXEIRA [20] e BRAGA [23] apresentam regras para a conversão grafema-fone, separação silábica e acentuação das vogais. Posteriormente, em [24], são publicadas regras de conversão para o português brasileiro (PB), formuladas por pesquisadores brasileiros com a participação de BRAGA [24]. Para o PB, encontram-se ainda diferentes trabalhos resultantes da associação de linguístas a profissionais das áreas de ciência de computação e engenharia, tais como [25] [26] [27] [28].

As regras de conversão são formuladas normalmente a partir da observação. Porém uma regra que a princípio possa parecer simples teoricamente, nem sempre é simples de ser implementada tendo disponível somente a cadeia de caracteres (ou grafemas). TEIXEIRA [20] relata a dificuldade em transcrever regras formuladas manualmente para um formato capaz de ser aplicado computacionalmente. É comum encontrar regras intuitivas que dependem de informações que não estão presentes diretamente na cadeia de caracteres, como por exemplo, a tonicidade ou limites de uma sílaba.

Em [24] e [20], a separação silábica e a acentuação das vogais são feitas a partir da sequência de caracteres e somente após são aplicadas as regras de conversão grafema-fonema. Os autores observam a dificuldade na separação silábica correta dos ditongos [20].

Nesta tese, é proposto um procedimento diferente para obtenção da transcrição automática, baseado nos seguintes conceitos:

- A conversão grafema-fonema é feita considerando somente a informação disponível na cadeia de caracteres da palavra. Neste caso, as regras que fazem referência à tonicidade ou à formação da sílaba não são consideradas. A partir da conversão grafema-fonema da palavra, todas as etapas seguintes fazem referência aos fones e não mais à cadeia de caracteres;

- A separação silábica é feita após a conversão grafema-fonema das palavras, como em [29], baseado no fato de que a separação silábica é mais natural

quando relacionada ao agrupamento dos fones do que de caracteres;

- As regras de acentuação das vogais são incluídas na conversão grafema-fonema, em um modo de acentuação preliminar. Após a separação silábica, a tonicidade das sílabas e a acentuação das vogais são determinadas definitivamente a partir da acentuação preliminar;

- Somente após a separação silábica e a acentuação das sílabas são aplicadas regras que fazem referência à tonicidade ou à sílaba, nomeadas de regras pós-silábicas;

- Após a transcrição das palavras são aplicadas regras de junção das palavras, nomeadas de regras pós-lexicais.

Nas subseções a seguir serão apresentadas as etapas de obtenção da transcrição automática conforme os conceitos descritos acima.

2.3.1 Conversão grafema-fonema

As regras de conversão grafema-fonema operam no domínio da palavra, somente na sequência de caracteres da palavra, quando todos os caracteres são convertidos para a representação fonética, mesmo que seja uma representação temporária a ser revista em etapas posteriores.

Um dicionário de exceções é usado para cobrir os casos de conversão grafema-fonema que são exceções às regras.

A representação fonética utilizada é formada por uma sequência de fones, identificados não somente por um símbolo, mas por uma estrutura contendo atributos próprios e pré-determinados. Estes atributos consistem em: um código, um símbolo, a classificação quanto à sonoridade e a informação sobre sonoro/surdo, conforme mostrado na tabela 2.2. Os símbolos considerados, seguem os caracteres em SAMPA[2] devido à simplicidade de representação numérica deste alfabeto em código ASCII de 7 bits.

Os fones descritos na tabela 2.2 seguem as ocorrências que foram possíveis de identificar, conforme [30], para o dialeto do autor, natural da cidade de Petrópolis/RJ, com dialeto típico do Carioca.

As regras de conversão grafema-fonema deste nível foram na maioria formuladas a partir de [24], [27] e [30]. Conforme a proposta de partição das regras em diferentes etapas, foram consideradas neste nível somente as regras que se aplicam nos limites da palavra e que fazem referência somente aos caracteres anteriores e/ou posteriores.

[2]http://pt.wikipedia.org/wiki/SAMPA

Tabela 2.2: Tabela de representação dos fones usados neste trabalho.

código	símbolo	sonoro	sonoridade	exemplo	
10	p	0	oclusiva	pato	[patU]
11	b	1	oclusiva	boca	[bok@]
12	t	0	oclusiva	teto	[tEtU]
13	d	1	oclusiva	dado	[dadU]
14	k	0	oclusiva	casa	[kaz@]
15	g	1	oclusiva	gato	[gatU]
16	tS	0	africada	tia	[tSi@]
17	dZ	1	africada	dia	[dZi@]
18	f	0	fricativa	faca	[fak@]
19	v	1	fricativa	vaca	[vak@]
20	s	0	fricativa	sapo	[sapU]
21	z	1	fricativa	zebra	[zebr@]
22	S	0	fricativa	chá	[Sa]
23	Z	1	fricativa	gente	[Ze tI]
24	h	0	fricativa	arpa	[ahp@]
25	H	1	fricativa	carga	[kaHg@]
26	m	1	nasal	mato	[matU]
27	n	1	nasal	nada	[nad@]
28	J	1	nasal	linha	[liJ@]
29	r	1	liquida	caro, prato	[karU], [pratU]
30	l	1	liquida	bala, planta	[bal@],[pla t@]
31	L	1	liquida	olho	[oLU]
32	i	1	vogal_alta	subida	[subid@]
33	i	1	vogal_alta	sinto	[si tU]
34	e	1	vogal_media_alta	sapê	[sape]
35	e	1	vogal_media_alta	quente	[ke tSI]
36	E	1	vogal_media_baixa	vela	[vEl@]
37	a	1	vogal_baixa	pá	[pa]
38	a	1	vogal_baixa	campo	[ka pU]
39	O	1	vogal_media_baixa	pó	[pO]
40	o	1	vogal_media_alta	fulô	[fulo]
41	o	1	vogal_media_alta	conta	[ko t@]
42	u	1	vogal_alta	cajú	[kaZu]
43	u	1	vogal_alta	mundo	[mu dU]
44	@	1	vogal_media_baixa	fina	[fin@]
45	I	1	vogal_alta	sobe	[sObI]
46	U	1	vogal_alta	como	[komU]
47	w	1	glide	pauta salta	[pawt@],[sawt@]
48	w	1	glide	mão	[ma w]
49	j	1	glide	caixa	[kajS@]
50	j	1	glide	tem	[te j]
51	_	0	silêncio		

As regras que fazem referência à palavra posterior, ou que se referem à informação de tonicidade, não foram consideradas nesta etapa.

Além das regras que são típicas, tais como as regras de conversão das consoantes, foi observado que as regras de acentuação das vogais, propostas em [24], podem ser incluídas nas regras de conversão. Além disso, as regras de conversão dos encontros vocálicos que formam ditongos e hiatos foram modificadas. Esta modificação tem por objetivo adequar a transcrição ao novo acordo ortográfico e ainda incluir nas regras de conversão a acentuação preliminar de palavras oxítonas.

A seguir serão apresentados alguns conceitos que balizaram estas modificações e ao fim da seção será apresentado o procedimento proposto para padronização e aplicação das regras de conversão.

Encontros vocálicos (vogal + glide)

Em [31] é citado que os verdadeiros ditongos são os decrescentes, (vogal + glide) e que os crescentes (glide + vogal) variam como hiato. Por exemplo, o encontro vocálico na palavra "iate" pode apresentar um ditongo crescente [ja.tSI] ou um hiato [i.a.tSI]. Outro ponto, segundo a autora, é que os ditongos decrescentes se formam no léxico, enquanto os crescentes se formam pós-léxico. Por estas razões, foi adotado o procedimento de determinar nesta etapa somente os ditongos decrescentes, enquanto que os ditongos crescentes são tratados como hiatos.

No meio de palavras, os encontros vocálicos terminados em "i" e "u" ("ai", "ei", "oi", "ui", "au", "eu" , "ou" e "iu") são sempre convertidos como ditongos decrescentes. Neste caso, as vogais "i" e "u" são convertidas respectivamente para os glides [j] e [w]. Por exemplo, as palavras "gai.tis.ta" [gaj.tSiS.t@] e "dei.ta.do" [dej.ta.dU]. As vogais "í" e "ú" quando acentuadas graficamente, configuram um hiato e são transcritas por vogais. Por exemplo, se não fosse colocado o acento gráfico na palavra "moída" [mo.'i.da], a palavra seria transcrita por [moj.d@].

Existem dois casos de exceção em que estes encontros vocálicos são tratados sempre como hiatos: quando as vogais "i" e "u" são seguidas por consoantes nasais ou por "ir". Por exemplo, as palavra "rainha" [ra.i.nh@], "concluir" [kõ.klu.ir] e "sairmos" [sa.iH.mUs].

No caso destes encontros vocálicos em posição final, a acentuação da vogal foi feita preliminarmente, de modo a caracterizar previamente as oxítonas. Conforme as regras de acentuação gráfica, as palavras paroxítonas terminadas em ditongo são acentuadas. Portanto, se uma palavra termina em um ditongo mas não possui acento gráfico, então estas palavras serão naturalmente oxítonas e o acento recairá sobre a vogal. Assim, quando ocorre um ditongo decrescente no fim da palavra, ou seja, uma vogal seguida de "i(s)" ou "u(s)", as vogais recebem acento preliminar e convertem o "i" ou "u" para glides. Por exemplo, as palavras "sumiu" [su.m'iw] e

"bebeu" [be.b'ew].

Os ditongos nasais que ocorrem em final de palavra são convertidos por uma vogal nasal e um glide nasal. Por exemplo, as palavras "alemães" [a.le.m'ãj̃S], "também" [tã.b'ẽj̃], "balões" [ba.l'õj̃S], "cantam" [kã.tãw̃] e "pão" [p'ãw̃]. Nota-se para a peculiaridade das palavras "muito" e "ruim", que são convertidas como [mũj̃.tu] e [Hu.ĩ], ao invés de [muj.tU] e [rũj̃].

Encontros vocálicos (vogal + glide + vogal)

Os encontros vocálicos formados por (vogal + glide + vogal), como nas palavras "praia" e "maio", são separados de forma a criar um ditongo e uma vogal sozinha depois. Por exemplo os agrupamentos:

- do glide [j] entre duas vogais: aia, eia, oia, uia, aie, eie, oie, uie, aio, eio, oio, uio, uiu.
- do glide [w] entre vogais.

Para estes encontros vocálicos, o encontro inicial (vogal + glide) utiliza as regras de transcrição dos ditongos decrescentes em posição não final. As palavras terminadas com estes encontros vocálicos são sempre paroxítonas e portanto, as vogais não recebem acentuação preliminar.

Uma exceção a esta regra é o encontro "uiu", onde um ditongo decrescente aparece após a vogal "i". Por exemplo, nas palavras "concluiu" [kõ.klu.'iw] (e não [kõ.kl'uj.u]) e "saiu" [sa.'iw] (e não [s'aj.u]).

Aplicação das regras

A grande quantidade de regras de conversão, assim como as inúmeras verificações de contextos anteriores e posteriores, induz a adotar um procedimento que simplifique tanto o modo de organização das regras, quanto o modo de aplicação destas regras. Assim, buscou-se formatar as regras seguindo a proposta de [28], descrita a seguir.

Para cada grafema g, existe uma regra base de substituição, que irá convertê-lo para o fone f, independente do contexto. Além desta regra base, podem haver outras regras que fazem referência aos grafemas anteriores ga e posteriores gp. Assim, para cada grafema têm-se as regras no seguinte formato:

$$\{g, f, \{ga, gp, f, i\}, \{ga', gp', f', i'\}, ...\},$$

onde i representa o incremento na cadeia de grafemas.

A busca por uma regra é feita sequencialmente por ordem de prioridade. Se são satisfeitas as condições dos contextos anteriores e posteriores, a busca termina e a regra é aplicada. Caso nenhuma regra seja encontrada, a regra base é aplicada.

Para simplificar a busca e a comparação dos contextos anteriores *ga* e posteriores *gp*, alguns símbolos são usados para indicar contextos recorrentes, onde os símbolos a seguir indicam:

- # - limite de palavras;
- ! - qualquer contexto;
- $ - qualquer vogal;
- * - qualquer consoante;
- % - consoantes nasais (m ou n);

Por exemplo, as regras de conversão do grafema *c* é dada por $\{c$,k, $\{!, h$,S, $2\}, ...\}$. Conforme a regra base, o grafema *c* será transcrito como [k] se nenhuma regra for encontrada. Porém, se o contexto anterior for qualquer (!), e o contexto posterior for o grafema *h*, então o fone [S] é emitido e a busca na cadeia de grafemas continua. Entretanto, o incremento 2 (dois) indica que a busca irá recomeçar dois grafemas a seguir, para que seja pulada a conversão do grafema *h*. Como as regras são buscadas em ordem sequencial, uma exceção a uma regra geral é criada inserindo a regra de exceção antes da regra geral.

Com esse procedimento, o algoritmo de busca e aplicação das regras resulta em poucas linhas. Além disso, a inserção de novas regras ou exceções é feita somente por alteração da tabela de regras, sem alteração no algoritmo de busca.

A lista completa das regras usadas na conversão grafema-fonema estão mostrada no Apêndice A, na tabela A.1.

2.3.2 Separação silábica

A separação silábica é geralmente aplicada no nível dos grafemas, antes da conversão grafemas-fonema, tal como em [24] e [20]. Porém, optou-se por aplicar a separação da sílaba após conversão grafema-fonema, baseada na composição da sílaba [31].

O modelo para a formação da sílaba consiste em um *núcleo*, que é sempre ocupado por uma vogal, os componentes adjacentes à esquerda do *núcleo*, que ocupam o *onset*, e os componentes à direita do *núcleo*, que ocupam a *coda*. O *onset* e a *coda* da sílaba podem ser vazios, porém sempre haverá uma vogal ocupando o *núcleo* [31].

Em [29] os autores descrevem um procedimento de separação da sílaba a partir da sequência de fonemas, baseado na formação da sílaba. Os autores definem a sílaba pela formação **C1(C2 ou W)V(C3 ou W)C4**, onde **V** indica a vogal, ocupando o *núcleo* da sílaba, **Cn** indica grupos de consoantes e **W** glides, que podem ocupar as posições de *coda* e *onset* da sílaba.

Neste trabalho, o modelo adotado para a formação da sílaba é tal como em [29]. Porém, um conceito diferente foi usado no algoritmo para a separação das sílabas, baseado em [32]. O algoritmo consiste em encontrar um ponto de separação entre duas vogais, de maneira que os segmentos à esquerda da vogal possam configurar uma *coda* para a sílaba anterior, e os segmentos à direita da vogal possam configurar um *onset* da sílaba seguinte. Assim, o algoritmo de separação percorre da esquerda para a direita o espaço delimitado entre duas vogais, onde o ponto de separação da sílaba é definido quando forem satisfeitas as condições de existência da *coda* e *onset* simultaneamente. A busca da esquerda para a direita faz com que o algoritmo selecione o *onset* máximo permitido [31].

As regras de formação do *onset* e *coda* que constam na literatura geralmente são formuladas usando a constituição das sílabas, para que sílabas inexistentes não sejam produzidas. No entanto, para a simples verificação da possibilidade de formação de *onset* ou *coda* estas regras podem ser simplificadas.

Na tabela 2.3 são apresentadas as regras consideradas para formação da *coda*, e na tabela 2.4 para a formação do *onset*. Dado um ponto de busca na posição n, entre duas vogais, as regras de *coda* restringem-se aos dois fones anteriores em relação a n, indicados por $n - i$, enquanto as regras de *onset* consideram os dois fones posteriores, indicados por $n + i$.

Tabela 2.3: Regras de formação da *coda*.

$n-2$	$n-1$	exemplo
-	-	sa[].po
-	glide	a[j].po
-	consoante	a[p].to, a[dZ].mito
glide	h S Z H	anua[jS].
consoante	h S Z H	a[dZS].tri ge te

Tabela 2.4: Regras de formação do *onset*.

$n+1$	$n+2$	exemplo
-	-	cu.[]eca
consoante	-	so.[p]a, sEk.[s]o
consoante	glide	a.[gw]a
p b t d k f g v	r l	a.[tr]opEla

Como exemplo de aplicação das regras, na tabela 2.5 é demonstrada a separação silábica da palavra “tungstênio”, convertida para [tũgStenio]. O ponto “.” indica a

posição da busca entre as vogais, e as condições de *coda* e *onset* são verificadas, respectivamente, nos fones contidos entre colchetes "[]", antes e após o ponto. Quando as duas condições são verdadeiras, então o ponto de busca determina a separação da sílaba e a busca reinicia no intervalo entre as próximas vogais.

Tabela 2.5: Exemplo do procedimento de separação silábica da palavra "tungstênio".

tũgSt'enio	coda	onset
t**ũ[].[gSt]**enio	verdadeiro	falso
t**ũ[g].[St]**enio	verdadeiro	falso
t**ũ[gS].[t]**enio	**verdadeiro**	**verdadeiro**
tũgS.te**[].[n]**io	**verdadeiro**	**verdadeiro**
tũgS.te.n**i[].[]o**	**verdadeiro**	**verdadeiro**

2.3.3 Determinação da tonicidade

Nas regras de conversão grafema-fonema, as vogais que possuem acento gráfico, assim como as vogais acentuadas preliminarmente nas palavras oxítonas, são marcadas pelo símbolo ('), inserido antes da vogal. Este símbolo indica, a *priori*, que a vogal pode ser tônica.

Após a separação silábica, uma sílaba é marcada temporariamente como tônica se há uma vogal marcada como tônica entre os fones que compõem a sílaba . Entretanto, ocorrerão palavras com mais de uma vogal acentuada, e ainda, palavras que não tiveram nenhuma vogal acentuada preliminarmente. Por exemplo, as palavra "órfão" ['Oh.f'ãw̃] e "peixe" [pej.Se]. Portanto, foi realizado um procedimento de correção da tonicidade das sílabas baseado nas seguintes observações:

a. O acento só pode cair sobre uma das 3 (três) últimas sílabas da palavra;

b. As proparoxítonas sempre são acentuadas graficamente;

c. Quando a palavra tem mais de uma sílaba indicada temporariamente como tônica, a sílaba à esquerda tem prioridade;

d. No português existe a preferência pelas paroxítonas [31].

Assim, o procedimento de determinação da tonicidade das sílabas verifica as três últimas sílabas da palavra, de maneira que a indicação temporaria de tonicidade da sílaba mais à esquerda tem sempre prioridade. Se não há acentuação preliminar, a segunda sílaba tem a preferência pela tonicidade (paroxítona). Como resultado deste procedimento adotou-se atribuir a tonicidade da sílaba conforme a posição em

relação à sílaba tônica, podendo assumir os valores de: pretônica, tônica, postônica medial, postônica final, monossílabo átono e monossílabo tônico.

Nas tabelas 2.6 e 2.7 são mostradas as regras aplicadas a palavras com duas sílabas e com três sílabas ou mais, respectivamente. O símbolo s indica a última sílaba, onde as i-ésimas sílabas anteriores à última sílaba s são indicadas pelo símbolo $s-i$. Na coluna "entrada" estão as combinações possíveis da tonicidade temporária atribuída às sílabas. O número 1 (um) indica a marcação de tônica, 0 (zero) indica que não foi associado acento algum e X indica qualquer valor, 0 ou 1. Na coluna "resultado" estão indicados os valores de tonicidade que são atribuidos definitivamente à sílaba.

Tabela 2.6: Regras de atribuição de tonicidade para palavras de duas sílabas.

entrada		resultado	
s-1	s	s-1	s
1	x	**tônica**	postônica final
0	1	pretônica	**tônica**
0	0	**tônica**	postônica final

Tabela 2.7: Regras de atribuição de tonicidade para palavras de três ou mais sílabas.

entrada			resultado			
s-2	s-1	s	...	s-2	s-1	s
1	x	x	pretônica	**tônica**	postônica medial	postônica final
0	1	x	pretônica	pretônica	**tônica**	postônica final
0	0	1	pretônica	pretônica	pretônica	**tônica**
0	0	0	pretônica	pretônica	**tônica**	postônica final

No caso de palavras com uma única sílaba, ou seja, monossílaba, a determinação se átona ou tônica é um caso particular. Os monossílabos átonos são na maioria tratados como clíticos, ou seja, como elementos que não possuem acentuação própria e se agregam à palavra vizinha. Em [33] a autora considera que as palavras funcionais monossilábicas são essencialmente desacentuadas sendo formas "fracas", apresentando propriedades de sílabas sem acento. Assim, tomou-se como critério determinar como monossílabos átonos aquelas palavras cuja classe gramatical se enquadra em artigos, preposições, contrações de preposições e artigos, conjunções, locuções, pronomes relativos, pronomes oblíquos átonos e verbos de ligação e auxiliares.

2.3.4 Regras pós-silábicas

Nesta etapa, ainda realizada no âmbito da palavra isolada, tem-se acesso à diferentes níveis de informação, desde a categoria gramatical da palavra até a tipologia silábica e a tonicidade. Desse modo, é possível implementar regras de maior complexidade. Por exemplo, os ditongos crescentes, que foram convertidos como hiatos na conversão grafema-fonema, agora quando em posição postônicas, são transformados em ditongos crescentes.

Outro caso que merece atenção nesta etapa é a neutralização das vogais. Nem todos os casos de neutralização citados em [31] foram tomados como regras, devido à dificuldade de generalização. Na tabela 2.8 são mostrados os casos em que foi possível estabelecer regras gerais.

Tabela 2.8: Regras de conversão pós-silábicas.

regra	**exemplo**
A nasalização da vogal tônica quando a sílaba seguinte inicia por uma consoante nasal	cama [k'ã.ma]
A formação dos ditongos crescentes átonos em posição postônica final	férias [f'E.rj@S]
A redução das vogais átonas [o][a][e], em sílabas postônicas finais, para [U][@][I]	caso ['ka.zU]
A redução do [e] para [i], em início de palavra, em sílaba pretônica terminada em [S]	esperto [iS.p'Eh.tU]
A redução do [ẽ] para [ĩ], em sílaba pretônica ou monossílabo átono em início de palavra	enxada [ĩ.S'a.d@]
A redução do [e] para [i], em início de palavra, em sílaba pretônica precedida por [d] e sucedida por [S][z][Z]	destaque [dZiS.t'a.kI] desabafo [dZi.za.b'a.fU]

2.3.5 Regras pós-lexicais

Após a transcrição de cada palavra, foram adotados como regras alguns fenômenos observados na junção entre as palavras.

CAGLIARI [34] descreve que na fala contínua, na juntura entre palavras, aplica-se a restrição de que quando a última sílaba de uma palavra terminar em consoante, a primeira sílaba da palavra seguinte nunca começará por vogal. Neste caso, a consoante final passa a pertencer à palavra seguinte.

Além disso, CAGLIARI [34] descreve que quando ocorrer o encontro de duas consoantes, a primeira em geral concordará em vozeamento com a segunda. Porém, quando as consoantes forem iguais, estas consoantes são reduzidas a uma única con-

soante, que pertencerá à sílaba seguinte. TENANI [35] conclui que estes processos são bloqueados somente quando houver uma pausa entre os segmentos.

TENANI [35] dedica um capítulo da tese ao estudo da ocorrência dos processos envolvidos na junção de palavras, chamados de *sândi externo* (ocorrem externamente à palavra). A autora denomina de "vozeamento da fricativa" o caso em que a fricativa no final da sílaba concorda em vozeamento com a consoante que inicia a sílaba ou palavra seguinte e nomeia ainda de "tapping" o caso em que a fricativa se transforma em [r] (*tap*).

Na tabela 2.9 são mostrados os casos citados acima implementados por regras.

Tabela 2.9: Regras pós-lexicais.

Fricativa + Vogal			
anterior	**substituição**	**exemplo**	
h	r	mar aberto [mah] [abEh.tU]	[ma.ra.bEh.tU]
S	z	mais amor [majS] [a.moh]	[maj.za.moh]
s	z	clips amarelo [klips] [a.ma.rE.lU]	[klip.za.ma.rE.lU]
Fricativa surda + Consoante sonora			
S	Z	luz mortal [luS] [moh.taw]	[luZ.moh.taw]
h	H	ser maior [seh] [maj.Oh]	[seH.maj.Oh]
s	Z	bíceps braquial [bi.seps] [bra.ki.aw]	[bi.sepZ.bra.ki.aw]
Fricativas iguais			
H	apagamento	ter razão [teH.Ha.za w]	[te.Ha.za w]
S	apagamento	mais chá [majS.Sa]	[maj.Sa]

Quando ocorre o encontro de vogais na junção de palavras, em algumas condições o encontro pode ser desfeito por um processo nomeado de *sândi vocálico externo*, podendo ocorrer um ditongo ou apagamento da vogal anterior. Vários autores têm-se dedicado à descrição destes processos, como visto em [31] [36] [34] [35]. Em [31] a autora caracteriza tais processos como uma *ressilabação*, classificando-os em:

- Degeminação: quando duas vogais iguais e adjacentes se fundem, como em "menina amada" [me.n'i.na.m'a.da];
- Elisão: quando o [a] final, seguido de uma vogal diferente, é apagado, como em "menina humilde" [me.n'i.nu.m'iw.dZI); e
- Ditongação: quando as vogais [i] e [u] finais se transformam em glides, como em "pêssego amarelo" [p'e.se.gwa.ma.r'E.lU].

Nos trabalhos citados é consensual que a ocorrência de tais fenômenos dependem da velocidade e ritmo da fala, variando ainda de pessoa para pessoa. Por isso, nem sempre estas ocorrências são passíveis de serem generalizadas por regras, a priori. Ao longo deste trabalho, ponderou-se considerar os processos de *sândi vocálico externos* ainda na trasncrição fonética ou deixá-los como um processo cuja estimação posterior de duração pudesse dar conta. Porém, ao longo da etiquetagem do *corpus* de análise, descrito no capítulo 4, verificou-se que considerar alguns casos como regra resulta em um posicionamento mais coerente das etiquetas.

Em [35], a autora observa que o *sândi vocálico externo* ocorre com mais frequência entre vogais átonas e que a presença de pausas bloqueia o processo. NOGUEIRA [36] cita ainda que a configuração **palavra + monossílabos** favorece o processo ("aceita o noivo" - "aceit[u] noivo"), enquanto que **monossílabos + palavra** desfavorece ("moro na esquina - *"moro n[i]squina").

Portanto, foi tomado como regra aplicar a Elisão do [a] e a Degeminação de vogais iguais somente entre vogais átonas, e excluindo a junção de monossílabos seguido de palavras.

2.4 Estrutura hierárquica

No sistema desenvolvido nesta tese, é proposto organizar as informações linguísticas em uma estrutura de dados hierárquica. Neste caso a organização dos dados foi feita no formato de elementos dispostos em listas encadeadas em diferentes níveis. Os elementos são dotados de atributos que são determinados e/ou consultados ao longo das etapas do processo de transcrição. Maiores detalhes da implementação desta estrutura de dados serão vistas no Capítulo 6.

Após a classificação POS, a sentença deixa de ser tratada como uma sequência de caracteres (*string*) para se tornar uma lista de *tokens*. A sentença, no formato de *string*, é submetida a um *parser* que recorta a sentença por espaços, isolando palavras e pontuações, montando a lista de *tokens*. As *tags* inseridas pelo classificador são identificadas e associadas como um atributo de cada *token*. Na figura 2.2 é ilustrado o resultado desta operação.

Após o *parser*, a lista de tokens é submetida à transcrição automática. Neste caso, cada *token*, quando identificado como uma palavra, é submetido à conversão grafema-fonema. O resultado da conversão grafema-fonema é uma lista de fones, representados por estruturas ao invés da simples representação por símbolos. Na figura 2.3 é ilustrado um exemplo da conversão da palavra "emitido", onde cada bloco representa um elemento de uma lista. A partir desta etapa todas as operações seguintes são feitas no domínio desta estrutura, onde é possível a consulta por quaisquer atributos dos elementos. Nota-se que a lista de fones não pertence dire-

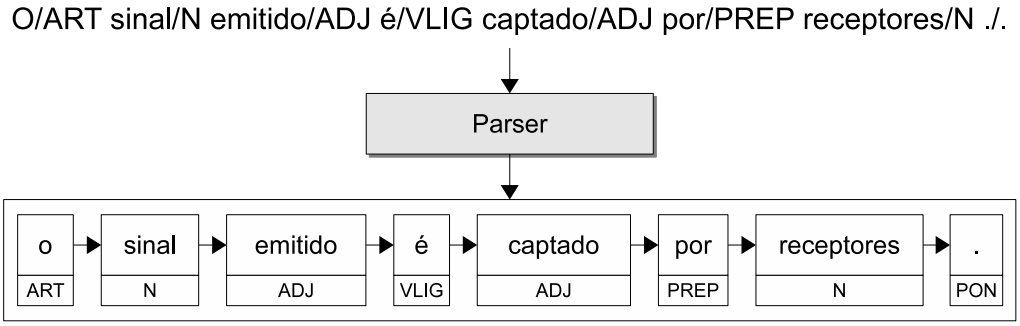


Figura 2.2: Exemplo do processo de classificação e *parser* de uma frase.

tamente à palavra, constando dentro de um elemento intermediário. Este elemento intermediário representa uma sílaba fictícia, única, que será repartida conforme o algoritmo de separação silábica.

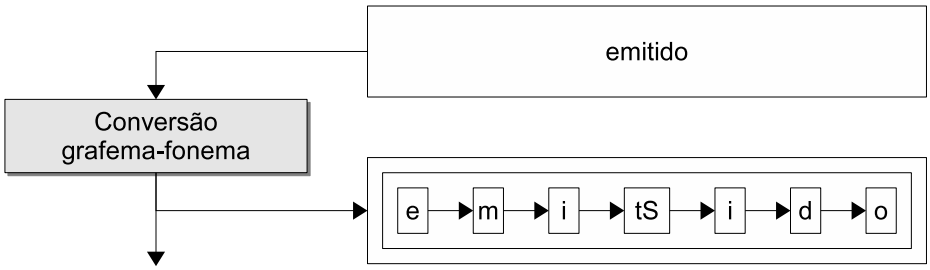


Figura 2.3: Ilustração da aplicação das regras de conversão grafema-fonema sobre a palavra “emitido”, resultando em uma lista encadeada de fones.

Conforme descrito no algoritmo de separação silábica, a busca pelo ponto que define a separação das sílabas é feita no intervalo entre vogais. Quando o primeiro ponto de separação é encontrado, a lista de fones é dividida em duas partes. A primeira parte representa uma sílaba original e a segunda é submetida novamente ao processo de separação. Isso é feito sucessivamente até que não seja mais possível a separação da parte final. Na figura 2.4, é ilustrado um exemplo da aplicação do algoritmo da separação silábica.

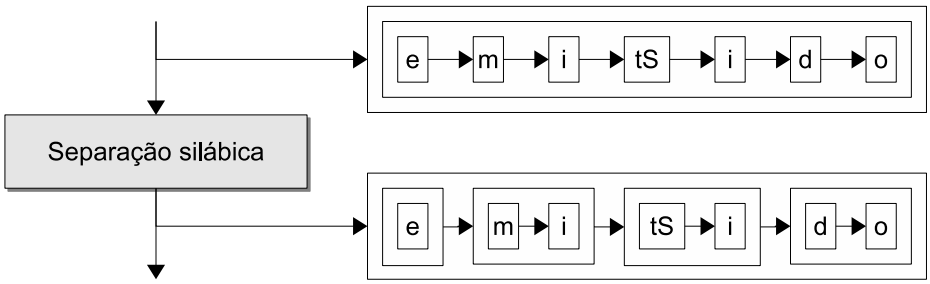


Figura 2.4: Ilustração da separação silábica pela partição da lista de fones.

Com este procedimento, tem-se a palavra como uma lista de sílabas que, por sua vez, é constituida de uma lista de fones.

Um dos atributos da sílaba é a tonicidade, que é atribuída na etapa de determinação de tonicidade. No exemplo da palavra “emitido”, usada até o momento, nenhuma indicação temporária de tonicidade foi atribuída às vogais. Assim, conforme as regras de determinação de tonicidade, a segunda sílaba é marcada como tônica, as anteriores marcadas como pretônicas e a última sílaba é identificada como postônica final. Na figura 2.5 é ilustrado um exemplo de atribuição da tonicidade, onde os diferentes tons de cinza atribuídos aos blocos indicam o atributo de tonicidade das sílabas.

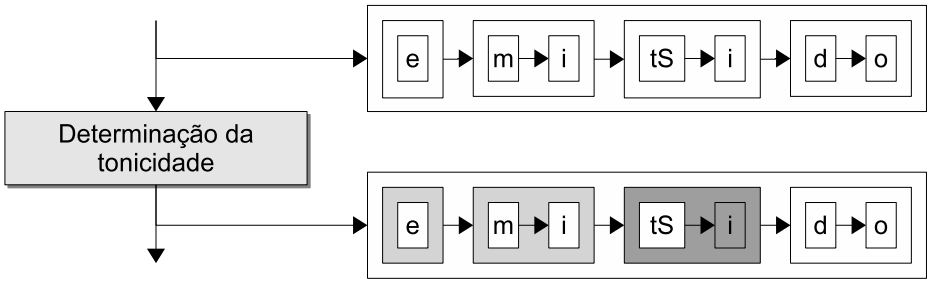


Figura 2.5: Ilustração da atribuição de tonicidade à sílaba. O bloco mais escuro indica a sílaba identificada como tônica, os blocos anteriores de cor mais clara indicam as sílabas pretônicas e o último bloco indica a sílaba postônica final.

Na aplicação das regras pós-silábicas, as regras de redução vocálica realizam somente a substituição de fones por outros. Somente no caso da transformação do ditongo, em postônicas finais, requer a reformulação da sílaba. Na figura 2.6 é ilustrada a redução do [o] pós-tônico final para [U].

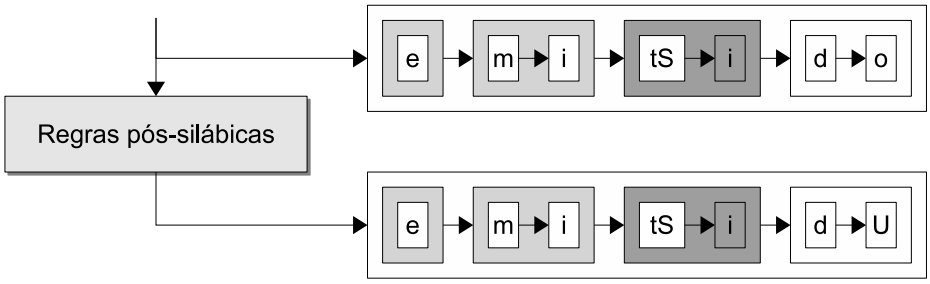


Figura 2.6: Exemplo da redução do [o] postônico final para [U].

Ao fim da conversão nos limites da palavra, as regras pós-lexicais atuam na fronteiras entre as palavras. A aplicação das regras pós-lexicais resulta no apagamento, substituição e deslocamento de fones. Por exemplo, quando o último fone da sílaba final de uma palavra é uma consoante, e a palavra posterior inicia por uma vogal, a consoante é retirada da sílaba à esquerda e movida para a sílaba à direita. Na figura 2.7 é dado um exemplo aplicado à fronteira das palavras “por receptores” [puh He.sep.to.rIS], onde duas regras pós-lexicais são aplicadas sucessivamente. Neste exemplo, a fricativa surda final [h], na sílaba [puh], é substituída pela equivalente

sonora [H], pois o primeiro fone da próxima palavra é uma consoante sonora. Entretanto, a substituição resulta em duas consoantes iguais, e então é aplicada a regra que apaga a consoante anterior.

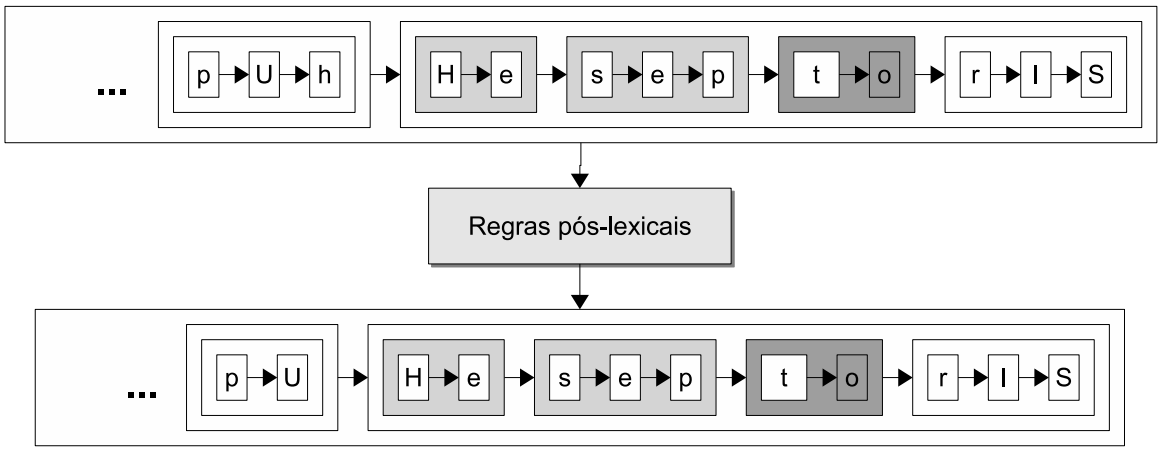


Figura 2.7: Exemplo da aplicação da regra pós-lexical que transforma a fricativa final surda [h], na palavra "por" [puh], em sonora [H].

2.5 Resultados parciais

Para avaliação do desempenho da etapa de processamento do texto, foram obtidas automaticamente a pronúncia das palavras de um léxico contendo aproximadamente 1.000 palavras e a transcrição foi verificada manualmente. As palavras que apresentaram erros de transcrição foram adicionadas a um dicionário de exceções, totalizando cerca de 90 palavras. Considerando a conversão como a compressão do léxico, teríamos uma compressão de 90%. Não foi implementado um esquema de busca eficiente no dicionário, que levasse em conta a similaridade entre palavras, o que poderia reduzir ainda mais o tamanho do dicionário de excessões.

Dentre os erros encontrados, a grande maioria se deve à conversão do [O] e [E], por exemplop nas palavras "posse" [pOsI] e "pede"[pEdZI], tal como encontrado em [24]. Optou-se por converter estes fones somente quando acentuados graficamente, pois estes casos requerem regras pós-silábicas mais aprimoradas, que considerem a classe gramatical das palavras.

Na figura 2.8 é ilustrado o resultado do processamento de texto da sentença "O sinal emitido é captado por receptores.". Nesta ilustração, a estrutura hierárquica é representada na forma textual, em três linhas de texto. Na primeira linha é mostrado o resultado da classificação morfossintática, onde as *tags* POS são inseridas após a palavras. Na segunda linha, é mostrado o resultado do *parser* da sentença e do agrupamento de palavras. Na última linha é representada a transcrição fonética, onde os símbolos que identificam os fonemas aparecem agrupados por sílabas

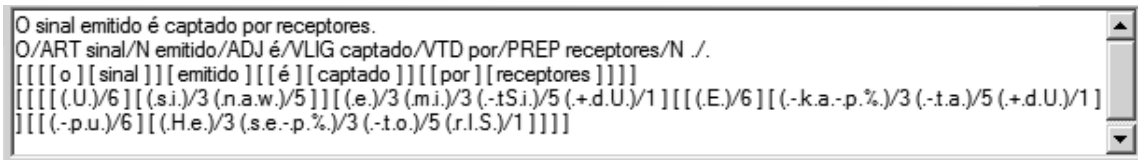


Figura 2.8: Ilustração do resultado do processamento de texto da sentença "O sinal emitido é captado por receptores.".

seguidas da indicação de tonicidade.

2.6 Conclusão

Neste capítulo, foi descrita a etapa de processamento do texto, resultando em uma estrutura hierárquica das informações linguísticas extraídas do texto. Um classificador morfológico foi agregado, tendo sido treinado por um *corpus* de textos em português, previamente etiquetado e conferido manualmente pelos autores do *corpus* [16].

Propôs-se um procedimento de obtenção da transcrição automática particionando as regras de conversão grafema-fonema, apresentadas em trabalhos anteriores, em diferentes etapas. Com este procedimento, as etapas de separação silábica e determinação de tonicidade das sílabas são realizadas no domínio dos fonemas e não mais no domínio dos caracteres, tal como é apresentado em trabalhos anteriores.

Conclui-se que com esta forma de implementação, a obtenção da transcrição automática pode ser vista como um processo de aplicação de regras de transformação em diferentes domínios. Apesar de não haver um rigor teórico nesta proposição, observa-se certa semelhança com a separação dos processos que estão no domínio da fonologia e da fonética.

As regras de conversão grafema-fonema foram ainda padronizadas em um formato específico permitindo a modificação e implementação de maneira simples. Além disso, um dicionário léxico de exceções foi criado para cobrir os casos de falha das regras de conversão. Sugere-se que, futuramente, o tamanho desse dicionário possa ser minimizado por um procedimento de busca por similaridade, tal como descrito em [28].

Nas regras de conversão nomeadas de pós-silábicas, foram implementadas regras básicas que fazem referência à formação da sílaba, tal como a redução das vogais. As diferentes informações disponíveis nesta etapa podem ser usadas futuramente para a formulação de novas regras, mais complexas, tais como as regras de alternância vocálica para a conversão do [O] e [E], que são as maiores fontes de erro de conversão.

Após a conversão de palavras, foram aplicadas regras pós-lexicais de junção de palavras, no qual foi necessário um estudo na literatura específica a respeito dos

processos decorrentes da fala contínua, resultando na proposição de um conjunto de regras julgadas como possíveis de serem generalizas, por observação de um *corpus* de fala transcrito, que será descrito no Capítulo 4.

Por fim, atinge-se o objetivo proposto de realizar um procedimento para extrair do texto informações que possam vir a ser consultadas nas etapas de análise e síntese da prosódia das sentenças.

Capítulo 3

Transplante de prosódia por alinhamento temporal

3.1 Introdução

Neste capítulo será apresentado o procedimento de transplante de prosódia baseado no alinhamento temporal de duas sentenças de mesmo conteúdo. O procedimento consiste em obter a curva de alinhamento, baseado na semelhança das características espectrais dos dois sinais, e em seguida utilizar um método de manipulação prosódica, para copiar ou transplantar as variações prosódicas de um sinal para o outro, geralmente de um sinal de teste para um sinal de referência. Quando o sinal de referência é um sinal sintético, gerado pela concatenação de unidades, esta técnica é útil na verificação da qualidade das unidades de concatenação.

Nesta tese, o transplante de prosódia será usado para copiar padrões prosódicos, típicos da expressividade de atitudes, para um sinal produzido sem expressividade. Neste caso, o transplante de prosódia será realizado pela associação do algoritmo de alinhamento DTW (*dynamic time warping*) com o algoritmo de manipulação prosódica TD-PSOLA [37] [38] (*time domain - pitch syncrhonous overlap and add*).

Desde os anos 80, o alinhamento entre sinais de fala por DTW tem sido proposto também como um meio de realizar a etiquetagem automática de um sinal. Neste caso, a partir de um sinal pré-etiquetado, as etiquetas deste sinal são mapeadas para um sinal que se deseja etiquetar, por meio da curva de alinhamento. O transplante de prosódia utiliza esta mesma curva de alinhamento para determinar as curvas de alteração nas variáveis de duração, pitch e intensidade. Portanto, por usarem a mesma curva de alinhamento, a eficiência do alinhamento entre os sinais para o transplante foi avaliada pela eficiência da etiquetagem automática.

Assim, na Seção 3.2 a seguir, será apresentado o mecanismo de alinhamento temporal dos sinais, no contexto da etiquetagem automática. Em seguida, na Se-

ção 3.3 será apresentado o procedimento a ser executado pelo TD-PSOLA para o transplante da prosódia.

Foi observado que ao usar o TD-PSOLA para a manipulação prosódica, se as marcas de pitch de ambos os sinais estão disponíveis, então o alinhamento pode ser realizado de maneira síncrona com o pitch, permitindo simplificar e otimizar o procedimento de OLA (*overlap and add*). Conforme este conceito, foi proposto um método de transplante denominado de PSDTW-OLA, que será descrito em detalhe na Subseção 3.3.2.

Na Seção 3.4 serão apresentadas avaliações entre diferentes medidas adotadas na execução do alinhamento e do transplante, e também das consequências do método proposto.

3.2 Etiquetagem automática

A proposta de alinhamento entre duas sentenças de mesmo conteúdo, uma de referência e outra de teste, de maneira que as marcas contidas no sinal de referência possam ser mapeadas no sinal teste, foi primeiramente apresentada em [39]. Neste caso, executando o alinhamento entre os sinais por DTW, os autores concluem que é possível obter um alinhamento bem sucedido, mesmo entre diferentes falantes. Posteriormente, outros trabalhos seguindo o mesmo princípio e com diferentes objetivos foram desenvolvidos em [40], [41] e [42]. Em [43], os autores utilizam a etiquetagem automática, baseada em DTW, para mapear as etiquetas presentes em um sinal gerado por síntese para um outro sinal de interesse.

Atualmente a técnica para etiquetagem amplamente usada é baseada em sistemas de reconhecimento contínuo por HMMs (*hidden Markov models*), em modo de alinhamento forçado, tal como em [44]. Entretanto, muitas vezes um sistema de reconhecimento não está disponível para desenvolvedores de sistemas de síntese.

Em [45], os autores comparam sistemas de etiquetagem automática baseados nos modelos HMMs com sistemas baseados no princípio de alinhamento, no qual utilizam como sinal de referência o sinal sintético gerado pelo sintetizador MBROLA (*multiband resynthesis overlap-add*), tal como [40] e [41]. Os autores argumentam que a grande vantagem dos sistemas de alinhamento baseados no DTW se deve à não necessidade de uma etapa de treinamento e apontam para a proposta de utilizar o sistema de alinhamento como *bootstrap* para o treino dos sistemas baseados em HMMs.

Na figura 3.1 está representado um diagrama em blocos do sistema implementado para etiquetagem baseado no alinhamento por DTW. Considere que o sinal de fala que se deseja etiquetar foi previamente gravado e que o conteúdo fonético deste sinal seja conhecido. Este será o sinal de teste $s(n)$, representado no diagrama pelo ramo

da direita. O sinal de referência $r(n)$, representado pelo ramo esquerdo do diagrama, pode ter sido gravado de fala natural e etiquetado previamente ou ser gerado por um sintetizador que forneça as marcas das etiquetas automaticamente.

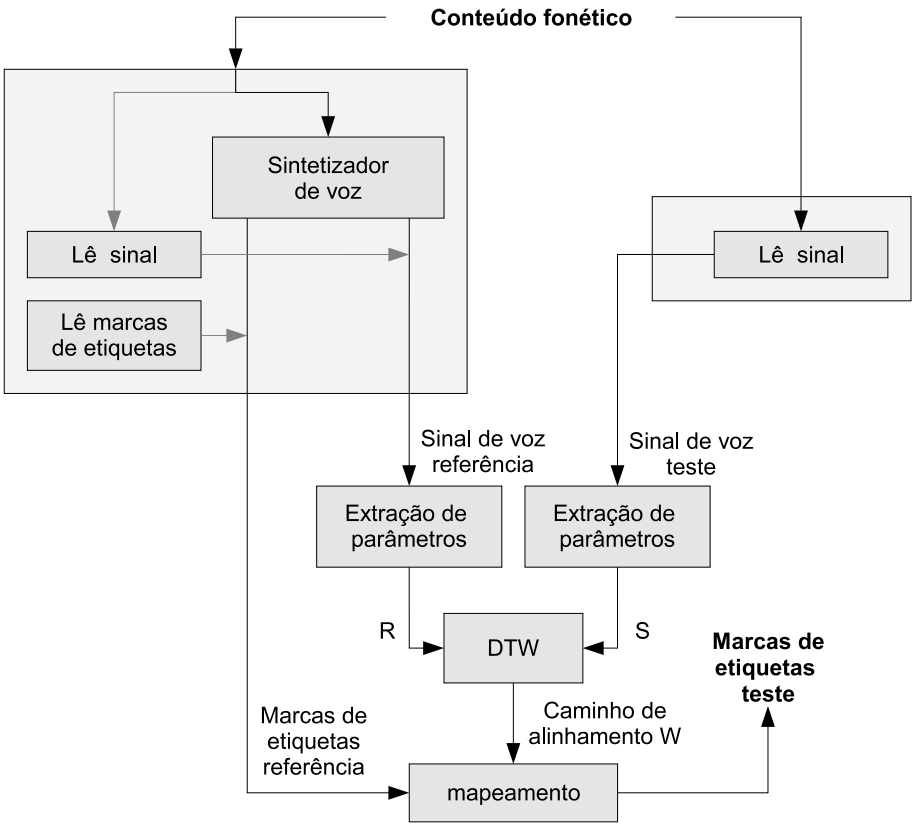


Figura 3.1: Diagrama em blocos de um sistema de etiquetagem automática baseado em DTW. No ramo esquerdo o processamento do sinal de referência e no ramo direito do sinal de teste.

Dados os sinais de referência e teste, estes sinais são submetidos à extração de parâmetros. Nesta etapa, ambos os sinais são segmentados em quadros, usualmente espaçados por um intervalo constante, e de cada quatro é obtido um vetor de parâmetros representativos do conteúdo espectral do quadro. Nesta tese, os quadros têm largura de 20 ms e são obtidos a cada 5 ms, de onde foram extraídos como vetor de parâmetros: 12 coeficientes mel-cepstrais, 12 delta mel-cepstrais, 12 delta-delta cepstrais, totalizando 36 parâmetros [45].

Suponha então que a partir dos sinais de referência e de teste, foram obtidos os quadros de índices j e i, respectivamente, para $j = 1, 2, ..., J$ e $i = 1, 2, ..., I$. Destes quadros foram obtidos os vetores de parâmetros $\mathbf{r}_j$ e $\mathbf{s}_i$ e portanto têm-se as sequências R e S dos vetores representativos do conteúdo espectral do sinais, dadas por $R = [\mathbf{r_1}, \mathbf{r_2}, ..., \mathbf{r_j}, ..., \mathbf{r_J}]$ e $S = [\mathbf{s_1}, \mathbf{s_2}, ..., \mathbf{s_i}, ..., \mathbf{s_I}]$. Estas sequências de vetores de parâmetros são então submetidas ao alinhamento pelo DTW.

3.2.1 Alinhamento temporal

O alinhamento entre as duas seqüencias de vetores R e S relaciona os índices j e i através de um *grid*, onde cada ponto de interseção do *grid* indica a similaridade entre os vetores de parâmetros. Podemos definir um caminho $W = [w_1, w_2, ..., w_K]$ formado por um conjunto de K pontos $w_k = (i_k, j_k)$. Na figura 3.2 é mostrado um exemplo de um caminho de alinhamento formado pelo pontos $W = [(1,1),(2,2),(3,3),(3,4),(4,5),(5,6),(6,7),(7,7),(8,8),(9,8),(10,10),(11,10)]$.

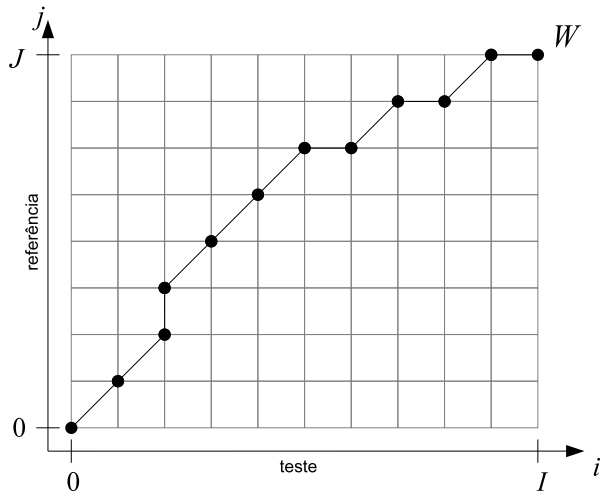


Figura 3.2: Exemplo de um *grid* e caminho de alinhamento W entre os vetores de um sinal de teste e de referência.

Para se determinar o melhor alinhamento são impostas restrições às transições do caminho a ser estabelecido, tal como apresentado em [46]. Neste caso, as transições permitidas são indicadas por uma sequência de movimentos unitários, especificadas por um par de coordenadas (p, q). Assim, um caminho W permitido, contendo K movimentos, será descrito pelas transições $P \rightarrow (p_1, q_1)(p_2, q_2)...(p_k, q_k)...(p_K, q_K)$, e um ponto $w_k = (i_k, j_k)$ do caminho será dado por $i_k = \sum_{n=0}^{k} p_n$ e $j_k = \sum_{n=0}^{k} q_n$. Na figura 3.3 são mostradas algumas restrições ao caminho consideradas nesta tese, das quais, os tipos A, B, C e D foram obtidos de [46] e renomeados aqui por questão de clareza. O tipo E é utilizado em [45] e o tipo F, foi proposto nesta tese.

Observa-se que as restrições impostas ao caminho, implicam em resultados diferentes na execução do alinhamento e também no mapeamento das etiquetas. Na Seção 3.4 serão apresentadas as implicações dos diferentes tipos de restrições no mapeamento das etiquetas.

Na escolha do melhor caminho de alinhamento, seja uma medida de distorção $d(\mathbf{r}_j, \mathbf{s}_i)$ entre os vetores $\mathbf{r}_j$ e $\mathbf{s}_i$, como por exemplo a distância euclidiana. A cada passo no *grid*, desde a origem, esta distorção entre os vetores vai sendo acumulada ao longo do caminho percorrido. Esta distorção acumulada pode ser interpretada como um "custo" para se sair da origem e se chegar até um ponto seguinte. Assim, o

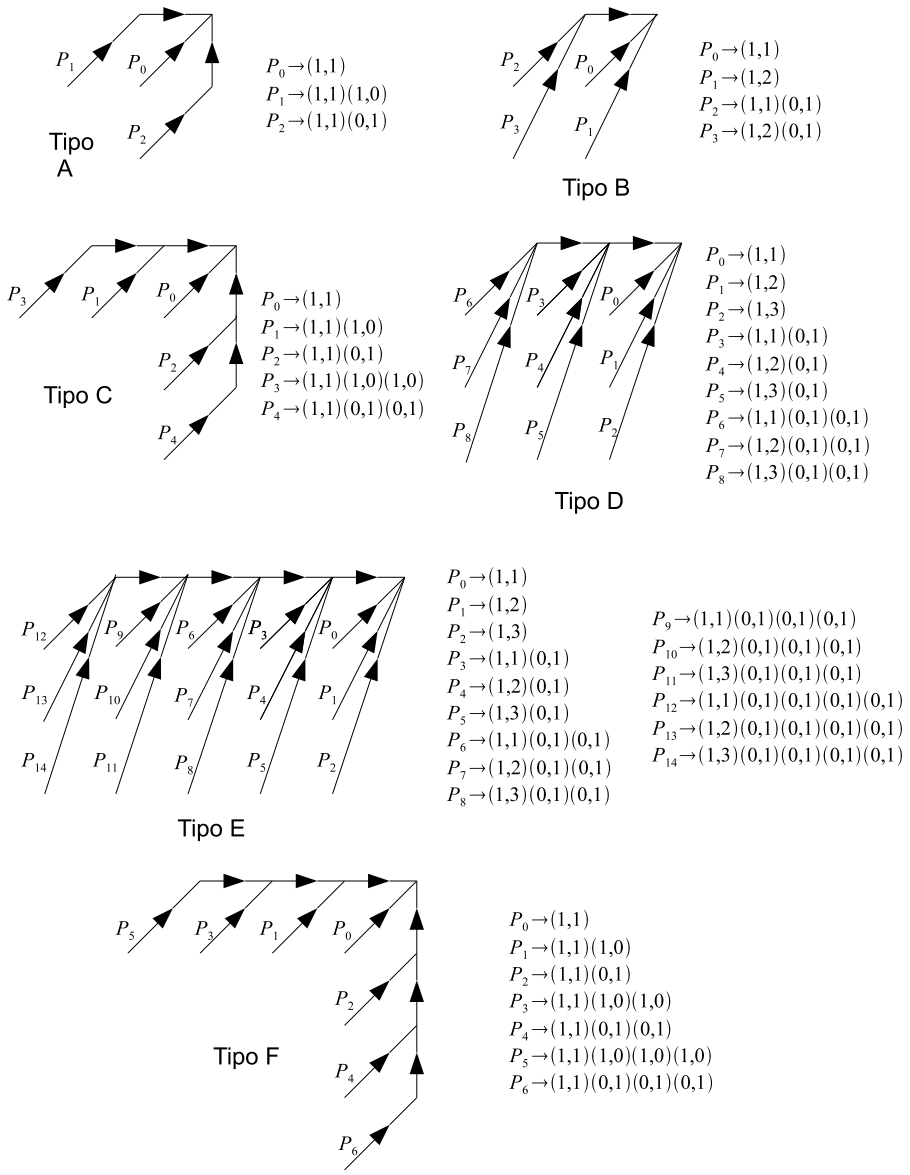


Figura 3.3: Exemplos de restrições impostas ao caminho de alinhamento.

algoritmo de DTW busca determinar, para cada ponto (i, j) no *grid*, o custo mínimo para se chegar a este ponto e por qual transição, conforme a restrição ao caminho adotada. No ponto final (I, J) de intersecção entre as sequências de vetores, tem-se portanto uma medida global de similaridade entre as sequências e retornando-se ao ponto $(0, 0)$, pelas transições que ofereceram menor custo, obtêm-se os pontos w_k do melhor caminho de alinhamento W.

Foi observado que se as fronteiras dos intervalos sonoros e surdos dos sinais estão disponíveis, o alinhamento pode ser executado por partes. Melhor dizendo, as fronteiras entre os intervalos sonoros e surdos impõem ao caminho pontos obrigatórios de passagem. Na figura 3.4 é ilustrado um exemplo, onde as áreas mais claras representam intervalos surdos e a área escura um intervalo sonoro. Então, os pontos (i_A, j_A) e (i_B, j_B) no *grid* definem as fronteiras entre os intervalos, e o alinhamento

é executado em 3 partes: do ponto $(0, 0)$ até (i_A, j_A); do ponto (i_A, j_A) até (i_B, j_B) e do ponto (i_B, j_B) ao fim (I, J). Este procedimento equivale a impor ao caminho de alinhamento pontos obrigatório de passagem, representados na figura por círculos ao redor do ponto de interseção. Este procedimento foi implementado, e na Seção 3.4 será observada a consequência de tal procedimento.

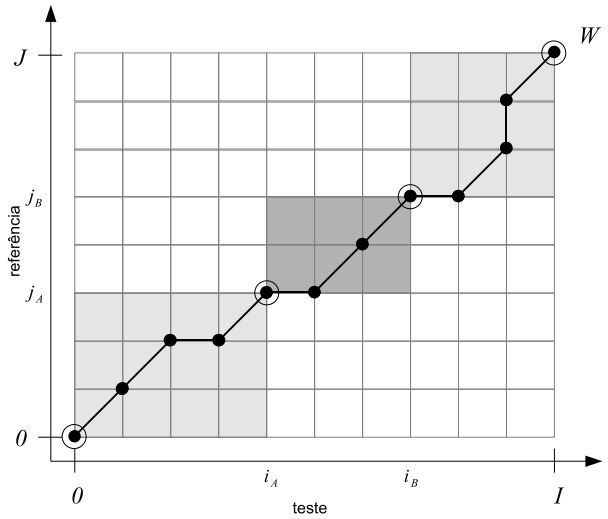


Figura 3.4: Exemplo de alinhamento usando a classificação surda/sonora, indicada pelas áreas claras e escuras, respectivamente. Os círculos ao redor dos pontos de interseção entre as áreas das áreas.

3.2.2 Mapeamento das etiquetas

Após determinada a curva ótima de alinhamento entre as sequências de vetores, então as etiquetas do sinal de referência são mapeadas para o sinal de teste através desta curva.

Considere que o sinal de referência possua L etiquetas inseridas nos instantes de tempo e_l^r, para $l = 0, 1, ..., L$. Estas etiquetas são mapeadas na curva de alinhamento, dando origem igualmente a L etiquetas no sinal de teste, posicionadas nos instantes e_l^s, como ilustrado na figura 3.5.

Os pontos $w_k = (i_k, j_k)$ da curva de alinhamento W relacionam os índices dos quadros i_k e j_k, centrados nos instantes de tempo t_i e t_j, dos sinais de teste e referência, respectivamente. Assim, para o mapeamento, é feita a busca ao longo dos pontos de W, de modo a encontrar o k-ésimo ponto $w_k = (i_k, j_k)$ que satisfaça a condição $t_{j-1} < e_l^r \leq t_j$. Em seguida, a etiqueta de teste e_l^s é obtida pela interpolação linear dos pontos t_{i-1} e t_i.

Na figura 3.6 é demonstrada a aplicação do alinhamento dos sinais da sentença "Renata jogava". O sinal de referência foi produzido sem expressividade, enquanto o sinal de teste foi produzido com uma atitude de aviso, que prolonga acentuadamente

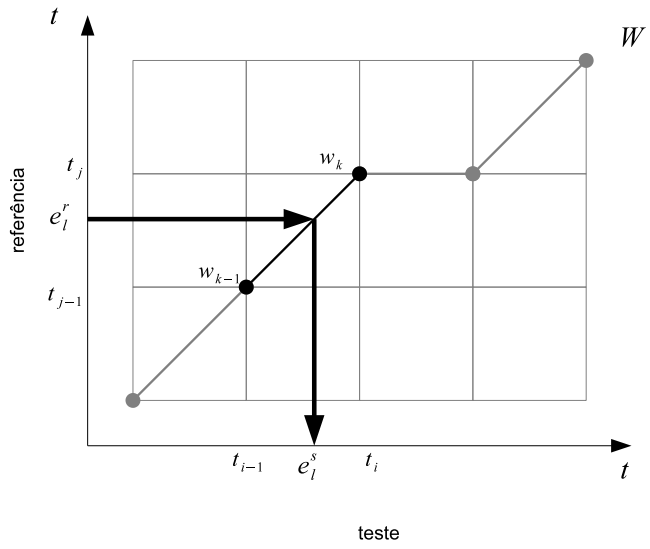


Figura 3.5: Exemplo do mapeamento da etiqueta e_l^r no sinal de referência para a etiqueta e_l^s no sinal de teste, através do caminho de alinhamento W.

a vogal da sílaba "ga", conforme será visto no capítulo 5. A imagem de fundo representa a matriz de similaridade entre os sinais e a linha cruzando a diagonal representa o melhor caminho de alinhamento selecionado. As regiões mais claras representam os pontos de maior similaridade, pelos quais o caminho de alinhamento busca percorrer. As linhas horizontais e verticais que cruzam a figura representam as etiquetas da referência sendo mapeadas para o sinal de teste.

3.3 Transplante de prosódia

O transplante de prosódia entre duas sentenças de mesmo conteúdo é feito pela associação do alinhamento por DTW com um método de manipulação de pitch e da escala temporal. Este procedimento tem tido diferentes aplicações, por exemplo, em sistemas de karaokê [47] e na validação subjetiva da qualidade das unidades de um sistema de concatenação [3], [45]. Em [40] os autores apresentaram uma ferramenta para transplante, chamada MBROLIGN, associando o alinhamento por DTW com o sintetizador MBROLA. Em [47] foi usado o algoritmo WSOLA (*waveform similarity based overlap-add*) para alteração das durações e o algoritmo PSOLA para alteração do pitch. Nesta tese, o alinhamento por DTW foi associado ao PSOLA, tanto para a alteração das durações quanto do pitch, pela simplicidade de implementação do PSOLA.

No transplante de prosódia a etapa de alinhamento usada é a mesma da etiquetagem automática. Adicionalmente, o mapeamento das etiquetas feito na etiquetagem automática também é feito no caso do transplante. Porém, neste caso o mapeamento é feito nas variáveis prosódicas e ainda, no sentido inverso, do sinal de teste para o

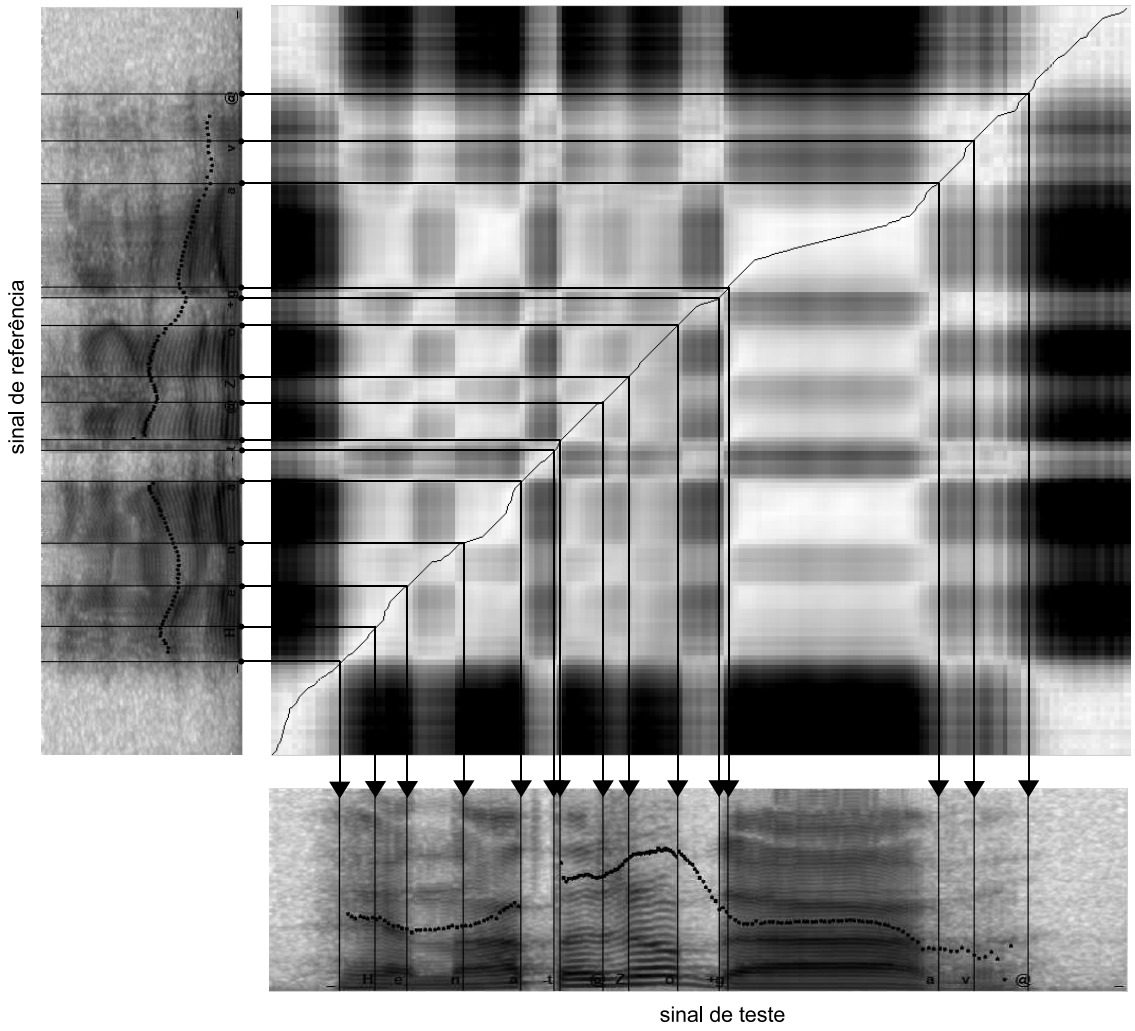


Figura 3.6: Exemplo da etiquetagem por alinhamento. Na vertical o sinal de referência contendo as etiquetas e na horizontal o sinal de teste etiquetado pelo mapeamento das etiquetas da referência para o teste, conforme o caminho de alinhamento.

sinal de referência. Deste modo, a curva de contorno de pitch, assim como a curva de intensidades, do sinal de teste, são mapeadas para serem usadas pelo PSOLA para impor um novo padrão de pitch e de intensidades ao sinal de referência. Ao mesmo tempo, a própria curva de alinhamento fornece ao PSOLA a informação de mudança da escala temporal.

Na seção a seguir será apresentado mais detalhadamente o procedimento de manipulação do pitch e durações executado pelo PSOLA.

3.3.1 PSOLA

O algoritmo PSOLA (*pitch synchronous overlap and add*) [37] é baseado na técnica de *overlap and add* (OLA) e permite reconstruir um sinal periódico com diferente escala temporal e/ou pitch [3] por meio do janelamento do sinal de maneira síncrona com o pitch.

O algoritmo apresenta baixa complexidade computacional e alta qualidade de voz [3]. Porém, por ser síncrono com o pitch, requer a determinação das marcas do pitch no sinal que se deseja modificar. As marcas de pitch são introduzidas, nos trechos sonoros, na posição de um evento específico no ciclo de pitch, idealmente no instante de fechamento glotal (*glotal closure instants* - GCIs). Nos trechos surdos, as marcas são regularmente espaçadas. Na construção de banco de unidades para sistemas de síntese por concatenação de unidades temporais, a detecção precisa dos GCIs tem sido feita por um equipamento eletroglotógrafo [48]. Este equipamento mede a atividade das cordas vocais, de onde os GCIs podem ser obtidos com maior precisão. Um método alternativo de detecção dos GCIs é apresentado em [10] no qual os GCIs são obtidos de um sinal auxiliar captado através de um microfone em contato com o pescoço.

O algoritmo PSOLA pode ser dividido em uma etapa de análise e outra de síntese, onde o sinal de análise, quando modificado, dá origem a um novo sinal, chamado de sinal de síntese. Considere que o sinal de análise possui as marcas de pitch posicionadas nos instantes de tempo p_j^a. Na etapa de análise o janelamento síncrono com o pitch é feito aplicando-se ao sinal de análise janelas de Hanning centradas nestas marcas de pitch p_j^a e com largura limitada pelas marcas vizinhas. Na etapa de síntese, as janelas obtidas do sinal de análise são usadas para compor o sinal de síntese por OLA. Para atender ao contorno de pitch desejado, as janelas são aproximadas ou afastadas, e para atender ao requisito das durações, as janelas são removidas ou replicadas.

A curva de período de pitch, que se deseja impor ao sinal de análise, é tradicionalmente representada por uma função do tempo $P^s(t)$, que fornece a posição das novas marcas de pitch, ou seja, das marcas de pitch do sinal de síntese p_i^s. Se $P^s(t)$ varia lentamente, então, dada uma primeira marca no instante p_i^s, as marcas posteriores são obtidas por:

$$p_{i+1}^s = p_i^s + P^s(p_i^s) \tag{3.1}$$

como lustrado na figura 3.7 para $P^s(t)$ linear e ligeiramente decrescente.

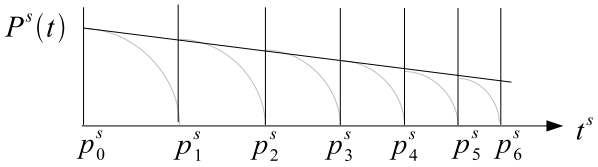


Figura 3.7: Exemplo da obtenção das marcas de pitch de síntese a partir de uma curva de período de pitch.

Na figura 3.8 é ilustrado o procedimento de análise e síntese executado pelo PSOLA [3]. No gráfico inferior, é mostrada a curva de período de pitch $P^s(t)$ que se deseja impor, de onde são obtidas as marcas de pitch de síntese p_i^s, tal como a

figura 3.7. No gráfico superior, no eixo das ordenadas, são mostradas as marcas de pitch do sinal de análise e as janelas a serem aplicadas, onde cada *j-ésima* janela foi representada em uma tonalidade distinta. Assim, para que o requisito de pitch seja atendido no sinal de síntese, cada janela de análise, centrada em uma marca p_j^a, precisa ser associada a uma marca p_i^s do sinal de síntese, conforme a curva W que representa a alteração nas durações que se deseja impor. As marcas de pitch de análise p_j^a são então mapeadas pela curva W para um eixo fictício de síntese no eixo das abscissas, dando origem às marcas $p_j'^a$. Em seguida, cada marca de pitch desejada para síntese p_i^s será associada à janela na marca $p_j'^a$ mais próxima. No exemplo da figura 3.8, as marcas de pitch de síntese de índices $i = 0, 1, 2, 3, 4, 5, 6$ são associadas às janelas de análise de índices $j = 0, 2, 3, 3, 4, 4, 4$. Então, o procedimento de OLA na etapa de síntese utilizará uma janela de análise j, centrada no instante p_i^s, conforme esta associação. Nesta figura, as janelas a serem usadas na síntese no procedimento de OLA estão desenhadas nas mesmas cores que as janelas de análise correspondentes.

Para o caso do transplante, conforme a nomenclatura utilizada até o momento, o sinal de referência que se deseja modificar trata-se do sinal de análise. O contorno de pitch a ser imposto será dado pelo sinal de teste e a alteração das durações será dada pela curva de alinhamento W fornecida pelo DTW. Por esta razão, na figura 3.8 foi utilizado o caminho W, representado sobre os pontos de um *grid*, tal como ilustrado na figura 3.2.

Na seção a seguir será proposto um método de transplante a partir da junção dos algoritmos DTW e PSOLA, de modo que a curva de alinhamento fornecida pelo DTW indicará diretamente a associação das janelas de síntese com as janelas de análise, atendendo simultaneamente os requisitos de pitch e durações e maximizando a similaridade entre os sinais.

3.3.2 Conjugação PSDTW-OLA

Se são conhecidas as marcas de pitch do sinal de referência e do sinal de teste, o alinhamento entre os sinais, quando realizado de maneira síncrona com o pitch, permite otimizar e simplificar elegantemente o procedimento de OLA descrito na seção anterior.

Uma vez que o procedimento de detecção dos GCIs no sinal de teste fornece diretamente o posicionamento das marcas de pitch, então não há necessidade da representação do contorno de pitch: as próprias marcas de pitch do sinal de teste são usadas como as marcas de pitch desejadas para o sinal de síntese. Isto elimina a representação intermediária do contorno de pitch. No entanto, a associação entre as janelas de análise e as marcas de pitch da síntese ainda terá que ser feita mediante

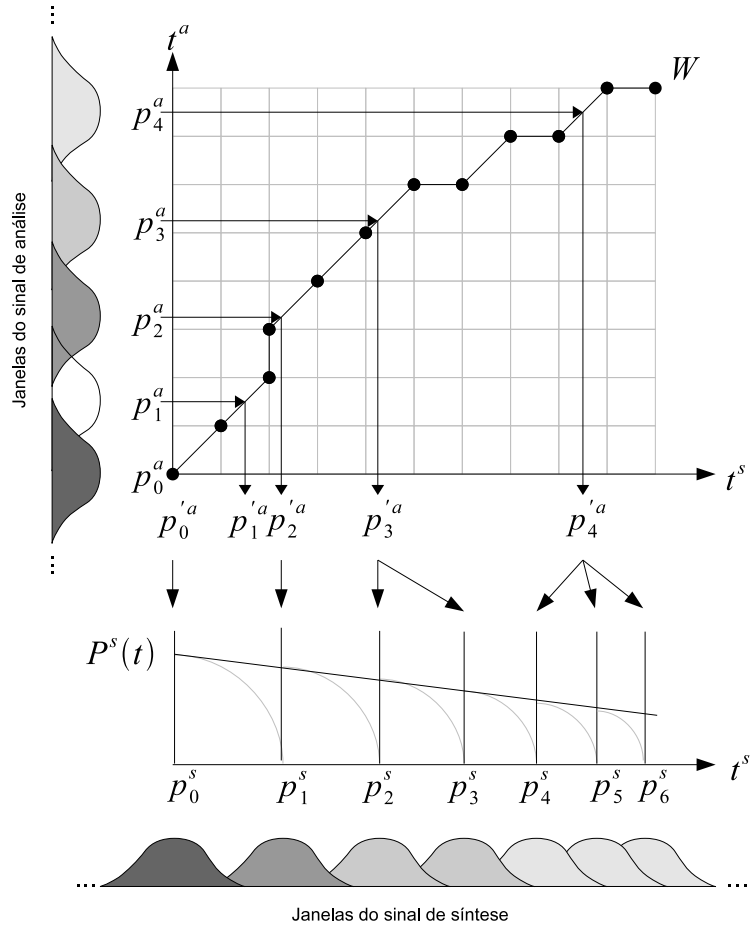


Figura 3.8: Exemplo do relacionamento entre as janelas de análise e as marcas de pitch do sinal de síntese, onde as marcas de pitch das janelas de análise são mapeadas no eixo do sinal de síntese pela curva de alinhamento e por aproximação são determinadas as janelas de análise que serão usadas por OLA para compor o sinal de síntese.

aproximação.

O que se observou é que se os pontos de W forem síncronos com as marcas de pitch de análise e de síntese, ou seja, se os pontos de W pertencerem a um *grid* que mapeia as marcas de pitch de ambos os sinais, então este caminho de alinhamento síncrono com o pitch W_{ps} estabelecerá diretamente o mapeamento das janelas de análise com as marcas de pitch do sinal de síntese. Deste modo, o procedimento do PSOLA poderá ser simplificado para executar o OLA conforme a associação das janelas indicada pelos pontos da curva de alinhamento.

Assim, para se obter o caminho W_{ps} síncrono com o pitch, é necessário que o DTW seja executado de maneira síncrona com o pitch, nomeado de PSDTW (*pitch synchronous dynamic time warping*).

Na figura 3.9 é mostrado um exemplo de manipulação das durações e pitch, similar ao mostrado na figura 3.8. Neste caso, porém, o *grid* representado é formado pelos pontos de interseção das marcas de pitch do sinal de referência

com as marcas de pitch do sinal de teste. Portanto, os pontos $w_k = (i_k, j_k)$ da curva de alinhamento W_{ps} indicam diretamente que a j-ésima janela de análise será utilizada no procedimento de OLA na i-ésima marca de pitch do sinal de teste para compor o sinal de síntese. Neste exemplo, a curva de alinhamento $W_{ps} = [(0,0),(1,1),(1,2),(2,3),(3,3),(4,4),(5,4),(6,4)]$ determina que as janelas de análise com os índices $j = [0,2,3,3,4,4,4]$ sejam posicionadas por OLA, na construção do sinal de síntese, nas marcas de pitch do sinal de teste de índices $i = [0,1,2,3,4,5,6]$.

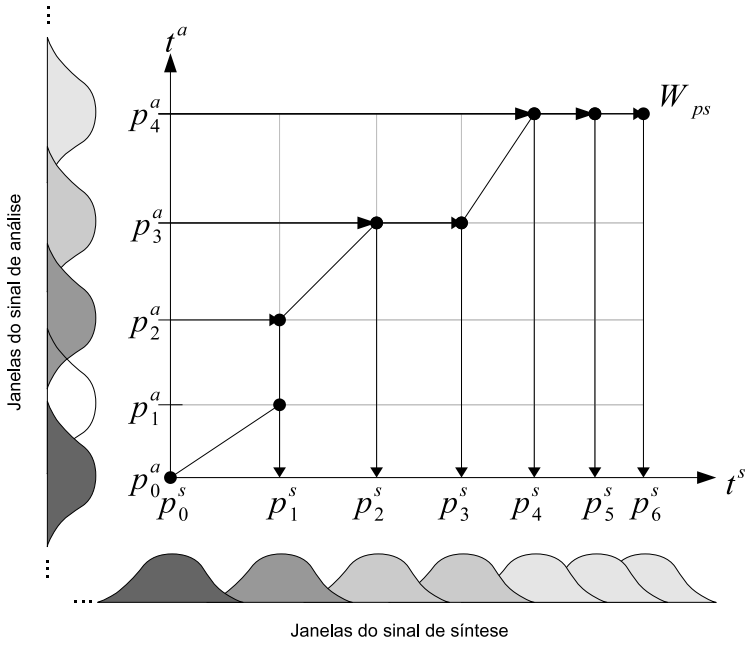


Figura 3.9: Exemplo do caminho síncrono com o pitch $W_{ps} = [(0,0),(1,1),(1,2),(2,3),(3,3),(4,4),(5,4),(6,4)]$ determinando que na construção do sinal de síntese, as janelas de análise com os índices $j = [0,2,3,3,4,4,4]$ serão posicionadas por OLA nas marcas de pitch do sinal de teste de índices $i = [0,1,2,3,4,5,6]$.

A aproximação que era feita na etapa de síntese, para associar uma janela de análise a uma marca de pitch de síntese, é agora transferida ao PSDTW que fará esta aproximação maximizando a similaridade entre os quadros. Além disso, o tipo de restrição imposta ao caminho de alinhamento passa a determinar diretamente um critério para a replicação e supressão de janelas.

Na figura 3.10 está representado o diagrama em blocos do método de transplante tal como proposto. O sinal de referência pode ser um sinal de fala natural ou gerado pelo sintetizador. As marcas de pitch no sinal gravado podem ser obtidas com o uso do eletroglotógrafo ou a partir do sinal alternativo do microfone de contato, conforme descrito em [10]. No caso do sinal gerado pelo sintetizador, as marcas de pitch já estão presentes nas unidades de concatenação.

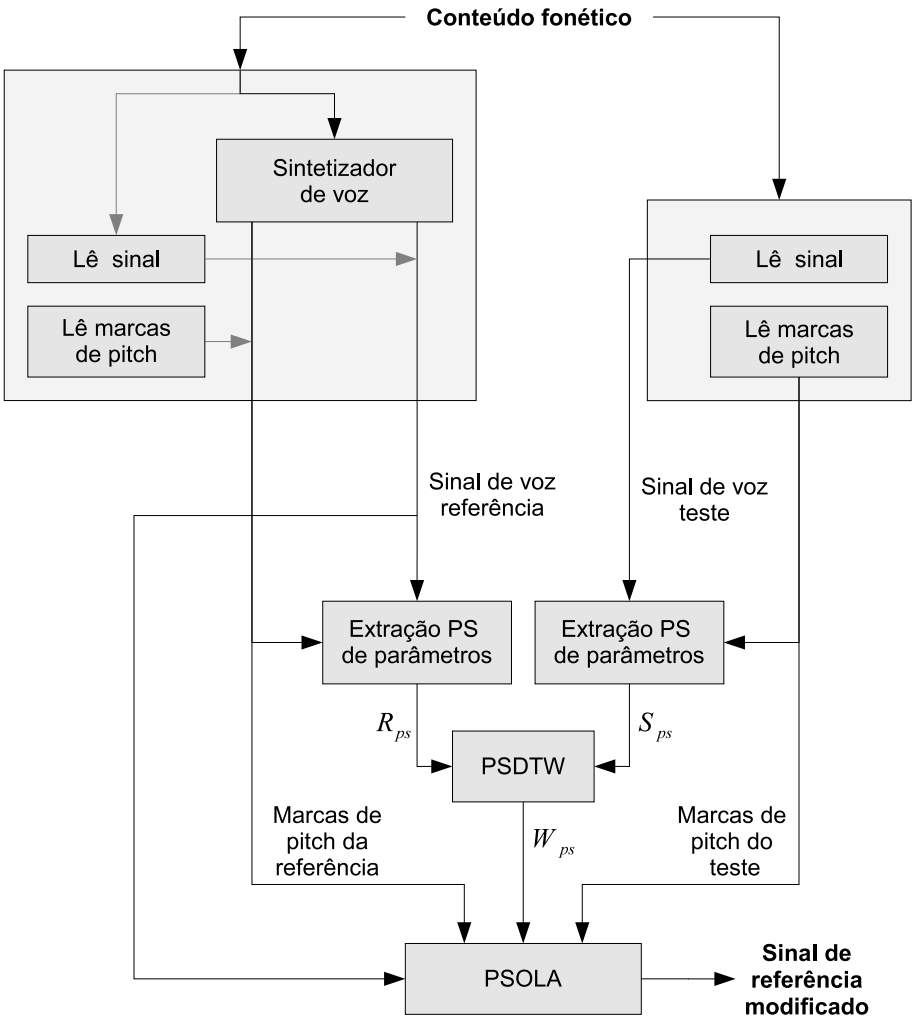


Figura 3.10: Diagrama em blocos de um sistema de transplante síncrono com o pitch.

Para o alinhamento, a extração de parâmetros é executada nos trechos sonoros, de maneira síncrona com o pitch, com quadros centrados nas marcas de pitch, e nos trechos surdos foram obtidos quadros uniformemente espaçados, com sobreposição de 50%, para permitir a reconstrução por OLA. De cada quadro foi extraído o mesmo vetor de parâmetros citados na Seção 3.2, dando origem a uma seqüência de vetores, nomeados por R_{ps} e S_{ps} por serem síncronos com o pitch. Então, o alinhamento entre R_{ps} e S_{ps} pelo PSDTW resulta no caminho de alinhamento W_{ps}.

Após o alinhamento, são repassados ao PSOLA o caminho de alinhamento W_{ps}, as marcas de pitch do sinal de teste e do sinal de referência, assim como o próprio sinal de referência que equivale ao sinal de análise. Na construção do sinal de síntese o PSOLA está limitado a realizar o procedimento de OLA conforme a associação fornecida pelos pontos de W_{ps}, tal como no exemplo da figura 3.9.

3.3.3 Exemplo de transplante

Na figura 3.11 está representado um exemplo de transplante do padrão prosódico da sentença "Renata Jogava" quando gravada com um atitude de aviso, para a mesma sentença gravada de forma declarativa. Os sinais usados são os mesmos usados no exemplo de alinhamento da figura 3.6. De cima para baixo, têm-se, respectivamente, os espectrogramas do sinal de referência (a), do sinal de teste (b), e o sinal resultante do transplante do sinal de teste para a referência (c).

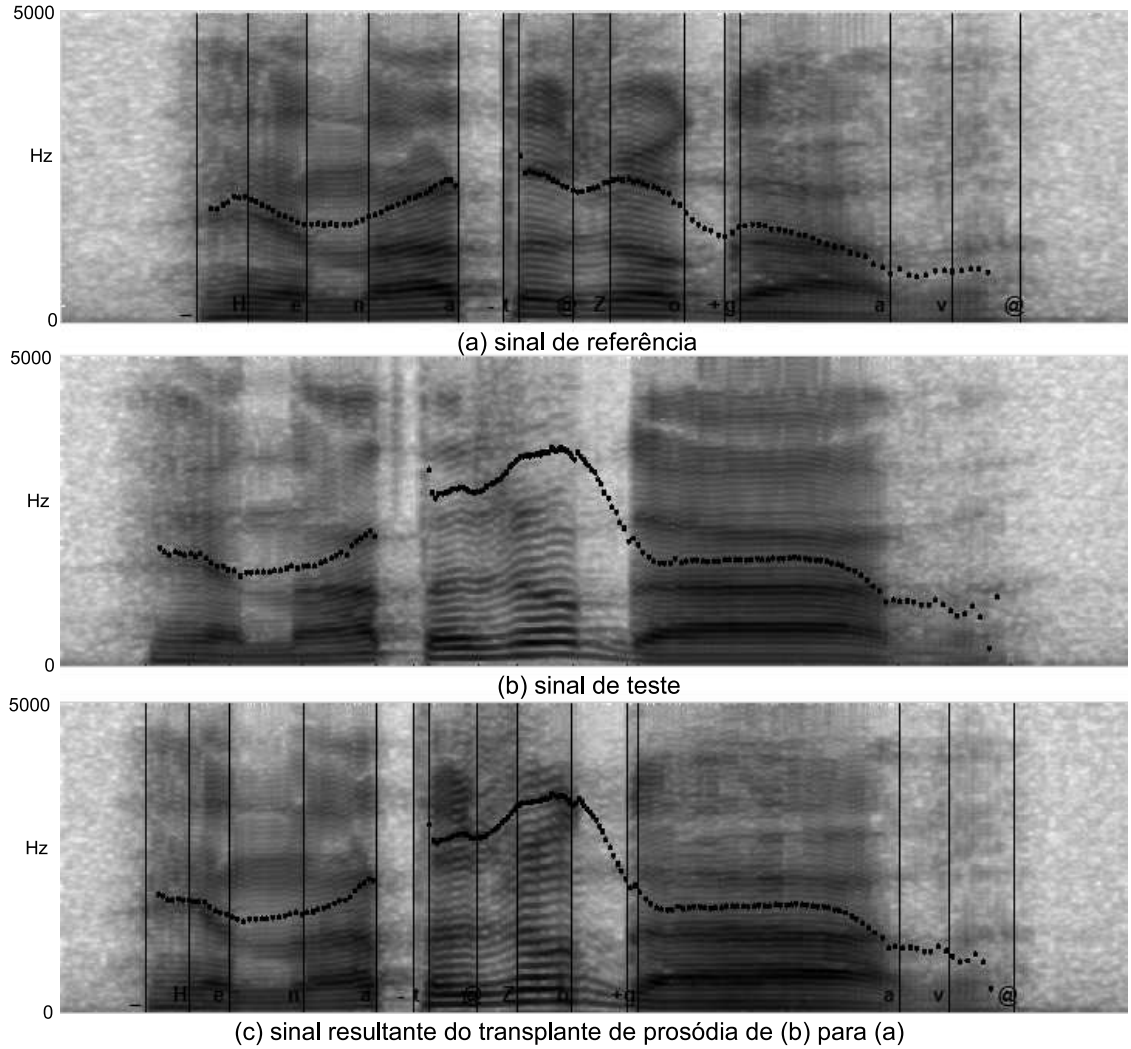


Figura 3.11: Exemplo do transplante de prosódia da sentença "Renata jogava" usando como referência (a) um sinal sem expressividade e como sinal de teste (b) um padrão prosódico de uma atitude de aviso.

A semelhança dos espectrogramas do sinal de teste (b) e do sinal resultante (c) aponta para um resultado bem sucedido do transplante. Comparano a reprodução dos sinais, nota-se que as características prosódicas foram inteiramente transferidas, transformando o sinal de prosódia neutra para uma atitude de aviso.

No exemplo da figura 3.11 estão representadas as etiquetas do sinal de referência

inseridas previamente. Na execução do transplante, a curva de alinhamento entre os sinais foi utilizada para a etiquetagem automática, tal como descrito na Seção 3.2, mapeando as etiquetas da referência para o sinal resultante do transplante, mostradas na figura inferior.

A qualidade do sinal resultante está sujeita à limitação do procedimento do PSOLA no qual é citado manter baixa distorção para uma taxa de variação do pitch em uma faixa de 0,5 a 2,0 e para uma contração e expansão das durações por um fator de 0,25 a 2,0 [3]. Assim, um transplante com variações extremas no pitch e nas durações poderá resultar em distorções, mesmo com o alinhamento e transplante bem sucedidos.

Por dividirem o mesmo caminho de alinhamento, podemos supor que os resultados obtidos na etiquetagem automática refletem o resultado do alinhamento a ser usado na execução do transplante. Assim, na seção a seguir serão descritos os resultados do alinhamento a partir da etiquetagem automática.

3.4 Avaliação quantitativa do algoritmo PSDTW-OLA

A eficiência da etiquetagem automática, em posicionar corretamente as etiquetas correspondentes às fronteiras fonéticas, foi avaliada comparando as etiquetas inseridas manualmente com as etiquetas inseridas automaticamente. Para esta avaliação, foi usado um conjunto de teste contendo 27 logatomas e 8 sentenças, incluindo 1 frase de outro falante, feminino. Os logatomas foram incluídos nesta análise para aumentar a variabilidade de transições fonéticas.

O sinal de referência foi obtido pelo sistema de síntese por concatenação de unidades que vem sendo desenvolvido, que fornece o instante de tempo das etiquetas contidas nas unidades de concatenação.

Para cada logatoma e frase do conjunto de teste foram inseridas etiquetas manualmente, com o auxílio visual do espectrograma e das amostras do sinal ao longo do tempo. No total, 313 etiquetas foram determinadas e gravadas. Em seguida, as etiquetas do conjunto de teste foram então inseridas automaticamente por alinhamento conforme diferentes parâmetros.

Dois critérios de análise foram utilizados para avaliar a eficiência do sistema em posicionar corretamente as marcas: o erro médio quadrático (MSE - *mean square error*), tal como usado em [39], e a verificação da taxa de erros menores que 5 ms, 10 ms, 15 ms, 20 ms e 30 ms, similar ao utilizado em [45].

Diferentes restrições ao caminho impostas ao DTW implicam em resultados diferentes na configuração da curva de alinhamento e também no mapeamento das

etiquetas. Assim, o alinhamento foi executado conforme as diferentes restrições ao caminho mostradas na figura 3.3. Na tabela 3.1 são exibidos os resultados obtidos com as restrições do tipo C, D, E e F. As restrições do tipo A e B não foram incluídas na tabela por oferecerem resultados muito inferiores aos apresentados.

Tabela 3.1: Resultados utilizando diferentes restrições ao caminho.

restrição	MSE (ms)	< 5ms	< 10ms	< 15ms	< 20ms	< 30ms
tipo C	10,80	57%	80%	88%	93%	97%
tipo D	13,46	40%	65%	80%	89%	96%
tipo E	11,01	52%	77%	87%	93%	97%
tipo F	10,71	55%	79%	89%	93%	97%

Observa-se na tabela 3.1 que as restrições que permitem somente passos unitários, do tipo C e F, apresentam melhores resultados globais do que as restrições do tipo D e E que permitem transições verticais duplas, triplas e quádruplas, conforme pode ser visto na figura 3.3. Comparando-se os tipos C e F, nota-se um desempenho razoavelmente equivalente. Porém, em um caso em que os sinais venham a ter durações muito diferentes, espera-se que o alinhamento seja mais bem sucedido com o tipo F, por permitir expansões de até 4 vezes.

Em uma segunda análise, a experiência acima foi refeita com o alinhamento realizado em partes, limitado pelos intervalos sonoros e surdo, conforme proposto na Seção 3.2. Na tabela 3.2 são mostrados os resultados obtidos com a modificação.

Tabela 3.2: Resultados utilizando alinhamento por partes sonoras/surdas.

restrição	MSE (ms)	< 5ms	< 10ms	< 15ms	< 20ms	< 30ms
tipo C	9,51	63%	81%	92%	95%	98%
tipo D	12,67	52%	70%	82%	91%	94%
tipo E	9,66	63%	80%	89%	94%	98%
tipo F	9,29	62%	82%	92%	95%	98%

Comparando as tabelas 3.1 e 3.2 nota-se uma redução global do erro médio quadrático e um aumento na precisão do posicionamento das etiquetas para todas as faixa percentuais consideradas.

Em [39] os autores atentam para a assimetria entre os sinais de teste e de referência, notando ser esta a maior fonte de falha no alinhamento e observando que em alguns casos é preferível trocar o sinal de teste com a referência. Assim, em uma segunda análise, foram repetidas as mesmas configurações anteriores, porém os sinais de referência e teste forma trocados, no qual observou-se que os resultados com o tipo C e F permanecem inalterados, por serem restrições simétricas, enquanto que os resultados com o tipo D e E foram degradados. Daí, conclui-se que a escolha pela restrição do tipo F, parece ser a mais adequada para a realização da etiquetagem, diferentemente da restrição do tipo E, adotada em [45].

Para o transplante de prosódia, foi proposto executar o alinhamento síncrono com o pitch de modo a simplificar o procedimento de OLA. Deseja-se portanto observar quais as conseqüências de se executar o alinhamento síncrono com o pitch. Uma vez que o transplante utiliza o mesmo procedimento de alinhamento executado na etiquetagem, então tais conseqüências podem ser observadas pela etiquetagem.

Na tabela 3.3 são mostradas as combinações executadas no alinhamento, onde UV representa o alinhamento por partes e PS quando realizado síncrono com o pitch. Comparando-se as linhas 1 e 2 da tabela 3.3 temos o resultado já observado de que o alinhamento parcial de intervalos sonoros e surdos (UV), apresenta a diminuição do erro médio.

Tabela 3.3: Resultado utilizando diferentes restrições ao caminho.

UV	PS	MSE (ms)	< 5ms	< 10ms	< 15ms	< 20ms	< 30ms
-	-	10,71	55%	79%	89%	93%	97%
X	-	9,29	62%	82%	92%	95%	98%
-	X	11,09	55%	76%	89%	92%	95%
X	X	9,51	58%	80%	92%	95%	97%

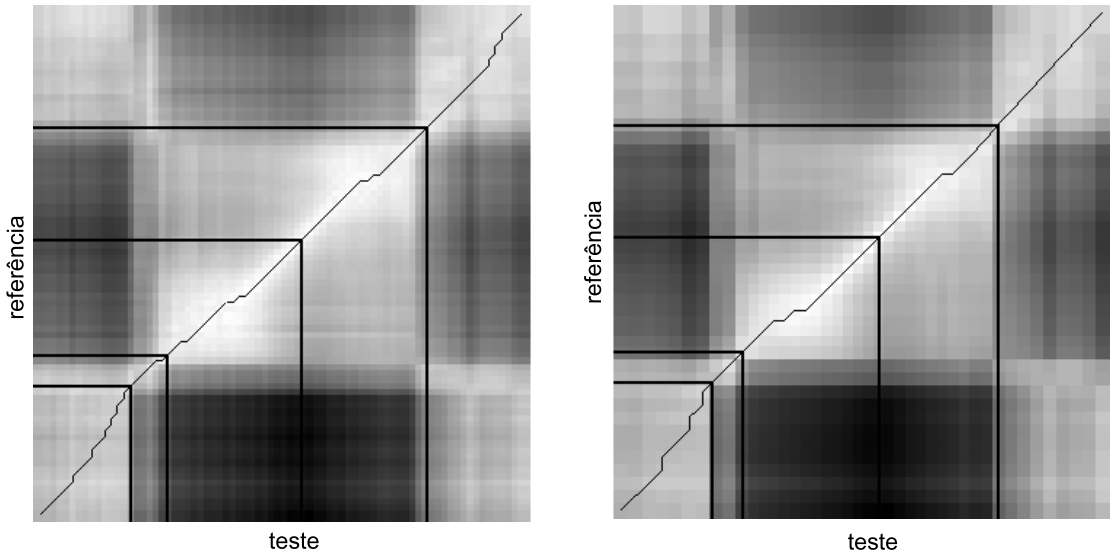
No entanto, comparando-se as linhas 1 e 3 da tabela 3.3 constata-se que a extração de parâmetros e o alinhamento síncrono com o pitch (PS) apresenta um aumento no erro médio e uma leve degradação na precisão da etiquetagem. Isto se deve à diminuição na resolução do *grid* de alinhamento, devido à diminuição da taxa de quadros e menor sobreposição entre quadros. Na figura 3.12 está representada a matriz de similaridade e o caminho de alinhamento, com a extração de características feita com quadros uniformes, obtidos a cada 5 ms, e com quadros síncronos com o pitch. Observa-se que quando é utilizada a extração síncrona com o pitch a resolução é menor.

Assim, constata-se comparando-se as linhas 1 e 4 da tabela 3.3 que apesar da perda de resolução devida à análise síncrona com o pitch, a execução do alinhamento em partes compensa boa parte desta perda.

Na execução do transplante pelo método convencional e pelo método proposto PSDTW-OLA, não se observou perceptualmente diferença entre os sinais, o que justifica o uso do método proposto.

3.5 Conclusões

Neste capítulo, foram apresentados os sistemas de etiquetagem automática por alinhamento e de transplante de prosódia desenvolvidos neste trabalho. Como base para a apresentação de um método de transplante proposto, foram também descritos



(a) Quadros uniformes.

(b) Quadros síncronos com o pitch

Figura 3.12: Comparação da resolução.

os princípios básicos no procedimento de alinhamento por DTW e de manipulação da duração e do contorno de pitch pelo algoritmo PSOLA.

Constatou-se que se o alinhamento é bem sucedido, a etiquetagem tem a principal característica de reproduzir o procedimento de etiquetagem da referência, ou seja, as marcações de fronteiras que serão inseridas na referência serão feitas em conformidade com o critério presente na marcação das fronteiras no sinal de teste, o que manualmente nem sempre acontece [46].

Na comparação entre marcas inseridas manualmente com marcas inseridas automaticamente, observou-se um erro médio quadrático de 10,7 ms, onde 93% das etiquetas apresentaram um erro menor do que 20 ms. Para a comparação fiel destes resultados com outros trabalhos, os dados de análise precisariam ser os mesmos, porém observa-se coerência com os valores obtidos em [39].

Foi observado ainda que no alinhamento, quando é utilizada uma sentença natural como referência, o desempenho é superior do que quando usada uma sentença gerada por síntese. Para obter-se um desempenho similar em ambos os casos, seria necessário um sintetizador de alta qualidade segmental.

Outro fato observado no alinhamento de certos segmentos é que certos parâmetros representativos do espectro parecem ser mais importantes do que outros para obter-se um alinhamento bem sucedido. Por exemplo, na separação das fricativas a energia é um fator importante, enquanto que nem tão importante para a fronteira entre vogais. Isto sugere um procedimento de otimização para a atribuição de pesos ao vetor de parâmetros, ou ainda variando-se tais pesos conforme a classificação dos segmentos, tal como em [43].

Foi proposto ainda que no alinhamento entre dois sinais, quando são conhecidos

os intervalos sonoros e surdos de ambos os sinais, o alinhamento pode ser realizado em partes, onde os resultados indicam uma diminuição do erro médio quadrático em 15% quando este procedimento é usado.

Na execução do transplante de prosódia, foi proposto ainda que se as marcas de pitch de ambos os sinais estão disponíveis, então a conjugação do DTW executado de maneira síncrona com o pitch com o PSOLA, nomeado de PSDTW-OLA [49], resulta na simplificação do procedimento de manipulação de janelas executado pelo PSOLA para atender os requisitos de pitch e durações simultaneamente. Foram avaliadas as consequências de se realizar a análise e o alinhamento com janelas síncronas com o pitch, ao invés de janelas obtidas uniformemente espaçadas. Pela monitoração do erro entre etiquetas inseridas manualmente com as etiquetas inseridas pelo alinhamento, constatou-se que a perda de resolução devido a análise síncrona com o pitch é compensada ao se utilizar o alinhamento parcial entre segmentos sonoros e surdos.

Portanto, conclui-se que, quando os sinais possuem exatamente o mesmo conteúdo, tanto a etiquetagem automática quanto o transplante por alinhamento são ferramentas úteis. O método de transplante PSDTW-OLA apresentado é um procedimento conveniente pela sua simplicidade. Nos capítulos posteriores, serão descritos os procedimentos para permitir o transplante entre sentenças de conteúdos diferentes.

Capítulo 4

Codificação das variáveis prosódicas

4.1 Introdução

Um dos objetivos deste trabalho é observar o comportamento das variáveis prosódicas em um nível maior de abstração do que por uma sequência de pontos ao longo do tempo. A solução proposta consiste inicialmente em codificar ou parametrizar as variáveis de duração, intensidade e pitch, em um dado intervalo de tempo, conforme limites que possam ser determinados a partir do texto, configurando uma "unidade de análise".

Vários trabalhos têm sugerido diferentes limites e configurações para estas unidades, tais como intervalos sonoros ou surdos, sílabas, etc. inclusive para o português [50], em que o autor considera o intervalo vogal-vogal como unidade. Neste trabalho foi adotada a sílaba como unidade, que sob o ponto de vista linguístico é considerada uma unidade mínima da estrutura prosódica da sentença [31]. Assim, cada sílaba é codificada por um conjunto de parâmetros obtidos pelo comportamento das variáveis prosódicas dos segmentos contidos, cujo procedimento de determinação destes parâmetros foi nomeado de análise.

A normalização implica em assumir que os segmentos possuem características próprias, capazes de serem estimadas pela distribuição estatística ao longo de um *corpus* de análise. Ao propor um modelo para estimação das durações, CAMPBELL [6] defende um ponto de vista que fundamenta todo o restante do seu estudo, e que também norteia este trabalho: "dois sons similares raramente têm a mesma duração, mas a análise estatística da distribuição das durações revela padrões consistentes. O ponto chave é que a medida de alargamento de um segmento pode fornecer muito mais informações úteis do que a medida da própria duração. A duração *per se* tem pouco significado e é de pouca utilidade, a não ser que saibamos a distribuição

exibida por segmentos similares e o grau de variação desta duração".

Assim, para a obtenção de dados para a normalização das durações e intensidade dos segmentos e codificação da sílaba, assim como para o estudo posterior do comportamento da sílaba ao longo de sentenças, optou-se por construir um pequeno *corpus* de análise. Na Seção 4.2 é descrita a construção deste *corpus* desde a gravação até a etiquetagem das fronteiras fonéticas. Ao longo do trabalho foi observada a importância de manter um critério coerente no posicionamento das fronteiras, e por isso optou-se por dar atenção à descrição deste critério, apresentado no Apêndice B, embora sem a pretensão de estabelecer um critério ótimo de etiquetagem.

Na Seção 4.3, é apresentado o procedimento de obtenção da distribuição estatística das variáveis de duração e intensidade para os segmentos etiquetados, permitindo a normalização destas variáveis.

Em seguida, na Seção 4.4 é descrito o procedimento que nomeou-se de análise, onde, a partir da normalização dos segmentos da sílaba, os parâmetros de codificação da sílaba são determinados. Na Seção 4.5, é descrito o processo nomeado se síntese, onde a manipulação direta destes parâmetros da sílaba resulta em curvas que descrevem a alteração das variáveis prosódicas no domínio do tempo. Os exemplos apresentados ao longo destas seções foram obtidos a partir do *software* desenvolvido ao longo deste trabalho, que executa tanto o procedimento de análise quanto síntese, permitindo ainda a ressíntese do sinal através do algoritmo TD-PSOLA [38].

Por fim, na Seção 4.6 será apresentada uma análise quantitativa da codificação de sentenças com diferentes conteúdos.

4.2 Construção do *corpus* de análise

O *corpus* é formado basicamente por 200 sentenças foneticamente balanceadas, descritas em [51]. A gravação foi feita pelo autor desta tese, natural da cidade de Petrópolis do estado do Rio de Janeiro.

4.2.1 Gravação do *corpus*

Para a gravação das sentenças do *corpus*, foi imposto pronunciá-las, na medida do possível, sem expressividade, ou seja, sem nenhuma manifestação de atitude, sendo algumas sentenças pronunciadas como declarativas finais e outras como declarativas não finais. Por se tratarem de sentenças de curta duração e sem interrupção por pontuação, buscou-se ainda manter uma fala contínua, sem a ocorrência de pausas.

Cada sentença foi lida a *priori*, antes da produção, havendo portanto um tempo de preparação, o que se difere de um processo de leitura onde a sentença é lida ao mesmo tempo em que é apresentada, sem o conhecimento prévio. Assim, podemos

considerar o *corpus* como sendo de fala semi-espontânea.

Pelo fato do falante não ser um locutor profissional, é esperado que alguns hábitos ou costumes próprios do dialeto do falante estejam expressos na gravação do *corpus*, apesar da tentativa de minimizá-los.

A gravação do sinal de fala foi acompanhada da gravação simultânea do sinal de atividade das cordas vocais, captado por um eletroglotógrafo. Os sinais foram adquiridos por uma placa de som convencional e um pré-amplificador externo, sendo gravados em formato .wav estéreo, com taxa de amostragem de 22.500 Hz e 16 bits.

Buscou-se manter a proximidade do microfone constante, usando um microfone de boca, fixo aos fones de ouvido, e ainda tentou-se manter constantes as condições de ruído e horário de gravação, de maneira a equalizar a influência de tais fatores na intensidade do sinal.

Após a gravação, cada sentença foi recortada para manter um período de silêncio inicial e final similares, em torno de 200 ms, e em seguida as marcas de pitch foram determinadas automaticamente a partir do sinal do eletroglotógrafo, podendo ser corrigidas. Por fim, cada sentença está então pronta para a inserção das etiquetas equivalentes às fronteiras fonéticas, conforme o critério descrito na seção a seguir.

Todo este procedimento de gravação, recorte, visualização da amplitude dos sinais e do espectro, assim como a edição das marcas de pitch e das etiquetas, é realizado através do *software* que vem sendo desenvolvido, que será descrito no Capítulo 6.

4.2.2 Critério de etiquetagem

Diferentes autores convergem para a mesma opinião de que tratar a fala como uma sequência delimitada e não sobreposta de elementos é uma simplificação do complexo processo de articulação da fala. Por exemplo, em [19] o autor cita que é notoriamente difícil definir a fronteira entre vogais ou entre vogais e semivogais *(glides)* e reitera que em muitos casos o posicionamento e inserção de uma marca de fronteira é arbitrário. No entanto, para a pesquisa na área de processamento da fala, é assumido que existem fronteiras que delimitam os segmentos e o processo de etiquetagem consiste em atribuir etiquetas a estas fronteira.

A etiquetagem deste *corpus* foi feita manualmente pelo autor deste trabalho, que conforme TAYLOR [1] pode ser caracterizada como uma etiquetagem intuitiva, onde o julgamento é baseado na própria habilidade com a linguagem.

São muitos os trabalhos que se propõem a segmentar ou etiquetar um *corpus* por meio de um procedimento de aprendizagem automática, cujo treino e/ou validação normalmente utilizam um *corpus* etiquetado previamente de forma manual. Porém, são poucos os trabalhos que descrevem detalhes sobre a metodologia para a inserção

manual destas etiquetas, a exemplo de [52] [19] [53] [44].

Para a etiquetagem do sinal de fala, é imposta a restrição de que as etiquetas a serem inseridas possuam correspondência unívoca com a transcrição obtida automaticamente a partir do texto. Um procedimento similar é utilizado em [19] onde cada etiqueta é associada a um fonema e, portanto, a mesma etiqueta é usada para os alofones de um mesmo fonema, apesar das diferenças acústicas. Assim, para cada sentença, a identificação das etiquetas a serem inseridas no sinal de fala são previamente determinadas pela transcrição automática obtida pelo processamento do texto, conforme descrito no Capítulo 2.

Este procedimento de restringir a etiquetagem à transcrição automática obtida do texto, se deve a duas razões principais: primeiramente, permite estabelecer o sincronismo entre o conteúdo do texto e o sinal de fala, isto é, como cada etiqueta é associada a uma referência de tempo do sinal de fala, então devido ao formato de listas encadeadas hierarquicamente, é possível determinar os limites das sílabas, palavras e grupos de palavras, sem a necessidade de etiquetagem em todos os níveis. Em segundo lugar, forçar a correspondência implica em adotar na etiquetagem a coerência das regras utilizadas na transcrição automática.

No Apêndice B, são apresentados exemplos do critério de posicionamento das etiquetas. Dentre os exemplos citados convém ressaltar alguns casos particulares:

- As oclusivas foram etiquetadas por duas fases: uma fase fechada e uma fase de oclusão. Igualmente, as africadas foram etiquetadas por duas fases: uma fase fechada e uma fase fricada.

- SEARA [54] (apud [55]) e SOUZA [56] concordam que foneticamente as vogais nasais configuram pelo menos duas fases: uma porção da vogal e o término em um "murmúrio nasal". SOUZA [56] observa que quando há uma consoante oclusiva posterior à vogal nasal, o tempo da oclusão é inegavelmente menor. Isto parece indicar que a vogal está "roubando" o tempo da oclusiva [56]. Por esta influência do murmúrio nasal na duração das oclusivas, a vogal nasal foi etiquetada em duas fases, isolando o murmúrio nasal, quando houver uma consoante oclusiva posterior.

- A vogal epentética e a vogal de apoio, conforme definidas em [57], foram etiquetadas em alguns casos. A vogal de apoio é etiquetada sempre que diante de [r]. A vogal epentética foi etiquetada quando a consoante posterior for uma nasal, quando estiver entre oclusivas, ou entre uma oclusiva e fricativa.

- No caso de supressão de um segmento, mesmo que o segmento não exista no sinal, foi feita a etiquetagem com duração desprezível.

Ao fim da etiquetagem das 200 frases, após algumas revisões, foram obtidos aproximadamente 7000 segmentos, cuja ocorrência relativa se apresenta distribuída conforme a figura 4.1.

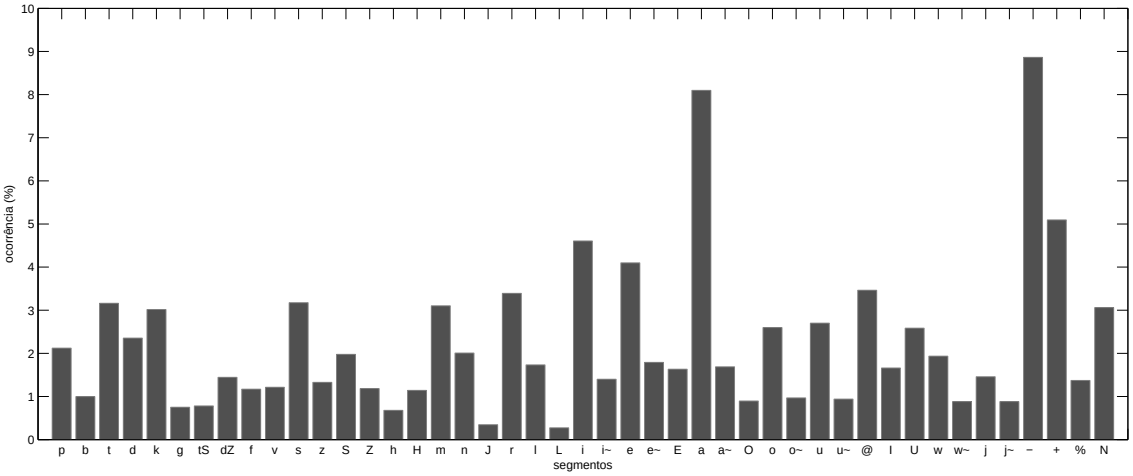


Figura 4.1: Ocorrência relativa dos segmentos presentes no *corpus*.

4.3 Normalização dos segmentos

A partir da etiquetagem dos segmentos foram obtidas as distribuições das variáveis de duração e intensidade dos segmentos presentes no *corpus*, permitindo normalizar estas variáveis.

4.3.1 Durações

A duração de um segmento é dada simplesmente pela diferença de tempo entre as etiquetas que delimitam as fronteiras do segmento.

CAMPBELL [58], ao apresentar uma primeira proposta para o modelo das durações, utiliza a distribuição normal para modelar a duração dos segmentos. Ao aprimorar este modelo, ao longo de publicações subsequentes, CAMPBELL [6] propõe utilizar distribuições diferentes da normal, tais como lognormal e gamma.

Neste trabalho, será usada a distribuição lognormal por ser considerada mais adequada aos dados do que a distribuição normal e mais simples de ser tratada do que outras distribuições. Na figura 4.2 é mostrado um exemplo da modelagem da distribuição da duração da vogal [a] por uma lognormal. Por conveniência de manipulação numérica, as durações obtidas em milisegundos (ms) foram convertidas para o domínio log e a modelagem foi feita por uma curva normal.

Para um dado segmento identificado por s e de duração d, o fator de normalização z_d, comumente chamado de *z-score*, para uma distribuição lognormal de média μ_d^s e

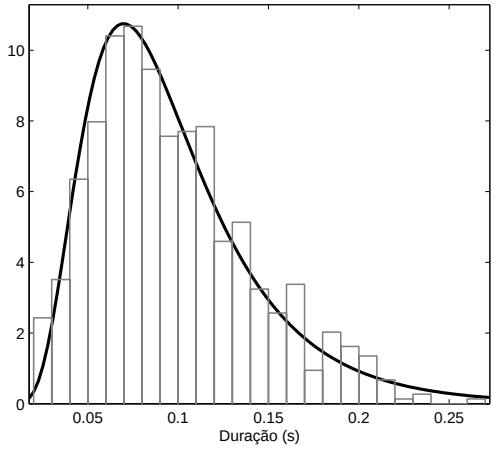


Figura 4.2: Distribuição lognormal das durações.

desvio padrão σ_d^s, é dado pela relação:

$$z_d(d, s) = \frac{\ln(d) - \mu_d^s}{\sigma_d^s}. \tag{4.1}$$

De modo inverso, a duração do segmento s, com fator de normalização z_d é obtida pela relação inversa:

$$d(z_d, s) = e^{(\mu_d^s + z_d \sigma_d^s)}. \tag{4.2}$$

Portanto, para a obtenção das durações normalizadas, é preciso primeiramente obter a distribuição das durações para cada tipo de segmento do *corpus*. Na figura 4.3 é mostrada a distribuição individual da duração de cada segmento. Nessa figura, é possível identificar os segmentos curtos, cuja média é baixa, tais como as oclusivas, e segmentos tipicamente longos, com média alta, tais como a vogais. Observa-se ainda que as vogais, por exemplo, exibem uma dispersão acentuada da média, o que indica serem segmentos mais elásticos, ou seja, mais suscetíveis ao alargamento e compressão. As oclusivas, por outro lado, possuem menor desvio da média, mostrando-se menos elásticas do que as vogais.

A partir desta análise estatística, seria trivial a determinação dos parâmetros μ_d^s e σ_d^s para cada segmento quando considerado isoladamente. Porém, dois problemas sugerem uma abordagem mais refinada na obtenção das distribuições das durações: a influência que o segmento sofre em função do contexto onde ele se insere; e a pouca variabilidade dos segmentos em diferentes contextos, tal como apontado em [59]. Considerando estes problemas, CAMPBELL [6] propõe aprimorar o modelo básico usando distribuições específicas conforme o contexto, e reagrupando os segmentos em classes mais amplas para suprir a pouca variabilidade.

Seguindo a proposta de CAMPBELL, dentre os diversos contextos de ocorrência de um segmento, foram considerados dois casos principais: a posição que o segmento

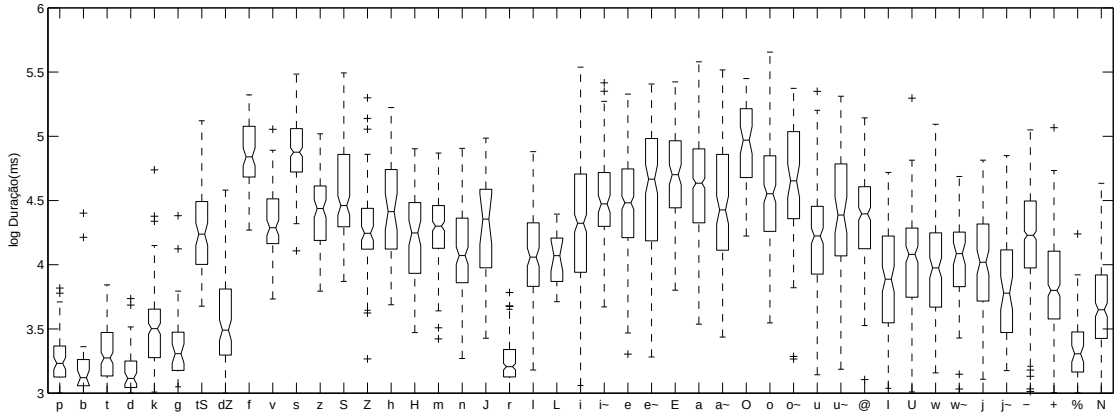


Figura 4.3: Distribuição do log da duração dos segmentos presentes no *corpus*.

ocupa na palavra e a influência dos segmentos adjacentes. São várias as combinações contextuais possíveis para os segmentos, o que tornaria inviável uma análise caso a caso. Assim, optou-se por observar alguns casos contextuais particulares, que se mostraram relevantes para a normalização, e são eles: o contexto de posição das vogais orais e nasais dentro da palavra; as vogais em ditongo e as oclusivas quando precedidas por nasais.

Para a indicação de agrupamento dos segmentos em classes mais amplas foi usada a medida de similaridade ANOVA para comparar as distribuições. Uma significância de $p > 0{,}05$ foi adotada como indicativa de agrupamento. Nos casos de ausência de amostras executou-se o agrupamento por analogia, ou conforme características articulatórias semelhantes.

As vogais

Em análise da duração das vogais, sabe-se a *priori* que a duração das vogais em contexto tônico é maior do que em contexto não tônico. Assim, um desequilíbrio na ocorrência das vogais em contexto tônico ou não tônico pode resultar em uma distribuição tendenciosa. A figura 4.4 mostra a ocorrência das vogais em posições pretônica, tônica, postônica medial e postônica final, quando observadas fora de encontros vocálicos.

A princípio, nota-se que somente as vogais [@ U I] ocorrem em posição postônica final. Isto se deve à redução vocálica dos segmentos [a o e] para [@ U I], respectivamente, quando em posição postônica final [30]. Pela mesma razão, os segmentos [a i u] apresentam ocorrência predominante em posição postônica medial, pois, devido à junção de palavras, os segmentos postônico finais [@ U I] são elevados para os postônicos mediais [a i u], respectivamente. Nota-se ainda que as vogais [E O] ocorrem predominantemente em posição tônica, o que se deve a uma restrição própria

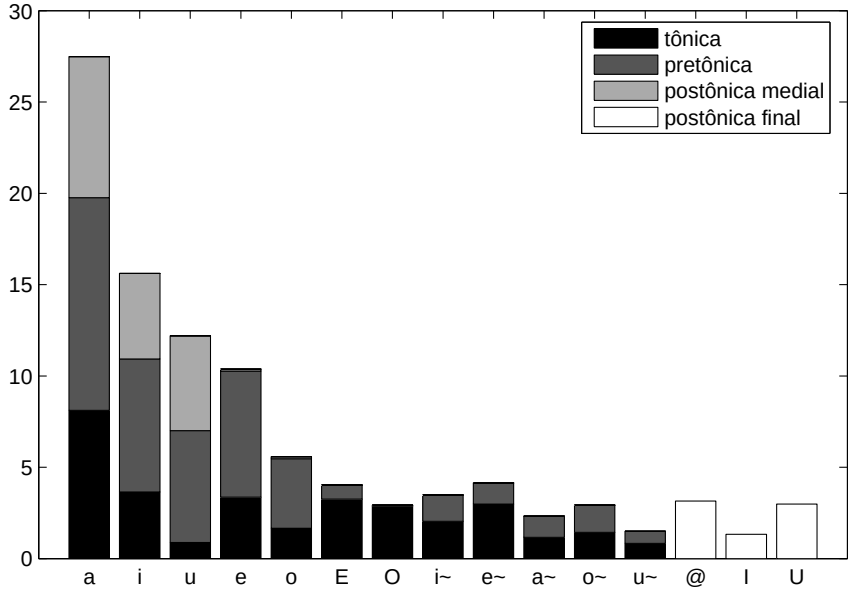
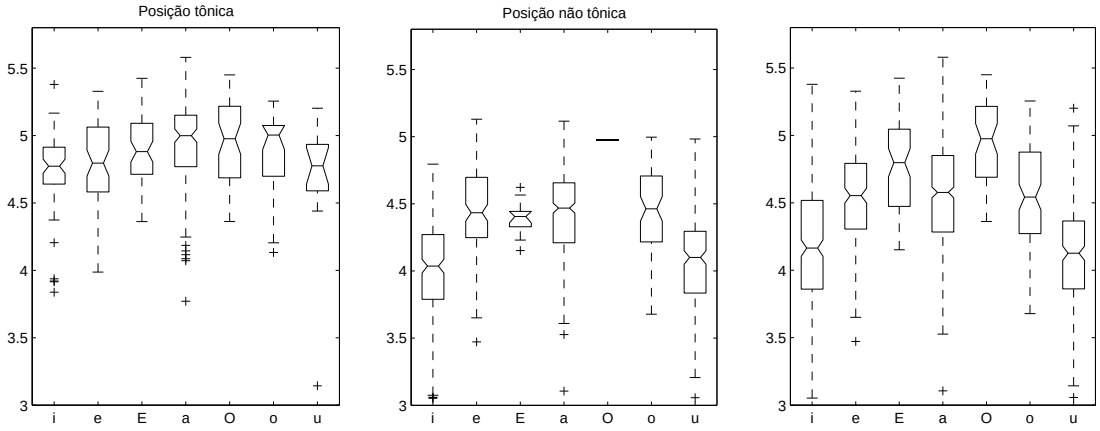


Figura 4.4: Ocorrência dos segmentos vocálicos conforme a posição em relação à tônica.

da língua.

Observando a distribuição da duração das vogais conforme a posição, a figura 4.5 mostra a distribuição das vogais em contexto tônico 4.5a, não tônico 4.5b e em qualquer contexto 4.5c.



(a) contexto tônico (b) contexto não tônico (c) Qualquer contexto

Figura 4.5: Distribuição do log da duração das vogais conforme o contexto tônico.

Na figura 4.4, todas as vogais ocorrem na posição tônica. Ao tomarmos a distribuição condicionada a tonicidade, conforme a figura 4.5a, nota-se curiosamente um formato triangular que sugere o triângulo de classificação das vogais conforme a articulação: [i u] altas, [e o E O] médias e [a] baixa. Porém, em contexto não

tônico, como mostra a figura 4.5b, a ocorrência praticamente nula das vogais [E O] leva a um desequilíbrio. Isto explica o fato de que nas distribuições em qualquer contexto estas vogais aparecem com média superior, pois predominam os valores cujo contexto é tônico.

Esta falta de amostras das vogais [E O] em contexto não tônico precisa ser compensada. Por exemplo, se considerarmos as palavras [bOla] e [bala], ao normalizarmos as durações dos segmentos, espera-se que os valores normalizados do [O] e [a] tônicos sejam bem próximos. No entanto, se a média na vogal [O] é relativamente mais alta do que a vogal [a], o fator de normalização não será próximo.

A significância entre as distribuições no contexto tônico indica que as vogais [e E o O] podem ser agrupadas. Em contexto não tônico, a significância também indica o agrupamento de [e a o]. Assim, para suprir a falta de dados do [E O] em contexto não tônico e ainda a falta de dados de [e o] em posição postônica final, as vogais [e E a O o] foram agrupadas em uma única distribuição.

As distribuição das vogais [i u] apresentaram significância em qualquer contexto e, portanto, também foram agrupadas em uma única classe.

Além das vogais orais, temos as vogais nasalizadas e as vogais nasais. Sabe-se de antemão que a duração das vogais nasais são em média maiores do que as vogais nasalizadas ou orais [55]. Porém, com a etiquetagem do murmúrio nasal, a média das vogais nasais se aproxima das vogais nasalizadas e orais.

Nota-se que as vogais nasais e nasalizadas ocorrem basicamente em contexto tônico e pretônico, como pode ser visto na figura 4.4. A falta de amostras em contexto postônico faz com que a média destas vogais, quando consideradas em qualquer contexto, seja predominantemente mais alta em comparação com as vogais orais. Ao compararmos a distribuição das vogais orais nasais e nasalizadas em contexto tônico, obtém-se uma alta significância, o que indica que em contexto tônico poderiam ser agrupadas. Pela falta de amostras em outros contextos, optou-se por agrupar as vogais nasais, nasalizadas e orais [ẽ õ ã e o a E O] em uma mesma classe. Do mesmo modo, as vogais [ĩ ũ i u] também foram agrupadas em uma única classe, amenizando o problema de falta de amostras das vogais nasais no contexto não tônico.

Ditongos

Outro caso problemático são as vogais em ditongos orais e nasais, onde a junção vogal+glide ou glide+vogal faz com que a duração média da vogal diminua. Por exemplo, em comparação das palavras “seita” [sejt@] e “seta” [sEt@], o ditongo [ej] na primeira palavra apresenta quase a mesma duração da vogal [E] na segunda palavra. A razão deste fato provavelmente deve-se à conjugação de fatores como a coarticulação entre os sons e uma tendência a manter a duração das sílabas simi-

lar (isocronia silábica). Alguns autores, tal como [20], decidem considerar certos ditongos como unidades próprias. Porém, nesta tese, como o glide e a vogal foram etiquetados separadamente, decidiu-se por adotar distribuições condicionais para as vogais em ditongos. Devido à baixa ocorrência dos ditongos, e seguindo o agrupamento das vogais, as vogais orais em ditongos crescentes e decrescentes, assim como as vogais nasais em ditongo decrescente, foram todas agrupadas em uma única distribuição condicional.

Na distribuição dos glides orais, nota-se que quando estão em um ditongo crescente eles possuem média menor do que quando em ditongos decrescente. Assim, foram criadas distribuições condicionais para os glides, quando sucedidos ou precedidos por uma vogal.

Os glides nasais aparecem sempre precedidos por uma vogal nasal e quando comparados aos glides orais na mesma posição, apresentam significância que sugere serem agrupados na mesma distribuição condicional.

As oclusivas

Na interação entre as vogais nasais e as oclusivas, o murmúrio nasal da vogal parece roubar a duração da fase oclusiva [56]. Por esta razão, foram criadas distribuições condicionais para as fases fechadas das oclusivas sonoras e surdas [+ -], quando precedidas pelo murmúrio nasal. Do mesmo modo, a distribuição do murmúrio nasal [N] foi condicionada a uma oclusão surda ou sonora.

Na tabela C.2, no Apêndice C são apresentados os valores de média e desvio padrão usado para a normalização da duração dos segmentos, conforme os agrupamentos e distribuições condicionais consideradas. A partir destes valores, o fator de normalização z_d obtido para todos os segmentos do *corpus* estão distribuídos por uma normal de média zero e variância unitária.

No *software* desenvolvido, foi construído um objeto de desenho gráfico para visualizar o fator de normalização da duração dos segmentos ao longo de uma sentença. Na figura 4.6 é mostrado um recorte da imagem gerada pelo *software*, na normalização dos segmentos da sentença "Os professores acreditam nesta teoria". No gráfico, a linha horizontal indica $z_d = 0$, quando a duração dos segmentos é dada somente pelo valor médio μ_d. A altura das barras indica o fator de normalização $z_d > 0$ ou $z_d < 0$ e a largura indica o desvio padrão σ_d das distribuições, isto é, o quanto o segmento é suscetível de ser alongado ou comprimido. Portanto, a área limitada pelas barras indica a alteração da duração do segmento além da média. As vogais foram também marcadas com cor diferente para ressaltar sua importância.

A princípio, a evolução do parâmetro de normalização ao longo da sentença não mostra prontamente uma aparência de regularidade. Porém, na Seção 4.4.1

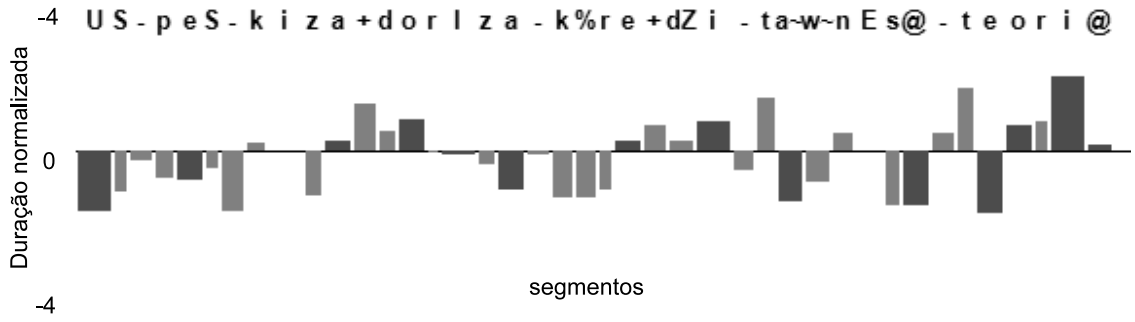


Figura 4.6: Gráfico da normalização da duração dos segmentos da sentença "Os professores acreditam nesta teoria". A altura das barras indica o valor do fator de normalização e a largura indica o desvio-padrão das distribuições. As vogais estão indicadas com um tom mais escuro.

de análise da sílaba será visto que ao considerarmos um comportamento uniforme interno à sílaba surgem padrões mais evidentes.

4.3.2 Intensidade

A intensidade do sinal de fala em um dado intervalo de tempo foi calculada a partir da energia média do sinal neste intervalo, após a aplicação de uma janela de Hamming para atenuar a descontinuidade nas bordas do intervalo. A energia média em um intervalo contendo N amostras de amplitude x_i é dada por:

$$e = \frac{1}{N} \sum_{i=0}^{N} x_i^2. \tag{4.3}$$

A intensidade I é representada em dB SPL (sound pressure level), ou seja, dB relativo a $2.10^{-}5$ pascal, que é o limiar da audição para uma onda de 1000 Hz [60]. Assim, a intensidade é dada por:

$$I = 10 \log_{10} e + 94. \tag{4.4}$$

A intensidade do sinal é influenciada pelas condições de gravação, tais como a proximidade do microfone, níveis de ruído e reverberação na sala de gravação. Porém, assume-se que estas condições sejam globais e constantes ao longo da gravação do *corpus*. Interessa-nos observar variações locais na intensidade ao longo de uma sentença, como por exemplo, o decaimento na intensidade dos segmentos em final de sentença ou o aumento da amplitude quando deseja-se dar ênfase. Assim, propõe-se o procedimento de normalização da intensidade dos segmentos de forma similar ao utilizado para as durações.

Para a caracterização da média e desvio padrão da intensidade dos segmentos foi considerada uma distribuição normal, como sugere a figura 4.7 que mostra a

distribuição da intensidade de vogais.

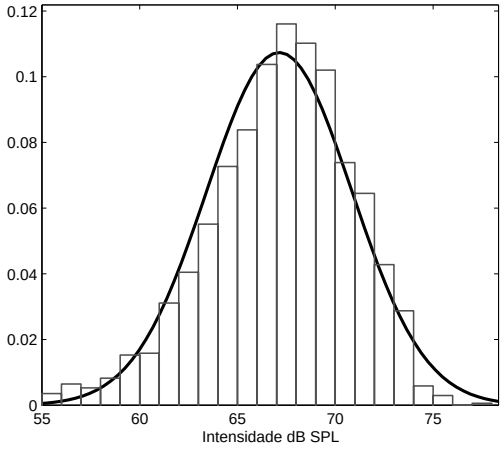


Figura 4.7: Distribuição da intensidade das vogais modelada por uma normal.

Analogamente às durações, para um segmento de identidade s cuja intensidade I é dada por uma distribuição normal de média μ_I^s e desvio padrão σ_I^s, o fator de normalização z_I é dado por:

$$z_I(I, s) = \frac{I - \mu_I^s}{\sigma_I^s}. \tag{4.5}$$

De modo inverso, a intensidade do segmento s a partir do fator de normalização é obtida por:

$$I(z_I, s) = \mu_I^s + z_I \sigma_I^s. \tag{4.6}$$

Na figura 4.8 está representada a distribuição da intensidade dos segmentos do *corpus*, onde observa-se que as vogais são as que exibem maior média e os segmentos oclusivos surdos apresentam médias mais baixas.

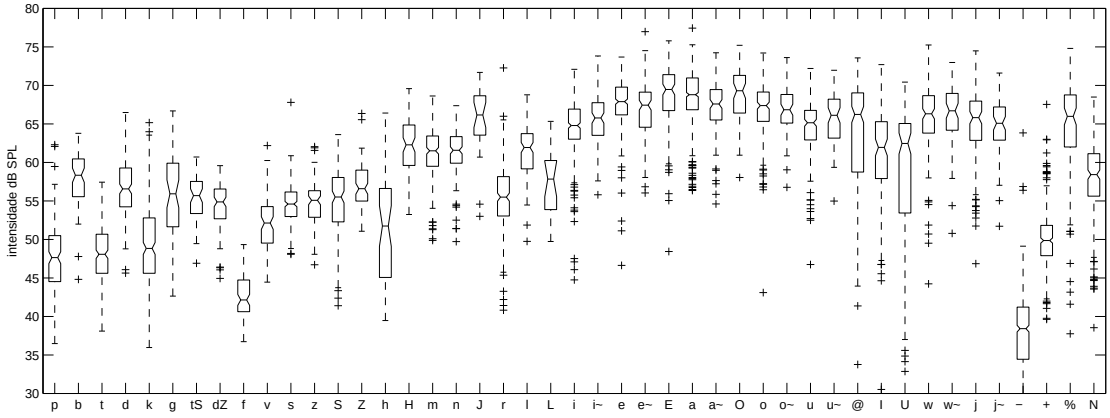


Figura 4.8: Distribuição da intensidade em dB SPL dos segmentos do *corpus*.

Buscando utilizar agrupamentos e distribuições condicionais, tal como nas durações, no caso da intensidade não foi possível identificar a influência contextual

entre segmentos adjacentes, e portanto, não foram consideradas distribuições condicionais. Entretanto, devido à influência do contexto tônico sobre a intensidade, buscou-se agrupar as vogais em classes mais amplas, para suprir a falta de amostras em diferentes contextos relativos à posição na palavra. A ocorrência das vogais conforme o contexto tônico e a significância entre as distribuições da intensidade das vogais, tal como nas durações, sugerem agrupar as vogais [ẽ õ ã e o a E O] em uma única distribuição, assim como o agrupamento de [i u ĩ ũ I U] e dos glides [w j w̃ j̃].

Na tabela C.1, no Apêndice C, são apresentados outros agrupamentos menos importantes e a média e o desvio padrão das distribuições consideradas.

No *software* desenvolvido, o fator de normalização das intensidade ao longo das sentenças também é representado. Na figura 4.9 é dado um recorte da imagem gerada pelo *software*, demonstrando a normalização dos segmentos da mesma sentença utilizada como exemplo para as durações: "Os professores acreditam nesta teoria". A linha horizontal indica $z_I = 0$, quando a intensidade dos segmentos é dada pelo valor médio μ_I. A altura das barras indica o fator de normalização $z_I > 0$ ou $z_I < 0$ e a largura indica o desvio padrão σ_I das distribuições, isto é, o quanto o segmento é suscetível de ter a intensidade aumentada ou diminuída. Portanto, a área limitada pelas barras indicam a alteração da intensidade em relação à média. Nota-se na figura a tendência do decréscimo na intensidade dos segmentos finais.

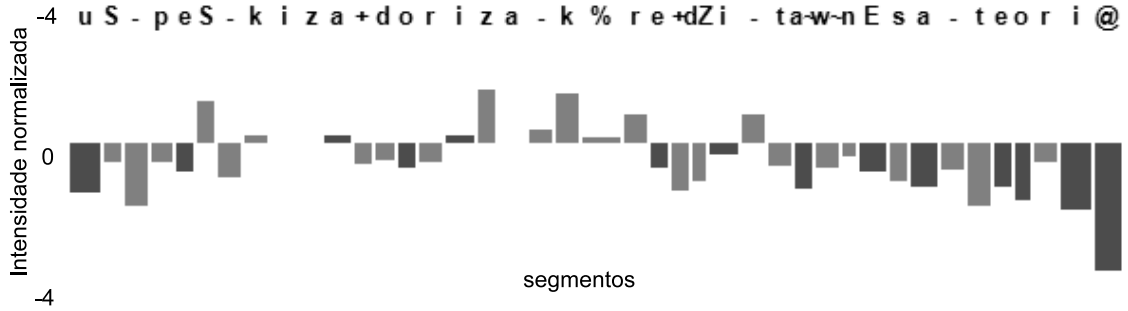


Figura 4.9: Gráfico da normalização da intensidade dos segmentos da sentença "Os professores acreditam nesta teoria". A altura das barras indica o valor do fator de normalização e a largura indica o desvio-padrão das distribuições.

4.4 Análise da sílaba

O objetivo desta seção é observar o comportamento da intensidade, durações e pitch em um nível mais abstrato do que os segmentos, ou seja, no nível da sílaba. É a partir dos fatores de normalização da intensidade e da duração dos segmentos que serão obtidos os parâmetros de normalização da intensidade e da duração da sílaba, procedimento no qual se nomeou de análise.

Os procedimentos de análise das durações e das intensidades das sílabas serão visto nas Seções 4.4.1 e 4.4.2 a seguir. A representação do contorno de pitch limitado à sílaba é um caso a parte que será visto na Seção 4.4.3.

4.4.1 Durações

O que diferentes modelos para as durações têm sugerido, tais como CAMPBELL [6] e BARBOSA [50] é que dentro de uma unidade limitada teoricamente, os segmentos exibem o mesmo fator de normalização. Este na verdade é o preceito básico do modelo de elasticidade proposto por CAMPBELL [6], que será apresentado a seguir.

O modelo de CAMPBELL

No modelo proposto em [58], CAMPBELL propõe que os segmentos limitados pela sílaba são alargados ou comprimidos pelo mesmo fator. Apesar da aplicação do mesmo fator de normalização a todos os segmentos, a duração de cada segmento é alterada em função da elasticidade própria de cada segmento, dada pelo desvio-padrão.

A título de demonstração, seja uma sílaba de duração D, contendo N segmentos s_j, de duração d_j, para $j = 1, 2, ..., N$. Considere a princípio que a distribuição da duração dos segmentos seja normal[1] com média μ_s e desvio padrão σ_s. Tem-se portanto que a duração D da sílaba, conforme a normalização da duração dos segmentos, é dada por:

$$D = \sum_{j=1}^{N} d_j = \sum_{j=1}^{N} (\mu_{s_j} + z_{d_j}\sigma_{s_j}). \tag{4.7}$$

Portanto, assumindo que todos os segmentos exibem o mesmo fator k_d, ou seja, $z_{d_j} = k_d$ para todo j, tem-se que:

$$D = \sum_{j=1}^{N} (\mu_{s_j} + k_d\sigma_{s_j}) = \sum_{j=1}^{N} \mu_{s_j} + k_d \sum_{j=1}^{N} \sigma_{s_j}. \tag{4.8}$$

Esta é a formulação original de CAMPBELL [58] em sua primeira proposta seguindo um modelo normal para a distribuição das durações. Assim, sendo conhecida a duração D da sílaba, o fator k_d de normalização da sílaba é obtido por:

$$k_d = \frac{D - \sum_{j=1}^{N} \mu_{s_j}}{\sum_{j=1}^{N} \sigma_{s_j}}. \tag{4.9}$$

[1] O índice d em μ_s^d e σ_s^d, que indicam ser distribuições referentes às durações, foi suprimido por simplicidade de representação na notação que segue.

Por outro ponto de vista, pode-se dar uma interpretação diferente ao substituirmos o valor de D na equação (4.9) pela equação (4.7):

$$k_d = \frac{\sum_j^N (\mu_{s_j} + z_{d_j}\sigma_{s_j}) - \sum_j^N \mu_{s_j}}{\sum_j^N \sigma_{s_j}} = \frac{\sum_j^N z_{d_j}\sigma_{s_j}}{\sum_j^N \sigma_{s_j}}, \tag{4.10}$$

o que indica que o parâmetro k_d é a média ponderada por σ_{s_j} dos fatores de normalização z_j dos segmentos.

Na dedução feita até o momento, foi suposta a distribuição normal para as durações. Porém, ao considerarmos a distribuição lognormal, a equação (4.9) para a obtenção de k_d não se aplica, pois a duração da sílaba passa a ser dada:

$$D = \sum_j^N d_{s_j} = \sum_j^N e^{(\mu_{s_j} + k_d\sigma_{s_j})} \tag{4.11}$$

Entretanto, k_d pode ser obtido por um procedimento numérico de minimização, já que a duração D da sílaba é conhecida a *priori*. CAMPBELL [61] descreve um algoritmo de aproximação, assumindo inicialmente $k_d = 0$ e então crescendo ou diminuindo em passos de $\pm 0{,}1$ até que o somatório atinja a duração da sílaba. Ao fim da recursão, o parâmetro k_d é retrocedido por uma constante de 0,075. Outro artifício usado pelo autor é associar à sílaba final um modelo diferenciado.

No *software* desenvolvido optou-se por uma implementação diferente. O valor de k_d é obtido por minimização do erro médio quadrático (LMS) entre a duração original D da sílaba e a duração determinada pelo modelo, dada pela equação (4.11), e o valor de k_d é corrigido recursivamente pelo erro de estimação.

Este modelo de elasticidade originalmente proposto por CAMPBELL [61] é considerado por alguns autores como uma hipótese teoricamente bem razoável, porém criticado em alguns pontos [59]. Por exemplo, no método de minimização para obtenção de k_d, o retrocesso executado ao fim da minimização é um indicativo de falha do modelo [59].

Na aplicação do modelo de CAMPBELL às frases do *corpus*, com prosódia neutra, o resultado é satisfatório. Porém, na aplicação do modelo à prosódia de certas atitudes, constatou-se que a duração dos segmentos se desviam de um comportamento uniforme. Assim, foi proposto neste trabalho um modelo de durações para se adequar à prosódia das atitudes, descrito a seguir.

Modificação do modelo de CAMPBELL

No estudo da prosódia característica de algumas atitudes, MORAES [5] observa que algumas atitudes são caracterizadas por um alongamento extra na duração da sílaba tônica, decorrente do alongamento da vogal, por exemplo, na atitude de Aviso. Por

outro lado, no estudo das atitudes de ênfase, por exemplo, MORAES [5] observa um comportamento de alongamento das consoantes.

Na figura 4.10 é mostrada a frase "Reanta jogava" produzida com prosódia neutra, sem expressividade, e na figura 4.11 é mostrada a mesma frase produzida com uma atitude de Aviso. Comparando os sinais, observa-se um alongamento acentuado da vogal [a] frente a consoante [+g], na sílaba [+ga].

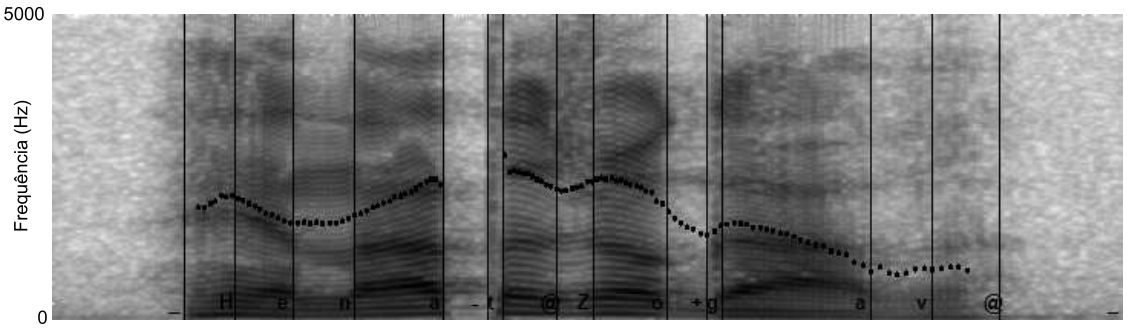


Figura 4.10: Prosódia neutra de referência.

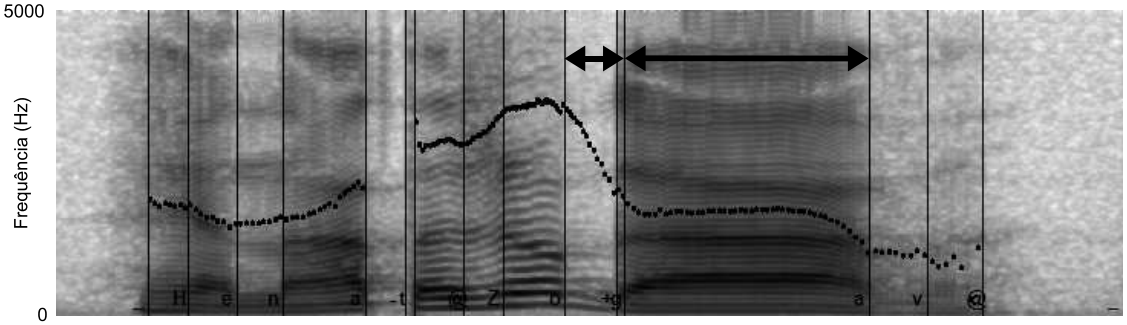


Figura 4.11: Alongamento da vogal [a] frente à consoante [+g], na sílaba [+ga], na atitude de Aviso.

Na figura 4.12 é mostrada a mesma frase com uma atitude de Ênfase. Comparando com a frase de prosódia neutra, observa-se um alongamento na consoante [Z] frente à vogal [o].

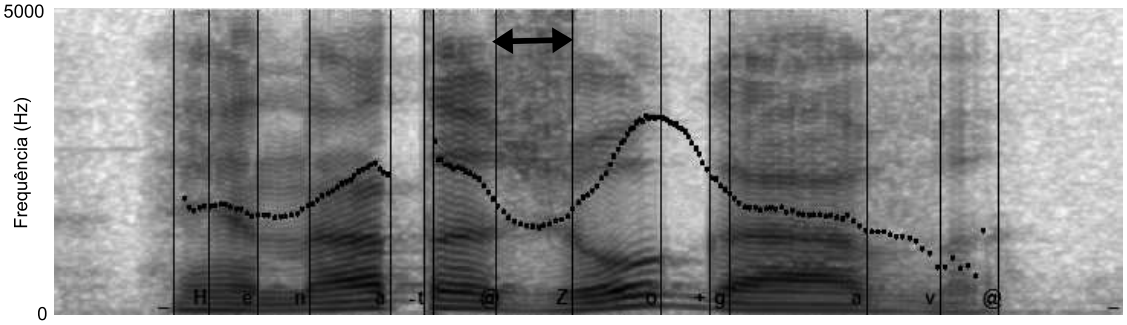


Figura 4.12: Alongamento da consoantes na sílabas [Zo].

Estes casos apresentados, onde ora ocorre o alongamento da vogal e ora o alongamento da consoante, nos leva a supor que estes segmentos possuem um comportamento independente dos demais ao invés de estarem totalmente correlacionados, como supõe o modelo de CAMPBELL [6].

Assim, propomos neste trabalho adotar dois parâmetros de normalização para a sílaba: um parâmetro para as consoantes em *onset*, nomeado de k_{do} e outro parâmetro para a normalização da vogal que ocupa o *núcleo* da sílaba, nomeado de k_{dn}. Deste modo, para as sílabas que possuem mais de um segmento na posição de *onset*, o parâmetro k_{do}, é obtido pela aplicação do modelo de CAMPBELL somente às consoantes que ocupam o *onset* da sílaba. O parâmetro k_{dn} é dado pelo fator de normalização da vogal do *núcleo*.

Para ilustrar o procedimento de análise das durações da sílaba, na figura 4.13 é mostrado um recorte da tela de manipulação prosódica do *software*, onde estão ilustradas as etapas do procedimento de análise. Os gráficos da figura representam diferentes níveis de análise, onde, de cima para baixo, têm-se respectivamente a representação no domínio do tempo, no nível dos segmentos e no nível das sílabas. As etapas do procedimento de análise decorrem tal como a ordem das figuras, ou seja, a partir das durações no domínio do tempo obtém-se a normalização dos segmentos e em seguida a normalização da sílaba.

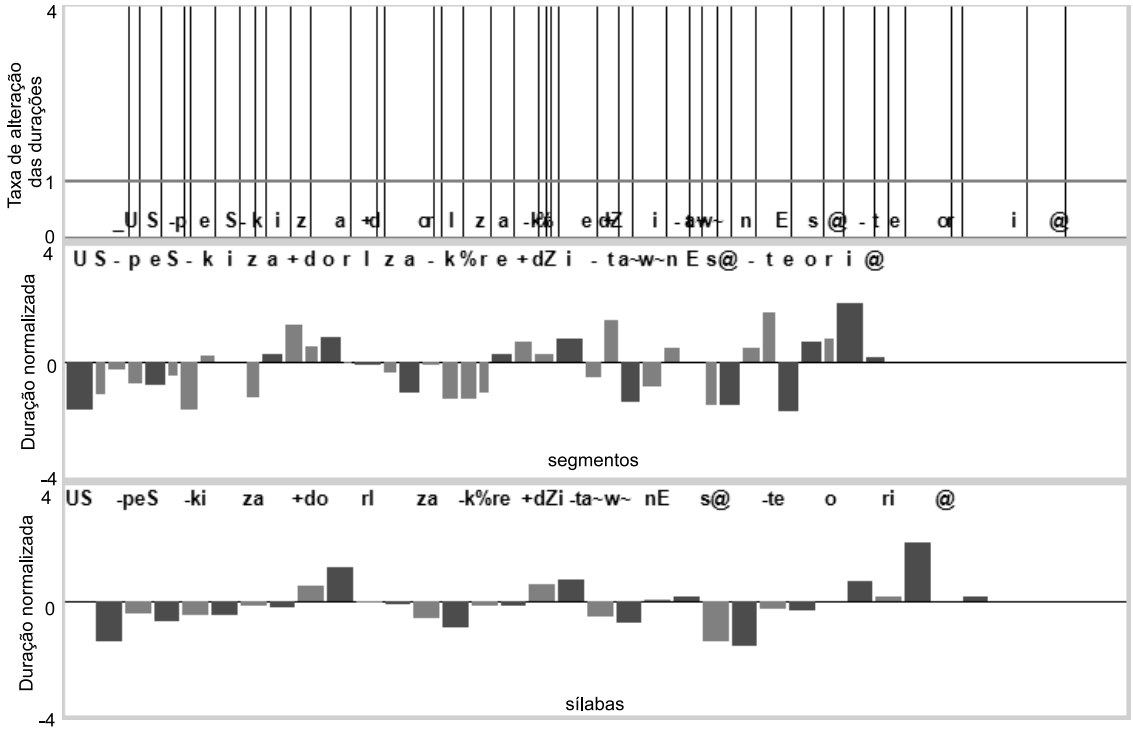


Figura 4.13: Exemplo de normalização das durações das sílaba.

No gráfico superior são representadas as etiquetas fonéticas como referência, e uma curva no domínio do tempo, que indica a alteração da duração do sinal original,

que será vista com maior detalhe na Seção 4.5.1. O gráfico do meio, no nível dos segmentos, é o mesmo mostrado na figura 4.6, onde estão representados os valores normalizados z_d das durações. No gráfico inferior, estão representados os parâmetros k_{do} e k_{dn} das sílabas, como resultado da análise da duração das sílabas.

Na seção a seguir será apresentado o procedimento proposto para a análise da intensidade da sílaba.

4.4.2 Intensidade

São poucos os trabalhos em modelagem da prosódia que dão atenção à intensidade, considerando-a como uma variável de menor relevância. CAMBPELL [6] observa que o modelo proposto para a duração da sílaba não se aplica à intensidade. De fato, a natureza da medida das durações é diferente da medida de intensidade. Enquanto a duração da sílaba é dada pela soma da duração dos segmentos, é falso supor que a intensidade da sílaba é a soma da intensidade dos segmentos.

Motivado pelo fato de que a vogal que ocupa o núcleo da sílaba é considerada como o pico perceptual da sílaba [31], propõe-se utilizar o fator de normalização da intensidade z_I da vogal como parâmetro k_I de intensidade da sílaba. Entretanto, resta saber como determinar a intensidade dos segmentos contidos pela sílaba a partir do parâmetro k_I. Contatou-se que a intensidade dos segmentos vizinhos à vogal exibem correlação com a intensidade da vogal. Então, esta correlação poderá ser usada para estimar a intensidade dos segmentos, o que será visto na etapa de síntese, na Seção 4.5.2.

Na figura 4.14, é ilustrado um exemplo de análise das intensidades, seguindo o mesmo critério utilizado na figura 4.13 para representar a análise das durações. No gráfico superior, a curva constante representa a alteração da intensidade do sinal ao longo do tempo. No gráfico do meio está representada a normalização da intensidade no nível dos segmentos e no gráfico inferior está representada a normalização da intensidade das sílabas. Comparando a figura do meio com a figura inferior, observa-se que os parâmetros das sílabas são iguais aos fatores de normalização mesmo das vogais do *núcleo* das sílabas, conforme o modelo adotado.

4.4.3 Pitch

A rigor, o termo pitch está associado à percepção da frequência fundamental (F0). No entanto, é comum no processamento da fala utilizar o termo pitch como sinônimo da frequência fundamental [3].

A curva do contorno de pitch é uma curva discreta no tempo, amostrada em intervalos não uniformes. Neste trabalho, a curva do contorno de pitch foi determinada a partir da indicação das marcas de pitch, obtidas do sinal do eletroglotógrafo.

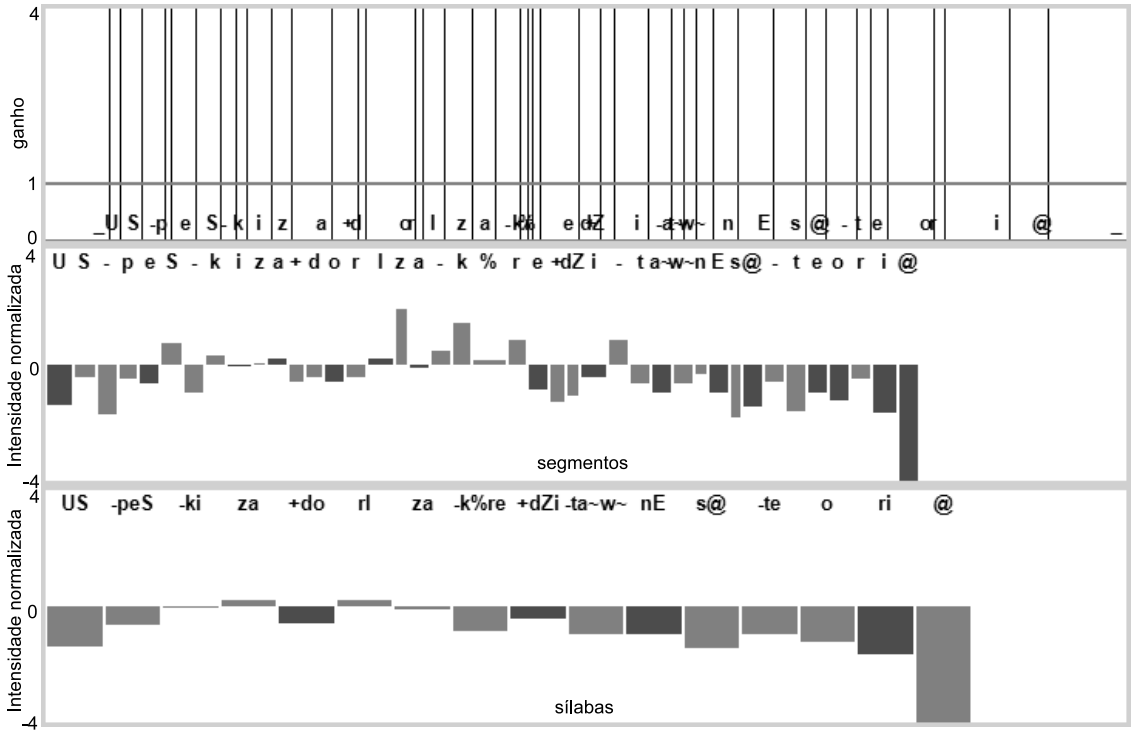


Figura 4.14: Exemplo de normalização da intensidade das sílabas.

Dadas duas marcas de pitch, localizadas nos instantes de tempo t_i e t_{i-1}, quando o intervalo entre as marcas é maior do que 20 ms o intervalo é considerado surdo, caso contrário, as marcas delimitam um período de pitch [60]. Assim, o pitch é obtido pelo inverso do período, $1/(t - t_{i-1})$, e é associado ao instante de tempo médio entre as duas marcas, dando origem aos pontos da curva do contorno de pitch.

Na vizinhança de intervalos surdos, a curva de pitch geralmente exibe pontos espúrios que ultrapassam a variação média ao longo da curva [62]. Por esta razão, a curva de pitch é submetida a um procedimento de suavização, de maneira que estes pontos são eliminados quando fora do intervalo de variação de 2% do pitch dos pontos vizinhos mais próximos.

O contorno de pitch pode ser interpretado como uma curva contínua, composta por uma sequência de eventos de subida, descida e patamares. Diferentes autores têm se dedicado à codificação do contorno de pitch, por um processo ora chamado de *stylization*, onde os eventos presentes na curva são codificados conforme um modelo. Algumas propostas consideram a priori a sincronia destes eventos com a estrutura da sentença, outros buscam estabelecer esta relação a posteriori. Dentre diversas propostas de modelagem e codificação do contorno de pitch, temos o modelo MOMEL [62] [63], aplicado ao português em [64], o modelo TILT [65], o modelo de Fujisaki [66], o *prosogram* [67], dentre outros.

Diante de tantas propostas, a escolha de um modelo para a modelagem da curva de contorno do pitch acaba por ser uma escolha conforme propriedades convenientes.

Por exemplo, na descrição do modelo que recebeu o próprio nome, FUJISAKI [66] indica que na análise de sentenças para o japonês, o exame do contorno de F0 de uma sentença, quando representado na escala logarítmica, sugere ser considerado como a superposição de curvas componentes, geradas por um conjunto de parâmetros. Por um procedimento de análise por síntese, é possível decompor um dado contorno de F0 nas curvas componentes e estimar os parâmetros das curvas.

Em adaptação do modelo de Fujisaki para o português europeu, TEIXEIRA [20] utiliza um procedimento empírico para encontrar os parâmetros das curvas componentes e em seguida utiliza uma rede neural para prever tais parâmetros a partir de informações obtidas na estrutura da sentença. Uma dificuldade do modelo de Fujisaki é obter uma solução única que víncule os parâmetros das curvas com a estrutura da sentença.

Um modelo interessante que agrega o conceito básico de superposição de curvas proposto por FUJISAKI, trata-se do modelo SFC (*superposicion of functional contours*), descrito como um modelo prosódico treinável [68]. Neste modelo, as curvas que se sobrepõem para formar o contorno de pitch são obtidas diretamente da estrutura linguística, sem nenhuma representação intermediária. Neste caso, o contorno de pitch da sílaba é representado por três pontos amostrados do contorno de pitch a 10%, 50% e 90% do intervalo do núcleo vocálico.

Nesta tese, adotou-se a proposta similar de observar o contorno de pitch limitado ao núcleo vocálico da sílaba, sem nenhuma representação intermediária. No entanto, na aplicação deste modelo à prosódia das atitudes, constatou-se que os três pontos são insuficientes para capturar certas nuances no contorno de pitch, que são importantes para caracterizar a atitude. Assim adotou-se amostrar o núcleo da sílaba em quatro pontos, obtidos em 12,5%, 37,5%, 62,5% e 87,5%.

Desta forma, os movimentos de subida e descida no contorno de pitch são observados nos limites da sílaba. Isto implica em assumir que tais eventos sejam, de algum modo, síncronos com a estrutura silábica da sentença. Além disso, ao reamostrar a curva do contorno de pitch em quatro pontos, tomados proporcionalmente à duração do núcleo vocálico da sílaba, implica que o contorno de pitch de vogais muito extensas é sub-amostrado, enquanto em vogais breves o contorno de pitch é super-amostrado. Consequentemente, a velocidade das variações no contorno de pitch acompanha a velocidade de emissão da vogal, ou seja, quando a sílaba é alongada, o intervalo entre amostras do contorno também é alongado, diminuindo a inclinação da curva e vice-versa.

Para tornar a análise do contorno de pitch menos dependente do pitch próprio do falante, os pontos amostrados do contorno de pitch foram normalizados, no domínio logarítmico. Conforme a distribuição dos valores de pitch ao longo de todas as sentenças do *corpus*, foi obtido um valor médio de 4,7 log(Hz) e desvio padrão de

0,14 log(Hz). Assim, os pontos amostrados do contorno de pitch, de valor p, são normalizados por:

$$k_p = \frac{\ln(p) - 4{,}7}{0{,}14} \tag{4.12}$$

Para manter coerência na notação, os quadro pontos após amostrados são nomeados por k_{p0}, k_{p1}, k_{p2} e k_{p3}.

Na figura 4.15 é dado um exemplo da análise do pitch. No gráfico superior é mostrado o contorno de pitch original. No gráfico inferior está representada a reamostragem do contorno de pitch, tomado somente no núcleo vocálico da sílaba, conforme o procedimento descrito acima. Os três eventos mais importantes de subida e descida estão indicados na figura, localizados nas sílabas tônicas da sentença.

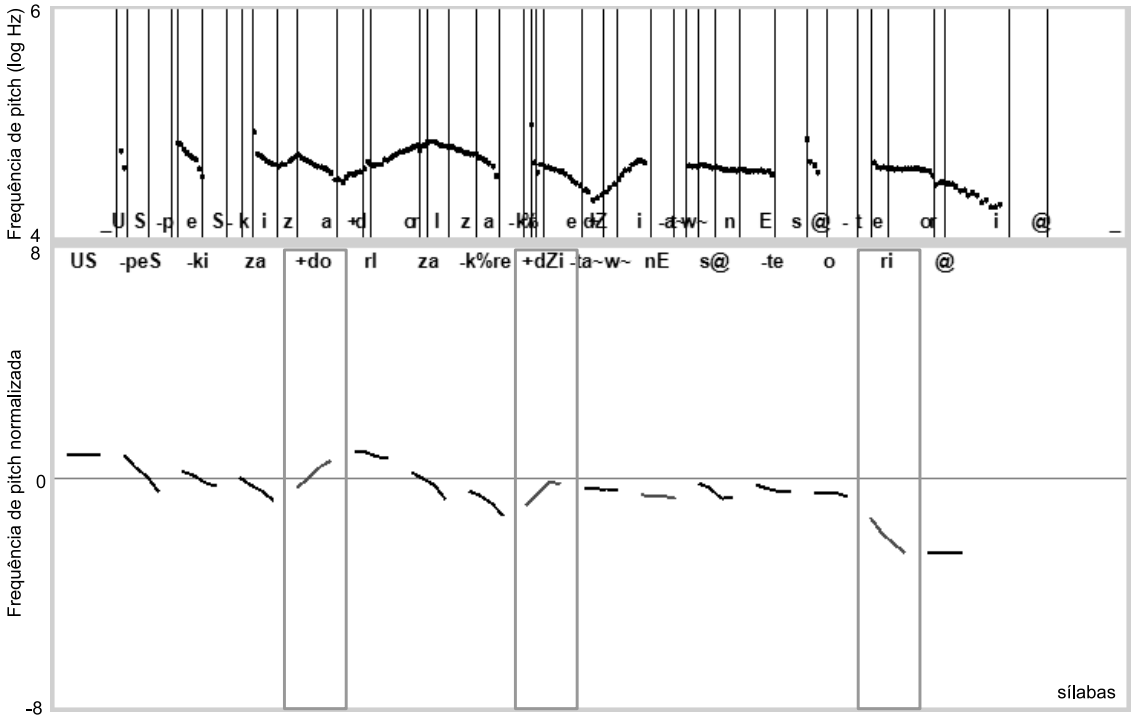


Figura 4.15: Exemplo da análise do contorno de pitch, normalizado pela distribuição do pitch ao longo do *corpus* e reamostrado por quatro pontos, no limite da vogal que ocupa o núcleo silábico.

Ao fim da análise da sílaba, têm-se portanto as variáveis prosódicas de duração, intensidade e pitch, codificadas por um vetor de parâmetros dado por:

$$\mathbf{k} = [k_{dn}, k_{do}, k_I, k_{p0}, k_{p1}, k_{p2}, k_{p3}].$$

Assim, a observação deste vetor de parâmetros ao longo das sílabas de uma sentença permite atingir o objetivo pretendido de observar o comportamento das variáveis prosódicas num nível mais abstrato.

Na seção a seguir será descrito como este vetor de parâmetros é decodificado para alterar a prosódia do sinal de fala.

4.5 Síntese da sílaba

Nesta seção será descrito o procedimento de síntese, que consiste em decodificar os parâmetros da sílaba para os valores nominais da duração e intensidade, a serem impostos aos segmentos, assim como o contorno de pitch a ser aplicado na sentença.

A etapa de síntese resulta em curvas no domínio do tempo que serão utilizadas pelo algoritmo PSOLA para transformar o sinal de uma sentença, conforme os parâmetros da sílabas. Como requisito, as fronteiras entre os segmentos desta sentença precisam ser conhecidas, e também a organização destes segmentos em sílabas.

Nas Seções 4.5.1 e 4.5.2 será descrita a síntese das durações e intensidades. Em seguida, na Seção 4.5.3, será vista a reconstrução da curva do contorno de pitch.

4.5.1 Durações

Na etapa de síntese, a duração nominal dos segmentos é obtida a partir dos parâmetros k_{dn} e k_{do} da sílaba, conforme definidos na seção 4.4.1.

Conforme pressupõe o modelo, o parâmetro k_{dn} da sílaba é repassado a todos os segmentos que pertencem ao *onset* da sílaba, ou seja, $z_{d_j} = k_d$. Por sua vez, a vogal que ocupa o *núcleo* recebe o valor de k_{do}. O fator de normalização da duração dos segmentos que ocupam a *coda* da sílaba são determinados por $z_{d_j} = \rho_j k_d$, onde ρ_j indica a correlação da duração do segmento j com a duração da vogal.

Para ilustrar o procedimento de síntese, na figura 4.16 é mostrada a execução da síntese logo após a análise da sentença ter sido executada. Tal como representado na etapa de análise, os gráficos da figura representam diferentes níveis. Entretanto, o procedimento de síntese ocorre no sentido inverso da análise, isto é, de baixo para cima, do nível da sílaba para o domínio do tempo.

No gráfico inferior da figura 4.16 estão representados os parâmetros das sílaba obtidos após a análise, tal como mostrado na figura 4.13. No gráfico do meio estão representados os fatores de normalização da duração dos segmentos, obtidos dos parâmetros da sílaba. No gráfico superior, está representada a curva temporal referente à alteração da duração original do segmento para a duração determinada.

A curva de alteração das durações é melhor representada pela curva de variação da duração ao longo do tempo [60]. Assim, a duração desejada em um determinado instante de tempo t é calculada pela área sob a curva desde 0 até o instante t. Assim, a duração de um segmento é determinada pela área da curva limitada pelas fronteiras do segmento. Por exemplo, se não há alteração na duração original, a curva é constante e igual a 1. Para se alterar a duração de um segmento a partir da duração original d_1, para uma nova duração d_2, a curva de alteração da duração nos limites do segmento pode ser constante e igual à razão $\frac{d_2}{d_1}$. Se considerarmos as durações normalizadas por z_{d_1} e z_{d_2}, a constante para alteração do segmento, seria

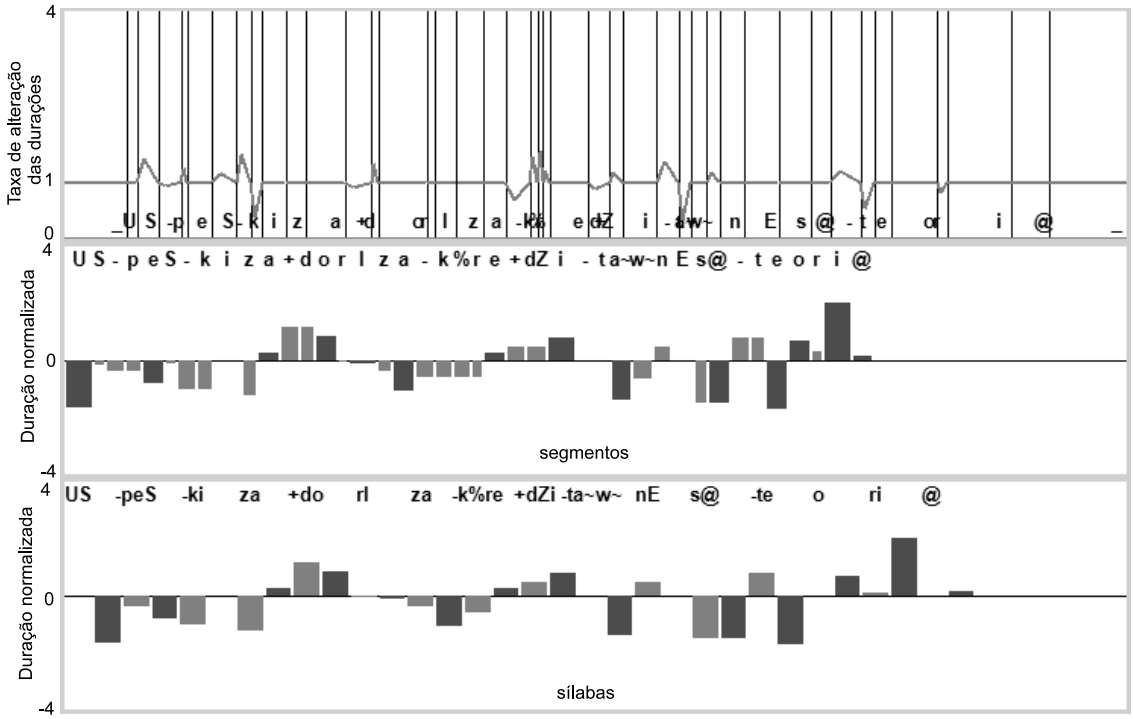


Figura 4.16: Exemplo do procedimento de síntese das durações a partir dos parâmetros de duração da sílaba.

dada por $\frac{d(z_{d_1},s)}{d(z_{d_2},s)}$. Ao substituir as durações pela equação inversa da normalização, tem-se:

$$\frac{d(z_{d_2},s)}{d(z_{d_1},s)} = \frac{e^{(\mu_d^s + z_{d_2}\sigma_d^s)}}{e^{(\mu_d^s + z_{d_1}\sigma_d^s)}} = e^{(z_{d_2} - z_{d_1})\sigma_d^s}. \tag{4.13}$$

Fazendo $z_{d_2} - z_{d_1} = \Delta_z$, que indica a alteração no fator de normalização, temos que o valor da curva para a alteração do segmento será uma constante dada por $e^{\Delta_z \sigma_d^s}$.

Entretanto, observou-se que em segmentos longos é preferível alterar o segmento de maneira não uniforme ao invés de constante, atenuando a alteração nas fronteiras e privilegiando a região inicial. Isto foi feito impondo uma alteração triangular nos limites do segmento, de modo que a área limitada pelo triângulo permanece a mesma do que com um valor constante. Deste modo, o valor do ponto da curva equivalente ao vértice do triângulo é dado por:

$$\frac{2d(z_{d_2},s)}{d(z_{d_1},s)} - 1 = 2e^{\Delta_z \sigma_d^s} - 1. \tag{4.14}$$

Aplicando-se a relação da equação (4.14) para todos os segmentos da sentença, temos a curva representada no gráfico superior da figura 4.16.

4.5.2 Intensidade

Na análise das intensidades, na Seção 4.4.2 foi proposto tomar o parâmetro de intensidade da sílaba como o fator de normalização da intensidade da vogal que ocupa o *núcleo* da sílaba. No entanto, ao impor a todos os segmentos o parâmetro de intensidade da sílaba, a síntese do sinal mostrou-se desastrosa, pois os segmentos são demasiadamente amplificados ou atenuados. Constata-se que o desvio padrão dos segmentos é insuficiente para conter a alteração da intensidade dos segmentos frente à alteração da vogal.

Se considerarmos que a intensidade I_j de cada segmento j, interno à sílaba, possui uma distribuição binormal com a intensidade da vogal I_v, então, dado o fator de normalização da intensidade da vogal z_I como parâmetro conhecido, a melhor estimativa para a intensidade dos segmentos vizinhos será dada por:

$$E[I_j|I_v] = \mu_{I_j} + \rho_j(\frac{I_v - \mu_{I_v}}{\sigma_{I_v}})\sigma_{I_j} = \mu_{I_j} + \rho_j k_I \sigma_{I_j} \tag{4.15}$$

onde ρ_j é a correlação entre a intensidade do segmento j e a intensidade da vogal. Conforme a equação de normalização dos segmentos, basta fazermos o fator de normalização dos segmentos $z_I = \rho_{sv} k_I$.

Os valores de ρ_{I_s} são apresentados na tabela C.1 no Apêndice C, juntamente com as distribuições. Foi considerada uma significância na correlação de $p > 0{,}05$, caso contrário é assumido que não há correlação, $\rho_{I_s} = 0$ e portanto a intensidade do segmento é inalterada, sendo dada somente pela média.

Na figura 4.17 é dado um exemplo, que tal como as durações, no gráfico inferior estão representados os parâmetros de intensidade k_I da sílaba, obtidos após a análise, tal como na figura 4.14. No gráfico do meio, estão representados os fatores de normalização da intensidade dos segmentos, obtidos a partir do parâmetro de intensidade da sílaba pela relação $z_{I_j} = \rho_s k_I$, para todo j pertencente à sílaba. E por fim, na figura superior, está representada a curva temporal utilizada para alteração das intensidades do sinal.

A curva de alteração da intensidade é dada por uma curva de ganho, onde a amplitude do sinal original é multiplicada pela curva. Considere um segmento de intensidade I_1 que se deseja alterar para a intensidade I_2. Este aumento na intensidade, implica na alteração da energia de e_1 para e_2. Portanto a curva de ganho, dentre as fronteiras do segmento, terá um valor constante dado por $\sqrt{\frac{e_2}{e_1}}$. Substituindo o valor da energia pelo inverso da equação de intensidade dado na equação (4.4), tem-se que:

$$\sqrt{\frac{e_2}{e_1}} = \sqrt{\frac{10^{0{,}1(I_2-94)}}{10^{0{,}1(I_1-94)}}} = 10^{0{,}05(I_2-I_1)} \tag{4.16}$$

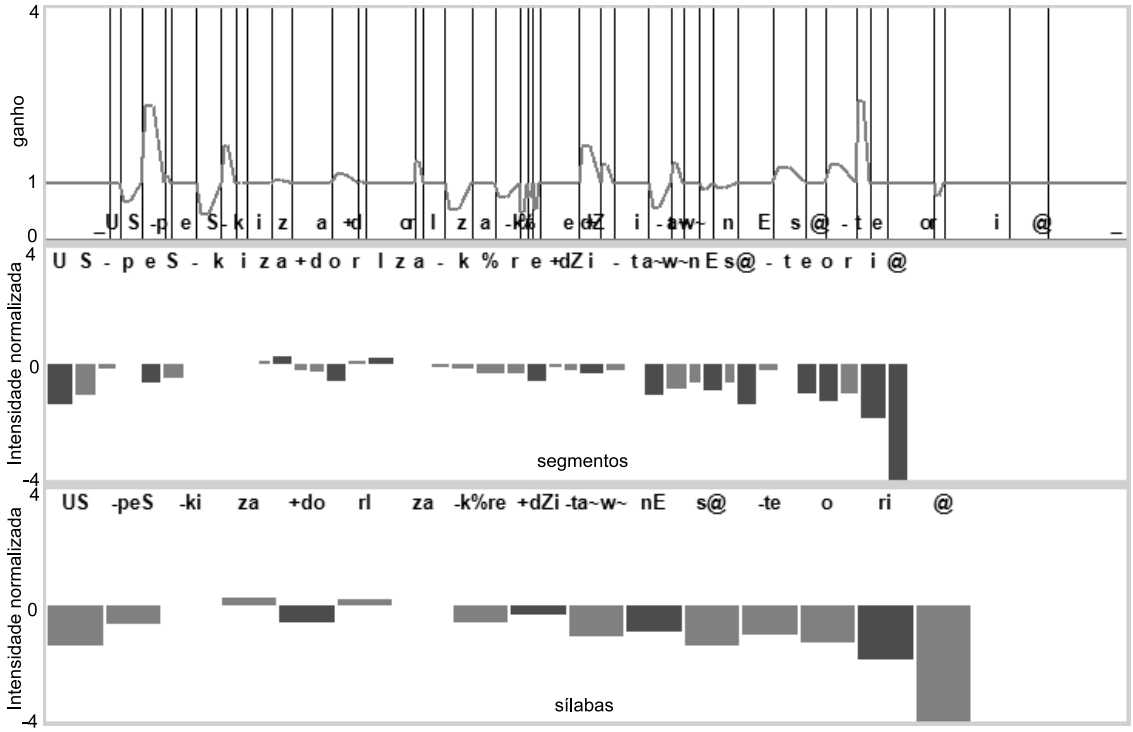


Figura 4.17: Exemplo do procedimento de síntese da intensidade a partir dos parêmtros da sílaba.

Sendo a intensidade normalizada dada pela equação (4.6), isto é, $I(z_I, s) = \mu_I^s + z_I\sigma_I^s$, tem-se que o valor do ganho em função dos fatores de normalização será dado por:

$$10^{0{,}05[(\mu_I^s + z_{I_2}\sigma_I^s) - (\mu_I^s + z_{I_1}\sigma_I^s)]} = 10^{0{,}05(z_{I_2} - z_{I_1})\sigma_I^s}. \tag{4.17}$$

fazendo $z_{I_2} - z_{I_1} = \Delta_{z_I}$, temos que o valor do ganho, entre as fronteiras do segmento, é dado por $10^{0{,}05\Delta_{z_I}\sigma_I^s}$.

Retornando ao caso onde a intensidade do segmento é normalizada por z_{I_j} e passa a ser normalizada por $\rho_{I_s}k_I$, tem-se que $\Delta_{z_I} = \rho_{I_s}k_I - z_{I_j}$. Aplicando-se a relação da equação 4.17 a todos os segmentos da sentença, temos a curva de ganho representada no gráfico superior da figura 4.17. Nota-se que manter o valor constante ao longo de todo o segmento resulta em alterações bruscas na amplitude. Deste modo, a constante de ganho de cada segmento foi atenuada nas fronteiras entre os segmentos.

4.5.3 Pitch

Na síntese, o contorno completo da sentença é gerado pela interpolação dos quatro pontos amostrados do contorno de pitch: k_{p0}, k_{p1}, k_{p2} e k_{p3}. A interpolação linear dos pontos resulta em um contorno demasiadamente simplificado por segmentos de reta, principalmente no caso em que há um grande número de consoantes entre nú-

cleos vocálicos, como por exemplo na palavra “transcrição” [tr**ã**S.kr**i**.sãw̃]. Assim, o contorno de pitch foi interpolado por curvas *splines* cúbicas, impondo que a primeira e segunda derivadas da curva sejam contínuas.

Na figura 4.18, é mostrado um exemplo da síntese do contorno de pitch para a frase “Os pesquisadores acreditam nessa teoria”. Na figura inferior é ilustrado o resultado da análise, tal como na figura 4.15. No gráfico superior, são mostrados os pontos do contorno de pitch original, sobrepostos por uma linha contínua que representa o contorno de pitch de síntese após a interpolação dos parâmetros de pitch da sílaba pela *spline* cúbica.

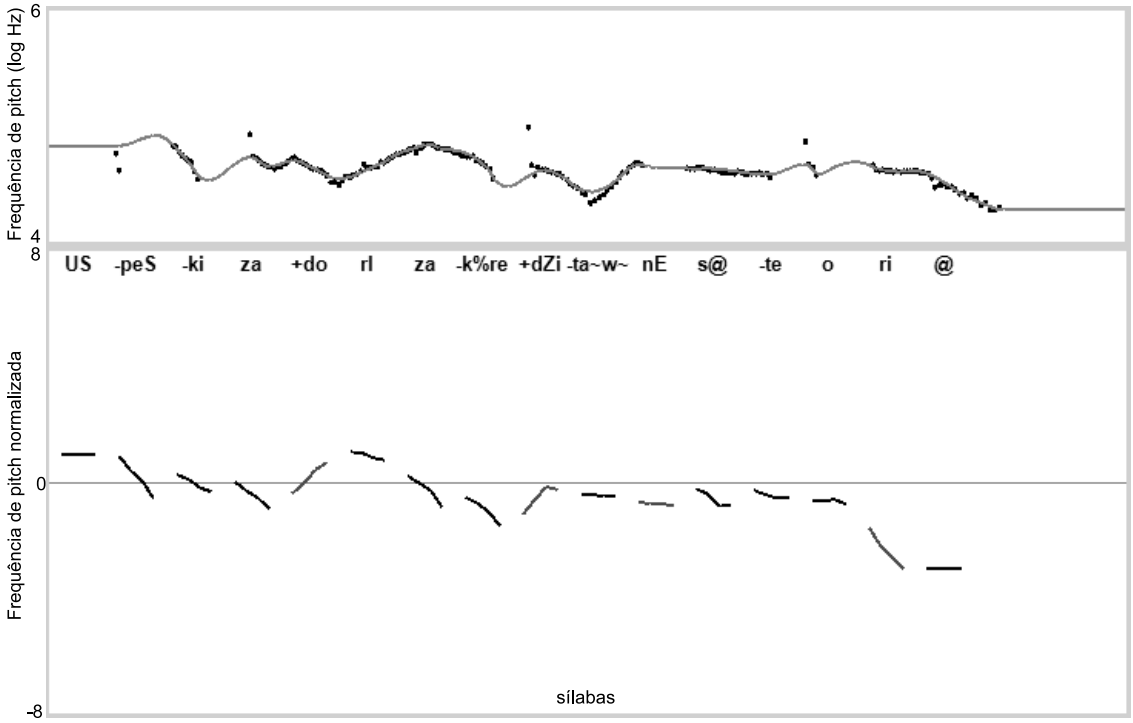


Figura 4.18: Exemplo da síntese da curva do contorno de pitch.

4.6 Resultados Parciais

Idealmente, o modelo adotado para a codificação das sílabas tem por hipótese que se a estrutura silábica de um conjunto de frases é similar, então o vetor de parâmetros **k** das sílabas depende somente da organização destas sílabas na frase, e ainda, que estes parâmetros são independentes do conteúdo segmental das sílabas.

Assim, para verificar esta hipótese, foram gravadas 25 frases, descritas na tabela 4.1, todas com estrutura silábica similar, isto é, todas com 6 sílabas, com as sílabas tônicas na segunda e quinta posições. Além disso, as sílabas das palavras foram escolhidas para conter diferentes conteúdos.

Tabela 4.1: Frases de teste.

renata jogava	lembrava da marta	joão caminhava
cristina lembrava	renata mostrava	que leite amargo
renata queria	estava cansada	gostava de sopa
renata saía	que torta gostosa	que horas começa
renata concorda	que lago bonito	Antônio demonstra
renata suporta	renata não gosta	Joana trazia
repete de novo	renata maltrata	Luísa teimava
e como jogava	será que aguenta	Amanda chorava
		Pedrinho calou-se

Quando observada a distribuição dos parâmetros das sílabas que ocupam a mesma posição, ao longo das diferentes sentenças, um pequeno desvio da média pode indicar o quanto este conjunto de sentenças se aproximam da hipótese assumida. Porém, este desvio da média reflete ainda variações aleatórias tais como alteração no ritmo, e a imprecisão na etiquetagem. Por esta razão, a mesma análise foi feita para um conjunto de dez repetições da frase "Renata jogava".

Nos tópicos seguintes, serão comparadas as distribuições dos parâmetros de duração, intensidade e pitch de cada sílaba, para o conjunto de repetições e o conjunto de frases de teste.

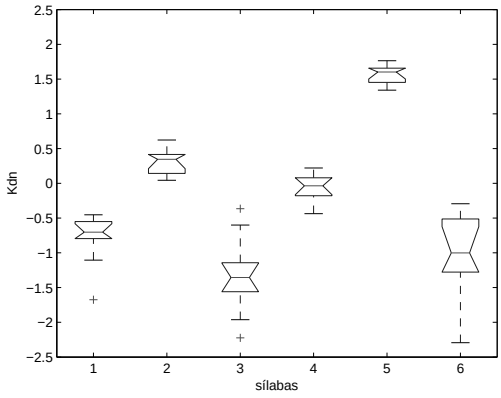
Durações

Na figura 4.19a é mostrada a distribuição do parâmetro k_{dn}, conforme a posição, para o conjunto de repetições. Nota-se que a segunda e a quinta sílabas tônicas são as que apresentam maior média, enquanto as sílabas que seguem a tônica apresentam as menores médias. O desvio padrão médio tomado ao longo das diferentes posições é de 0,34, que representa as flutuações aleatórias da etiquetagem e das diferentes gravações.

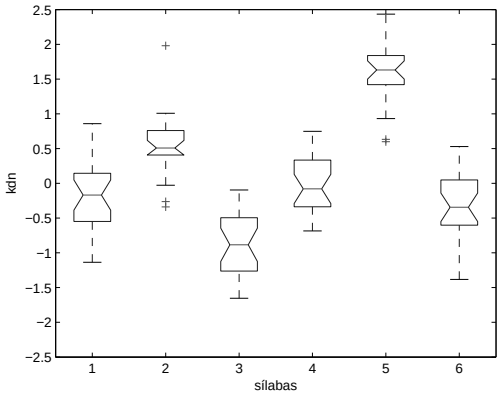
Na figura 4.19b é mostrada a distribuição do parâmetro k_{dn} para o conjunto de frases de teste, de conteúdos diferentes. Nota-se um padrão similar no comportamento da média, embora um desvio padrão médio maior, de 0,45, devido principalmente à diferença de conteúdo das frases.

Na figura 4.20a é mostrada a distribuição do parâmetro k_{do} para o conjunto de repetições, onde obteve-se um desvio padrão médio de 0,38. Neste caso o desvio médio é próximo ao desvio encontrado para k_{dn}. No entanto, para as frases de teste, obteve-se um desvio médio de 0,59, superior ao obtido em k_{dn}.

Para verificar o benefício do procedimento de agrupamento e da inclusão das distribuições condicionais, a normalização foi recalculada considerando somente a distribuição individual de cada segmento. Neste caso, o valor do desvio padrão médio encontrado anteriormente para k_{dn} e k_{do} aumenta para 0,53 e 0,60, respectivamente.

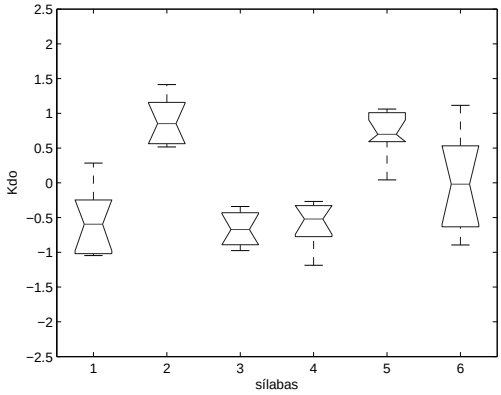


(a) Repetições.

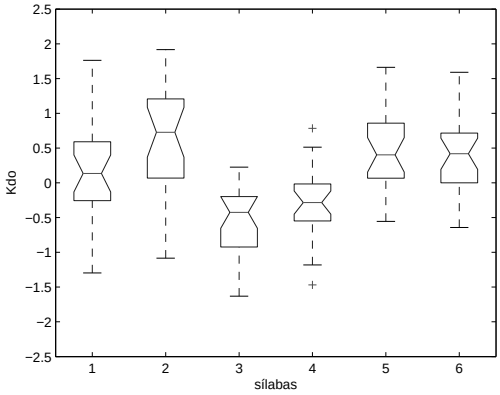


(b) Frases de teste.

Figura 4.19: Distribuição do parâmetro k_{dn} para um conjunto de frases.



(a) Repetições.



(b) Frases de teste.

Figura 4.20: Distribuição do parâmetro k_{do} para um conjunto de frases.

Isto justifica o uso dos agrupamentos, principalmente no caso das vogais.

Intensidades

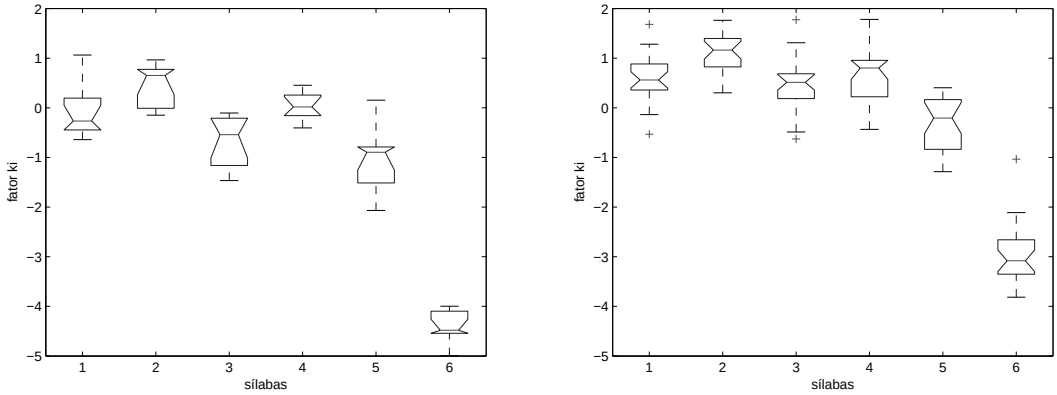
Tal como as durações, deseja-se verificar a hipótese do parâmetro k_I de intensidade ser independente do conteúdo das sílabas. Do mesmo modo, foi tomado o desvio da média de k_I ao longo das repetições e das frases de teste.

Na figura 4.21a é mostrada a distribuição do parâmetro k_I ao longo das sílabas das frases repetidas. O valor médio de k_I ao longo das sílabas apresenta um aumento na intensidade da primeira sílaba tônica em relação às sílabas vizinhas e um decaimento iniciando na segunda tônica e chegando à última sílaba com intensidade bem reduzida.

O desvio em k_I de 0,45, encontrado nas repetições, representa principalmente as flutuações das condições de gravação, tais como a proximidade do microfone e níveis

de ruído.

Nas frases de teste verifica-se o mesmo padrão na média observado nas frases repetidas, como mostra a figura 4.21b. Porém, ocorre um aumento geral da média e um desvio padrão médio de 0,52.



(a) Repetições. (b) Frases de teste.

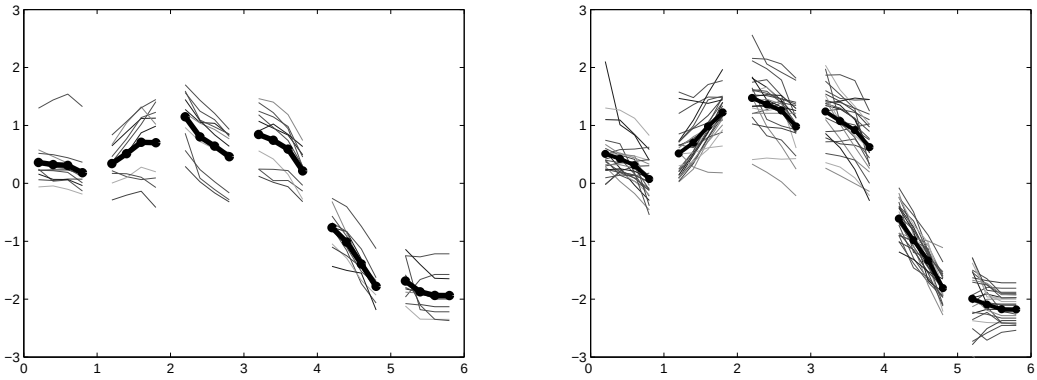
Figura 4.21: Distribuição do parâmetro k_I de intensidade das sílabas para um conjunto de frases.

Contorno de pitch

O contorno de pitch da sentença tem sido tradicionalmente tratado com certa independência do conteúdo. Para verificar também esta hipótese, na figura 4.22 foram representados os quatro pontos que caracterizam o contorno de pitch normalizado da sílaba, para cada sílaba das frases de mesmo conteúdo e das frases de teste.

As flutuações mostradas no contorno de pitch da figura 4.22a demonstram a imprecisão em se manter constante um mesmo formato do contorno de pitch em gravações executadas em momentos distintos. Todavia, observando os valores médios do contorno, nota-se um padrão médio característico de uma sentença de prosódia neutra. Observa-se um movimento de subida na primeira sílaba tônica, seguido de um decaimento que termina na segunda sílaba tônica. Na figura 4.22b referente às frases de teste observa-se o mesmo padrão característico, o que confirma a hipótese de certa independência do contorno de pitch com o conteúdo. Dentre as frases de teste foi encontrado um desvio padrão médio de 0,38.

Os resultados de média e desvio padrão nos parâmetros das sílabas das sentenças de teste fornecem um padrão para uma prosódia neutra e ainda uma medida do quanto estes parâmetros se desviam da média. No Capítulo 6, estes resultados serão usados no estudo da prosódia das atitudes.



(a) Repetições.

(b) Frases de teste.

Figura 4.22: Gráfico do contorno de pitch da sílaba para um conjunto de frases.

4.7 Conclusão

Neste capítulo, com o objetivo de observar a evolução dos parâmetros prosódicos tomando a sílaba como uma unidade mínima e independente de seu conteúdo, foi primeiramente estabelecido um procedimento de normalização das variáveis de duração e intensidade dos segmentos da sílaba.

Para a obtenção da distribuição das intensidades e durações dos segmentos foi construído um *corpus* de análise de 200 frases, no qual foi descrito o critério de etiquetagem dos segmentos, devido à influencia direta na obtenção das distribuições das variáveis de duração e intensidade. No procedimento de etiquetagem foi imposta a associação direta das etiquetas com a transcrição automática das sentenças permitindo vincular as diferentes estruturas linguísticas do texto com o sinal.

Na determinação da distribuição das durações, foram incluídas distribuições condicionais e foi feito o agrupamento dos segmentos em classes mais amplas, para minimizar os efeitos contextuais e suprir a falta de dados em certos contextos. Este procedimento foi feito para alguns casos particulares, ora indicados pela literatura, ora conforme sugerido pelos próprios dados. Os casos considerados não cobrem os vários efeitos contextuais que podem afetar a distribuição das durações, entretanto, o benefício de considerá-los foi demonstrado quantitativamente.

Após definidos os meios de normalização das variáveis de duração e intensidade dos segmentos, foram apresentados os conceitos envolvidos na análise e síntese dos parâmetros prosódicos da sílaba. O modelo proposto para as durações é uma adaptação do modelo proposto por CAMPBELL [6], considerando fatores observados ao longo de certos padrões de atitudes. Neste caso, a duração da sílaba passou a ser representada por um parâmetro relativo ao *onset* e outro relativo ao *núcleo* da sílaba, ao invés de um parâmetro único.

Para o caso das intensidades, a intensidade da vogal foi tomada como representante da intensidade da sílaba e a intensidade dos segmentos foi obtida pela correlação da intensidade dos segmentos com a intensidade da vogal.

O contorno de pitch foi normalizado pelo pitch característico do falante e reamostrado por quatro pontos, tomados nos limites da vogal que ocupa o núcleo da sílaba.

Por fim, foi apresentada a análise quantitativa dos parâmetros prosódicos obtidos de um conjunto de frases de diferentes conteúdos, mas de mesma organização silábica. Neste ponto, foram observados valores médios e desvios médios nos parâmetros das sílabas, onde foi possível observar um padrão característico dos parâmetros prosódicos que definem uma prosódia neutra.

No Capítulo 5 o procedimento de análise e síntese apresentados neste capítulo serão usados para o estudo das atitudes e no transplante de atitudes por meio da transferência dos parâmetros das sílabas.

Capítulo 5

Transplante paramétrico de prosódia

5.1 Introdução

No Capítulo 3 foi viabilizado o transplante de padrões prosódicos de diferentes atitudes para uma sentença de prosódia neutra, baseado no alinhamento temporal dos sinais. Porém, foi visto que o transplante por alinhamento temporal está limitado a sentenças de mesmo conteúdo segmental.

No Capítulo 4 buscou-se parametrizar as variáveis prosódicas no domínio da sílaba, de maneira a permitir a independência do conteúdo da sílaba e ainda permitir observar a evolução das variáveis prosódicas de forma paramétrica.

O objetivo deste capítulo é apresentar o transplante de forma paramétrica, permitindo que o transplante de prosódia seja feito no domínio da sílaba, entre sentenças de conteúdos diferentes. Na Seção 5.2 será apresentado o trasnplante por cópia dos parâmetros entre sílabas. Na Seção 5.3 será apresentada uma outra forma de transplante de forma paramétrica, nomeado de transplante por superposição. Nesta mesma seção será apresentada uma forma de representação de modelos que transformam o padrão de uma prosódia neutra no padrão típico das atitudes.

No artigo que motivou este trabalho, MORAES [5] apresenta, por meio de uma avaliação perceptiva, quais características presentes nas variáveis de duração, pitch e intensidade são determinantes para o reconhecimento de um conjunto de atitudes. Na Seção 5.4 é proposto observar as caracteristicas encontradas por MORAES [5] a partir pos parâmetros dos modelos das atitudes.

Na Seção 5.5 será apresentada a possibilidade da transformação entre atitudes, onde será ainda apresentada a semelhança entre as atitudes.

5.2 Transplante de prosódia por cópia

A normalização das variáveis prosódicas permitiu a abstração do conteúdo segmental das sílabas, permitindo que o transplante seja executado entre sílabas de diferentes conteúdos. Porém, os parâmetros prosódicos das sílabas dependem da função e da posição das sílabas na sentença. Isto implica que o transplante no domínio da sílaba, apesar de permitir a independência do conteúdo, está restrito à sentenças que possuem a mesma "estrutura silábica".

O termo "estrutura sílabica", usado neste trabalho, refere-se à organização das sílabas na sentença, ou seja, ao posicionamento das sílabas tônicas, à posição e distância das sílabas em relação à tônica, e ainda, à posição ocupada dentro das palavras e/ou grupos.

Ao tomarmos a sílaba como unidade básica, o transplante de prosódia é feito simplesmente pela cópia dos parâmetros prosódicos entre as sílabas correspondentes.

Na figura 5.1 é mostrado o esquema completo de execução do transplante de prosódia por cópia. No ramo à esquerda da figura estão representadas as etapas do processo da sentença de referência, enquanto que no ramo à direita estão representadas as etapas do processo da sentença de teste.

Nos blocos, maiores nomeados de "Análise", o sinal de fala, as marcas de pitch e as etiquetas de fronteiras dos segmentos, de ambas as sentenças são lidos do disco. Na etapa de processamento de texto a estrutura hierárquica de cada sentença é montada, conforme descrito no Capítulo 2. Em seguida, na etapa de análise das sílabas, conforme descrito no Capítulo 4, os parâmetros prosódicos das sílabas de ambas as sentenças são determinados.

Ao fim destas etapas, a estrutura hierárquica das sentenças contém toda a informação da estrutura silábica das sentenças, assim como os parâmetros prosódicos de cada sílaba. Então, na etapa de cópia dos parâmetros prosódicos, os parâmetros das sílabas da sentença de teste são copiados para as sílabas da referência.

Por fim, o procedimento de síntese das sílabas da sentença de referência, tal como descrito no Capítulo 4, gera as curvas de alteração das durações, do pitch e da intensidade, que serão usadas pelo PSOLA para executar a modificação do sinal da sentença de referência.

Na figura 5.2 é mostrado um exemplo da aplicação do transplante paramétrico da sentença "Renata jogava", usada como teste, para a sentença "Cristina lembrava", usada como referência. Neste caso, a sentença "Renata jogava" foi produzida com uma atitude de Aviso, que é caracterizada principalmente pelo formato típico no contorno de pitch e pelo alongamento da duração da vogal da sílaba tônica final. Ao comparar os gráficos 5.2b e 5.2c, verifica-se que o alongamento da vogal e o formato do contorno de pitch da sílaba tônica final foram devidamente transplantados para o

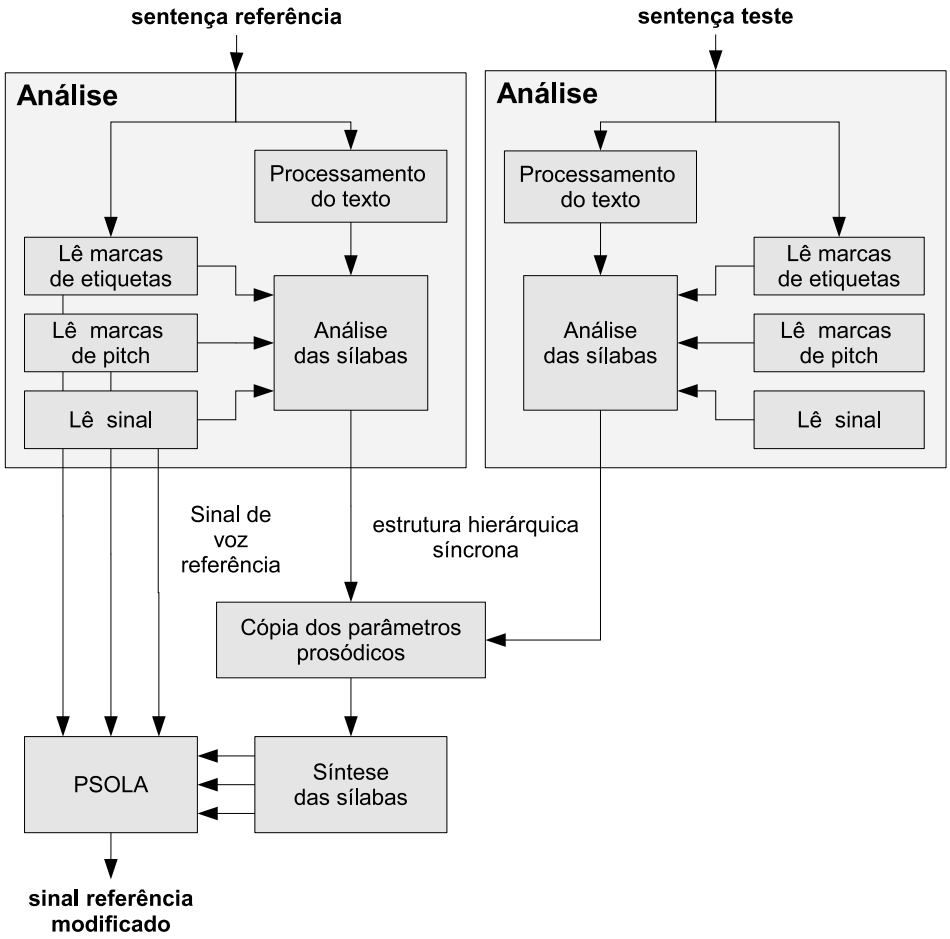


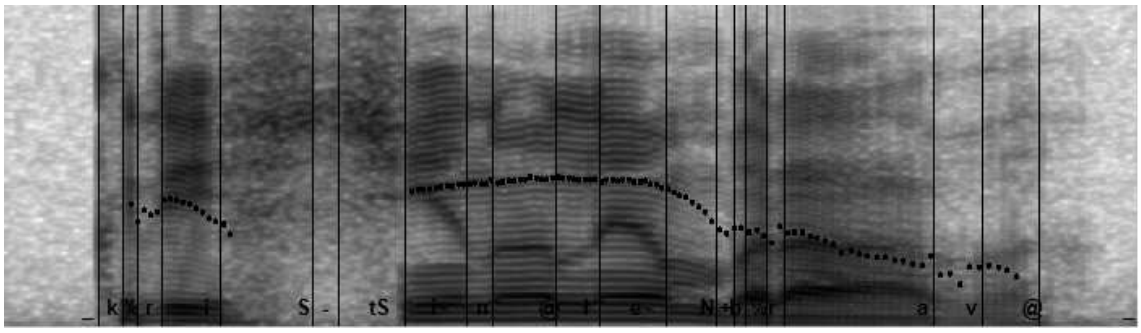
Figura 5.1: Esquema do transplante paramétrico de prosódia por cópia.

sinal de referência. Nota-se que não há sincronia entre as segmentos, pois as sílabas possuem formações diferentes.

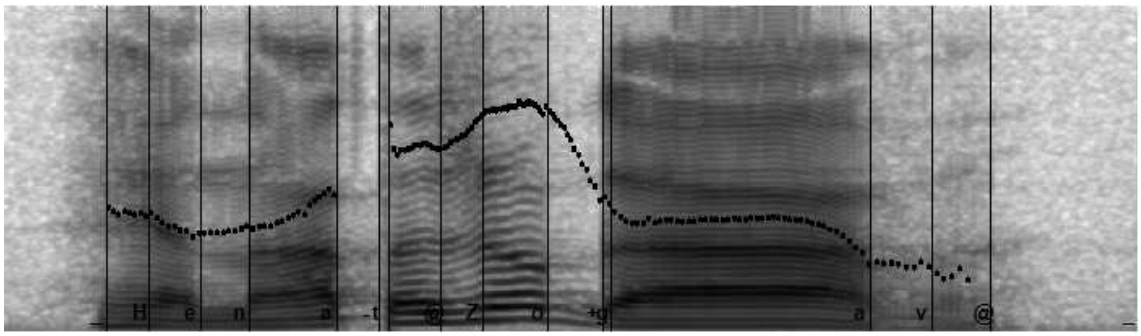
É desejável que no transplante por cópia as sílabas da sentença de teste possuam pelo menos uma consoante em *onset*, pois caso contrário o parâmetro de duração do *onset* será igual a zero, $k_{do} = 0$. Quanto isto ocorre, no momento do transplante a sílaba da refêrencia recebe $k_{do} = 0$ e, portanto, se houver segmentos em *onset*, a duração destes segmentos será sintetizada com a duração média característica.

Alternativamente, o sinal da sentença de referência pode ser obtida pelo sintetizador de fala e todo o procedimento é executado da mesma maneira.

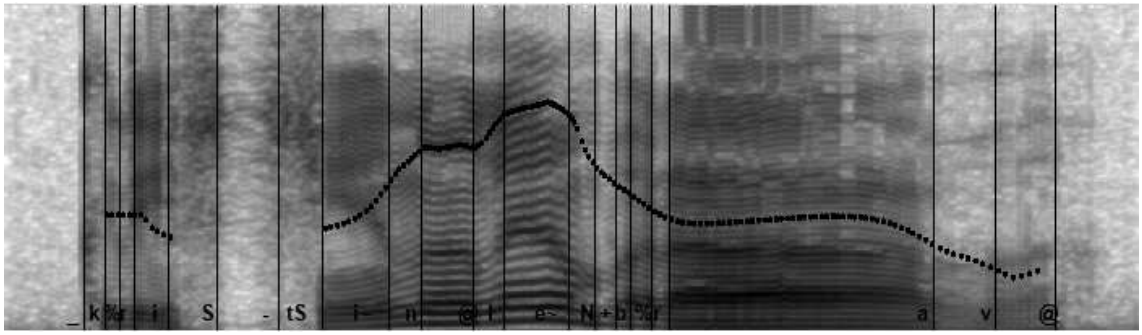
No transplante dos parâmetros por cópia, a hipótese de independência entre os parâmetros prosódicos e o conteúdo da sílaba é assumida totalmente quando os parâmetros de uma sílaba são copiados para outra, não importanto o conteúdo segmental. Esta hipótese é uma aproximação bastante conveniente, como verificado na Seção 4.6. Porém, podemos supor uma forma de transplante que considere certa dependência com o conteúdo como será visto a seguir no transplante por superposição.



(a) Sinal de referência da sentença “Cristina lembrava”.



(b) Sinal de teste da sentença “Renata jogava”, produzida com uma atitude de Aviso.



(c) Sinal resultante do transplante paramétrico de (b) para (a).

Figura 5.2: Exemplo do transplante paramétrico por cópia da prosódia da sentença “Renata jogava”, em um atitude de Aviso, para a sentença “Cristina lembrava”.

5.3 Transplante de prosódia por superposição

Alguns modelos prosódicos, baseados no conceito do modelo de Fujisaki [66], pressupõem que a curva do contorno de pitch pode ser construída pela superposição de curvas. O modelo SFC [68] é um exemplo que estende este conceito também às durações.

Se considerarmos que o padrão da prosódia neutra de uma sentença está intrinsicamente ligado à estrutura silábica desta sentença, podemos supor por hipótese que duas sentenças de mesma estrutura silábica dividem o mesmo padrão de prosódia neutra, a não ser por nuances particulares relativas ao conteúdo.

Assim, tomando as atitudes consideradas neste trabalho como exemplo, propomos um procedimento de transplante por superposição baseado no conceito de que a prosódia quando produzida com uma determinada atitude, pode ser modelada

pela superposição do padrão de uma prosódia neutra com alterações típicas desta atitude. Neste caso, o transplante por superposição pode ser entendido como um procedimento de transformação ao invés de cópia.

Em outras palavras, como as sentenças estudadas possuem a mesma organização silábica, então cada sílaba pode ser identificada pela sua posição na sequência. Assim, para uma sílaba na posição j, os parâmetros prosódicos desta sílaba são organizados no vetor $\mathbf{k}_j = [k_{dn}, k_{do}, k_I, k_{p0}, k_{p1}, k_{p2}, k_{p3}]$, conforme proposto no Capítulo 4. Portanto, o padrão prosódico das sentenças observadas neste estudo, de seis sílabas, é definido por uma sequência de vetores $\mathbf{K} = [\mathbf{k}_0, ..., \mathbf{k}_5]$. Então, se o padrão da prosódia neutra é dado por $\mathbf{K}_{neutra}$ e o padrão prosódico de uma atitude é dado por $\mathbf{K}_{att}$, o transplante por superposição considera que:

$$\mathbf{K}_{att} = \mathbf{K}_{neutra} + \Delta\mathbf{K}_{neutra \rightarrow att}$$

onde $\Delta\mathbf{K}_{neutra \rightarrow att}$ representa o que chamaremos de "modelo" da atitude, que transforma, por superposição, o padrão da prosódia neutra no padrão prosódico da atitude.

Assim, se considerarmos que a influência do conteúdo sobre a prosódia se aplica tanto na prosódia neutra quando na prosódia da atitude, então, a componente relativa ao modelo da atitude $\Delta\mathbf{K}_{neutra \rightarrow att}$ poderá ser menos suscetível ao conteúdo.

Na figura 5.3 é mostrado o esquema do transplante de prosódia por superposição. Neste caso, o transplante é realizado em duas etapas: primeiro é feita a determinação do modelo da atitude e em seguida o modelo pode ser aplicado a qualquer outra sentença de prosódia neutra, com a mesma estrutura silábica.

Conforme a figura, seja uma "sentença de conteúdo A" produzida em duas versões: uma versão produzida com prosódia neutra e outra versão produzida com uma atitude. Os blocos nomeados de Análise executam o mesmo procedimento descrito na seção anterior, dando origem à estrutura hierárquica de cada sentença, com os parâmetros prosódicos determinados. Portanto, a diferença entre os parâmetros prosódicos da versão neutra e da versão contendo a atitude determina o modelo $\Delta\mathbf{K}_{neutra \rightarrow att}$ da atitude.

A partir deste ponto, a prosódia neutra de qualquer "sentença de conteúdo B" pode transformada para o padrão da atitude modelada, sobrepondo os parâmetros prosódicos das sílabas da sentença B aos parâmetros do modelo da atitude, determinados previamente.

Sob este ponto de vista, o modelo $\Delta\mathbf{K}_{neutra \rightarrow att}$ das atitudes indica as alterações nos parâmetros prosódicos que transformam a prosódia neutra em uma determinada atitude. Na seção a seguir será apresentada uma proposta de representação para o como o modelo das atitudes.

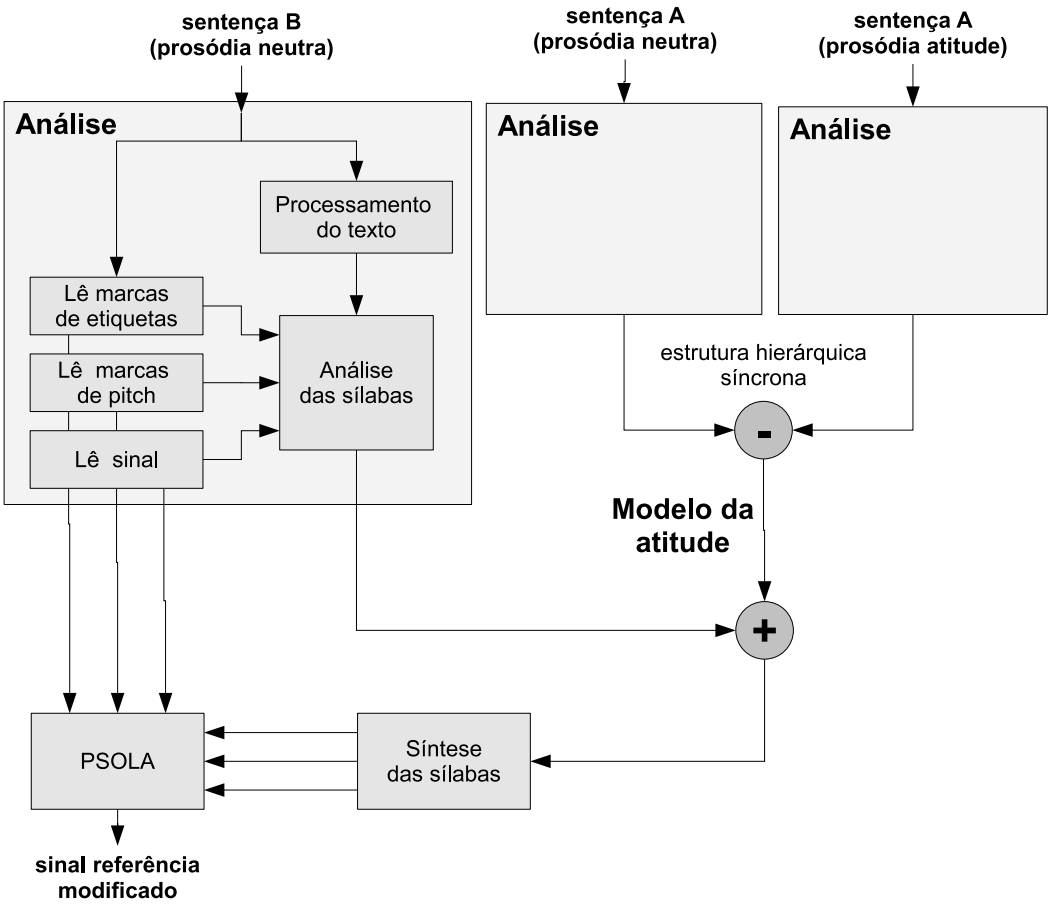


Figura 5.3: Esquema do transplante paramétrico de prosódia por superposição.

5.3.1 O modelo das atitudes

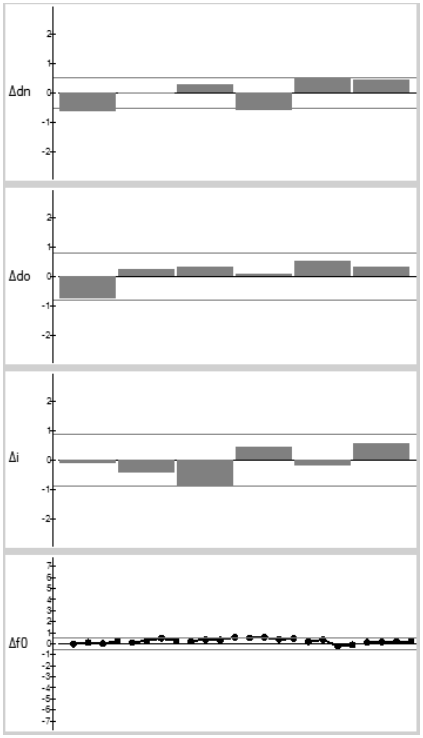
Foi visto que o modelo de uma atitude é dado pela diferença entre os parâmetros de uma prosódia neutra e os parâmetros da prosódia de uma atitude. No entanto, na Seção 4.6 foi visto que existem flutuações consideráveis nos parâmetros da prosódia neutra em relação a valores médios. Portanto, a partir destes valores médios, podemos definir um padrão de prosódia neutra médio $\bar{\mathbf{K}}_{neutra}$, onde os parâmetros prosódicos de cada sílaba na posição j serão dados pela média obtida ao longo das frases de teste, ou seja, por $\bar{\mathbf{k}}_j = [\bar{k}_{dn}, \bar{k}_{do}, \bar{k}_I, \bar{k}_{p0}, \bar{k}_{p1}, \bar{k}_{p2}, \bar{k}_{p3}]$. Deste modo, os desvios em torno deste padrão médio fornecem limiares que podem ser usados como indicativo de alterações nos parâmetros que fogem ao padrão neutro.

Assim, o modelo das atitudes pode também ser obtido usando o padrão de prosódia neutra médio como referência, ou seja, $\Delta\mathbf{K}_{att} = \mathbf{K}_{att} - \bar{\mathbf{K}}_{neutra}$.

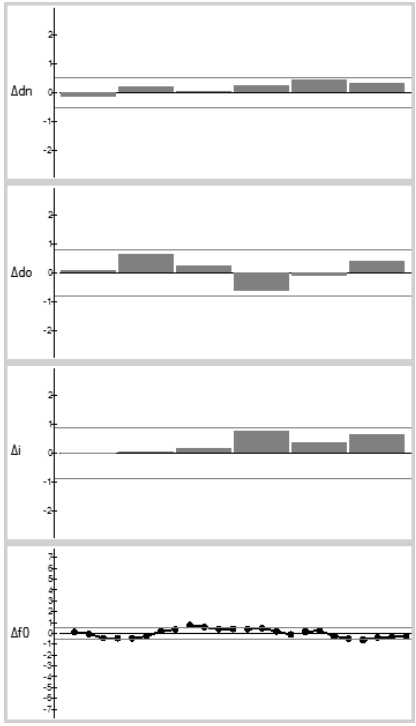
Para a representação dos modelos $\Delta\mathbf{K}_{att}$ das atitudes, foi feita uma interface de visualição no *software* desenvolvido. Na figura 5.4b são mostrados dois exemplos de obtenção do modelo das sentenças "Renata jogava" e "Repete de novo", ambas produzidas com prosódia neutra.

Os quatro gráficos da figuras representam, respectivamente, a alteração das du-

rações da vogal do *núcleo*, da durações dos segmentos em *onset*, da intensidade e dos quatro pontos de pitch. As duas linhas horizontais mais claras, acima e abaixo do nível zero, indicam os desvios da média considerados. Verifica-se que para estas sentenças de exemplo, produzidas com prosódia neutra, todos os parâmetros do modelo se encontram dentro dos limiares, o que indica não haver alterações relevantes entre a prosódia neutra média e a prosódia neutra produzida nestas sentenças.



(a) Parâmetros do modelo da prosódia da sentença "Repete de novo", produzida com prosódia neutra.



(b) Parâmetros do modelo da prosódia da sentença "Renata jogava", produzida com prosódia neutra.

Figura 5.4: Exemplo gerados pelo *software* para representação dos modelos prosódicos.

No artigo de MORAES [5], o autor introduz mudanças em um sinal de prosódia neutra para simular cada um dos padrões das atitudes. Em seguida, os sinais ressintetizados foram submetidos a um teste subjetivo para avaliar a significância destas alterações para o reconhecimento das atitudes. Na seção a seguir, propõe-se validar a proposta de codificação dos parâmetros prosódicos comparando as alterações indicadas pelo modelo das atitudes com as alterações encontradas por MORAES [5] como significantes para o reconhecimento das atitudes.

5.4 A prosódia das atitudes

Sem a pretensão de classificar rigorosamente as atitudes consideradas por MORAES [5], buscou-se descrever intuitivamente o entendimento a respeito da intenção do

falante ao produzir as atitudes:

- A atitude Óbvia (*self-evident assertion*) indica uma afirmação evidente, no sentido de "é claro".
- A atitude de Sugestão (*suggestion*) indica uma ação a ser tomada, onde o falante expressa ser uma ação evidente.
- A atitude de Ênfase Contrastiva (*contrastive emphasis*) representa uma afirmação em que o falante, mediante diferentes afirmações de mesmo teor, deseja destacar uma delas. Por exemplo, como resposta à pergunta: - Renata cantava ou jogava?, o falante destaca que Renata "jogava".
- As atitudes nomeadas de "Questão QU" (*wh-question*), referem-se às questões iniciadas por Que, Quem, Qual, Quando, cujo termo "QU", foi adaptado do termo *wh* que refere-se às palavras em inglês Who, What, Why,... Nota-se o termo também se aplica às palavras Porque e Como, apesar de não iniciarem por QU.
- A atitude de Comando (*command*) indica uma ação a ser cumprida.
- A atitude de Exclamação QU (*wh-exclamation*) trata-se de uma sentença exclamativa, porém iniciada tal como uma Questão QU. Por exemplo, a exclamação: - Que grande bobagem!
- A Questão SN (*yes-no question*) refere-se a uma pergunta na qual não há expectativa de resposta positiva ou negativa.
- Na atitude Questão SN Retórica (*rhetoric yes-no question*) a intenção do falante não está em obter uma resposta em si, pois ele já possui uma opinião formada a *priori*. Normalmente esta atitude é usada como uma proposta ao debate.
- A atitude de pedido (*request*) indica uma ação na qual será realizada ou não, conforme a vontade de quem é dirigida a ação. Um favor.
- A Questão SN Incrédula (*incredulous sn question*) é uma pergunta em que o falante expressa um conhecimento prévio contrário.
- A atitude de Aviso (*warning*) representa uma afirmação de alerta.
- Na atitude Irônica (*ironic assertion*) o falante deseja expressar que sua opinião é contrária ao conteúdo da afirmação.
- A atitude Incrédula (*incredulous assertion*) é uma afirmação em que o falante demonstra claramente não acreditar na proposição.

- A atitude de Ênfase Intensiva (*intensive emphasis*) se aplica somente em nomes e verbos quantificáveis [5], com o propósito de indicar uma qualidade ou ação acima da média.

Ao longo deste trabalho, tem-se usado regravações dos sinais considerados no estudo original de MORAES [5]. Isto foi feito para executar a extração precisa do pitch por meio do eletroglotógrafo e buscar manter constante as condições de gravação das sentenças. As frases gravadas por MORAES [5] foram imitadas fielmente, e portanto, considera-se que o teste subjetivo aplicado por MORAES [5] na distinção das atitudes se reproduza nas regravações desta tese.

Conforme as observações de MORAES [5], o autor divide o conjunto de atitudes em três grupos: as atitudes que apresentam um movimento de descida no contorno de pitch; um movimento de subida; e com uma marcante alteração na duração. Estes grupos serão observados a seguir na mesma ordem apresentada por MORAES [5].

5.4.1 Atitudes com movimentos de descida no pitch

Estão incluídos neste grupo as atitudes Óbvia, de Sugestão, Ênfase Contrastiva, Questão QU, Comando e Exclamação QU. Foi descartada desta análise a atitude de Pedido de confirmação (*request for confirmation*) devido ao baixo reconhecimento desta atitude, tal como identificado por MORAES [5], e pela dificuldade em reproduzi-la.

Óbvia (*self-evident assertion*)

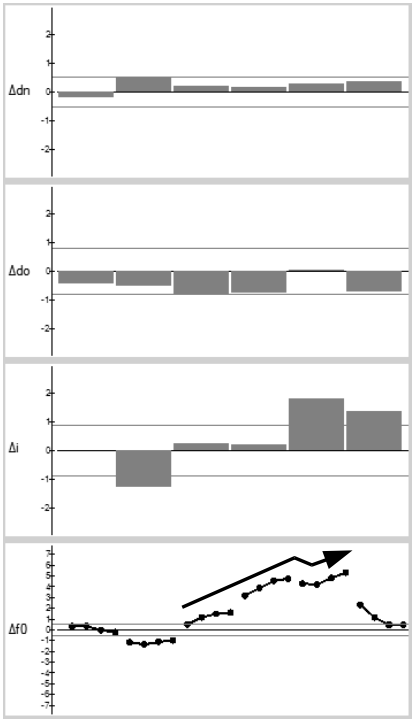
MORAES [5] observa que a atitude de obviedade se contrasta da neutra por um movimento de subida geral do contorno de pitch até a parte final. No entanto, o nível alto de pitch atingido pela tônica é o fator responsável pela identificação deste padrão de atitude.

Na figura 5.5a, a seta indica o movimento de subida do contorno de pitch e o movimento observado na tônica final. Adicionalmente, observa-se um aumento na intensidade na tônica e postônica, demonstrado pelo aumento em Δk_I ultrapassando os limiares de desvio máximo da sentença neutra média. Os parâmetros de duração não mostraram indícios de alteração, mantendo-se dentro dos limiares.

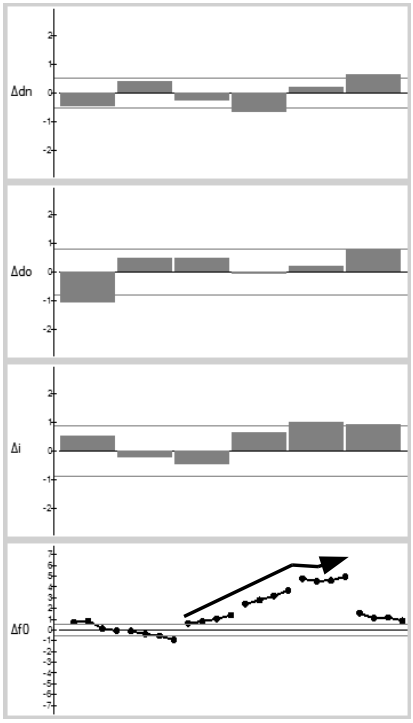
Sugestão (*suggestion*)

MORAES [5] indica que o padrão da atitude Óbvia quando aplicado a sentenças imperativas tem a força de sugestão.

Na figura 5.5b é mostrado o modelo da atitude de Sugestão quando produzido com a sentença "Repete de novo". Neste caso, nota-se a similaridade com o modelo da atitude Óbvia, mostrado na figura 5.5a.



(a) Atitude Óbvia contento um movimento de subida na parte final, onde a tônica atinge um nível alto.



(b) Atitude de Sugestão produzida com uma sentença imperativa, apresentando o mesmo modelo da atitude Óbvia.

Figura 5.5: Parâmetros dos modelos das atitudes Óbvia e de Sugestão.

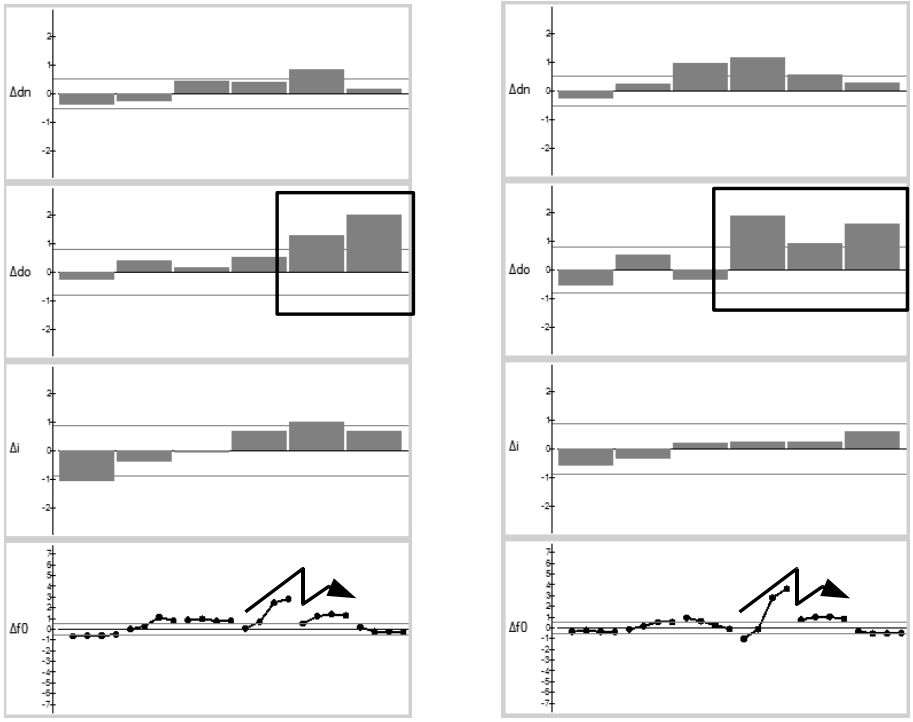
Ênfase Contrastiva (*contrastive emphasis*)

MORAES [5] observa que as principais alterações no contorno melódico desta atitude são o alto nível no contorno de pitch atingido pela sílaba pretônica final e um formato convexo na sílaba tônica final, em um nível baixo. Porém, MORAES descreve ainda que estas alterações não são fortes o bastante para o reconhecimento desta atitude e somente com um aumento na duração e intensidade o reconhecimento desta atitude chega a 80%.

Na figura 5.6a observa-se o formato descrito no contorno de pitch, e ainda, o aumento em Δk_{do} indicando o prolongamento das consontes em *onset* nas sílabas tônica e na postônica finais. Nota-se ainda um aumento na intensidade nas sílbas da palavra que recebe a ênfase, demonstrado por Δk_{do}, mas os valores ainda se apresentam dentro dos limiares de variação.

Na regravação desta atitude, observou-se que é recorrente uma outra forma de Ênfase Contrastiva que alonga acentuadamente o *onset* da primeira sílaba da palavra

em foco, conforme demonstra a figura 5.6b por um aumento expressivo em Δk_{do}.



(a) Atitude de Ênfase Contrastiva com um movimento de subida no pitch na sílaba pretônica e um formato convexo na sílaba tônica em um nível baixo. Além disso, o aumento em Δk_{do} indica o alongamento da duração das consoantes.

(b) Exemplo de outra forma de produção de Ênfase Contrastiva com alongamento adicional da consoante do *onset* da sílaba pretônica.

Figura 5.6: Parâmetros do modelo da atitude de Ênfase Contrastiva.

Questão QU (*wh-question*)

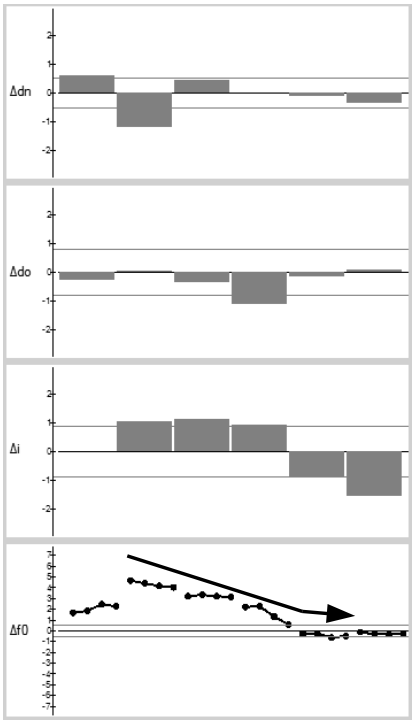
Para a produção desta atitude, MORAES [5] utiliza a sentença "Como ela jogava". Porém, neste caso ocorrem duas tônicas, nas palavras "como" e "ela", interferindo na observação desta atitude. Assim, para manter a estrutura silábica semelhante, nesta tese foi usada a sentença "e como jogava".

Na figura 5.7a é mostrado o modelo desta atitude, caracterizado por um um nível alto de pitch na palavra QU "como", seguido de um movimento de descida gradual até a última sílaba tônica, tal como descrito por MORAES [5]. As alterações acentuadas na intensidade, observadas na figura 5.7a, não representam alterações significativas para o reconhecimento desta atitude.

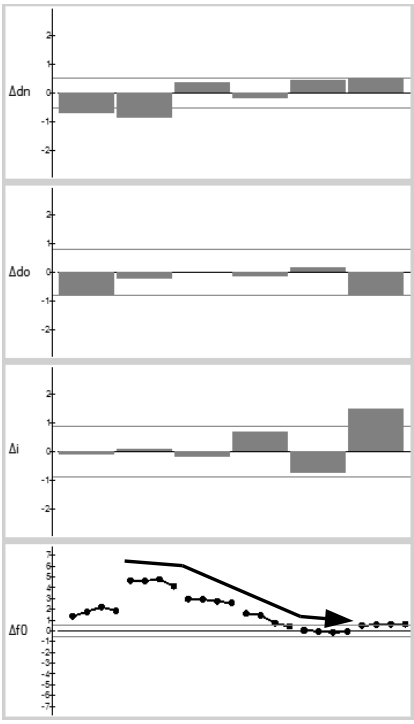
Comando (*command*)

Para a produção desta atitude foi utilizada a sentença "Repete de novo". O modelo observado nesta atitude, mostrado na figura 5.7b, é similar à atitude de Questão

QU, tal como considerado por MORAES [5].



(a) Atitude Questão QU, caracterizada por um nível alto no pitch na palavra "como" e um movimento de descida gradual até a última sílaba tônica.



(b) Atitude de Comando, cujo contorno de pitch é similar à Questão QU.

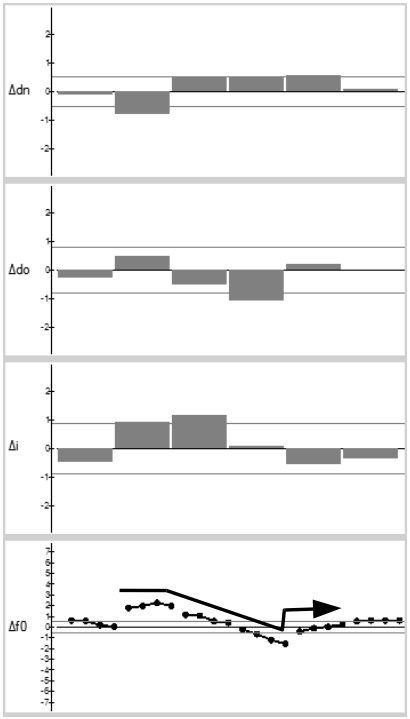
Figura 5.7: Parâmetros dos modelos das atitudes Questão QU e Comando.

Exclamação QU (*wh-exclamation*)

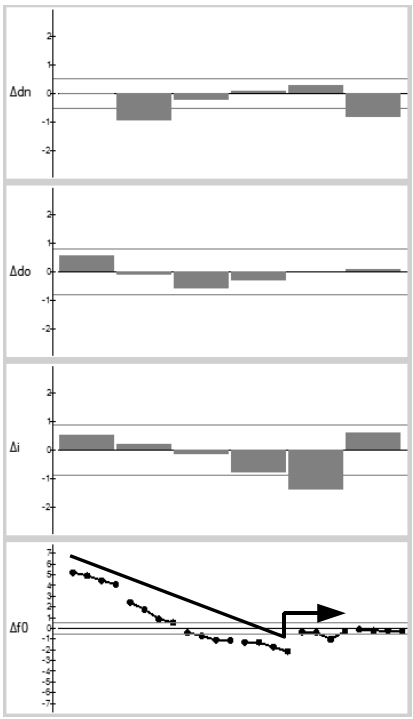
A atitude de Exclamação QU também foi produzida com a sentença "e como jogava", pela razão citada anteriormente. MORAES indica que o padrão prosódico desta atitude apresenta um contorno de pitch muito similar à Questão QU. Porém, o nível de pitch mais baixo na pretônica tem um forte efeito no reconhecimento desta atitude.

Curiosamente, ao compararmos o padrão melódico da atitude Exclamação QU, na figura 5.8a com o padrão da atitude Questão QU, na figura 5.7a, observa-se que o contorno de pitch antes da sílaba tônica final são similares. Porém a Exclamação QU apresenta um deslocamento para um nível mais baixo.

Um outro exemplo do padrão prosódico desta atitude foi produzido com a sentença "Que torta gostosa". Neste caso, tal como mostra a figura 5.8b, o declínio do contorno de pitch inicia na palavra QU e não na sílaba tônica da palavra "torta". Constata-se que o início deste declínio é opcional, podendo ocorrer tanto na sílaba tônica quanto na palavra QU.



(a) Atitude de Exclamação QU, caracterizada por um movimento de descida gradual a partir de um nível alto de pitch na tônica da palavra QU "como".



(b) Exemplo de produção da atitude de Exclamação QU na sentença "que torta gostosa", onde o movimento de descida do pitch inicia na palavra QU.

Figura 5.8: Parâmetros do modelo da atitudes de Exclamação QU.

5.4.2 Atitudes com movimento de subida no pitch

Nas atitudes agrupadas por MORAES [5] por conterem um movimento de subida no pitch estão as atitudes de Questão SN, de Questão SN Retórica, de Pedido e a Questão SN Incrédula. Na ressíntese destas atitudes, [5] toma a Questão SN como base para a manipulação das variáveis prosódicas. No entanto, ao longo desta seção será mantida a comparação feita até o momento com o padrão neutro médio.

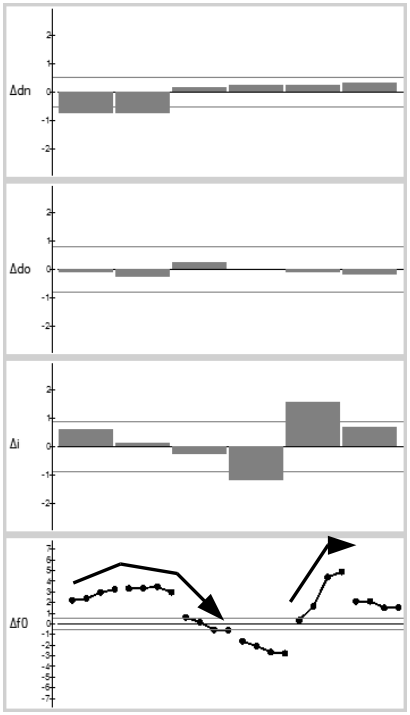
Questão SN (*yes-no question*)

MORAES [5] descreve que o contorno melódico desta atitude é caracterizado por uma subida na sílaba tônica inicial, seguida por uma descida até a postônica inicial. Na figura 5.9a observa-se o movimento inicial no contorno de pitch, indicado pela primeira seta. Porém, MORAES [5] acrescenta que o fator predominante é o movimento de subida entre as sílabas pretônica e tônica finais, indicado pela segunda seta na figura 5.9a.

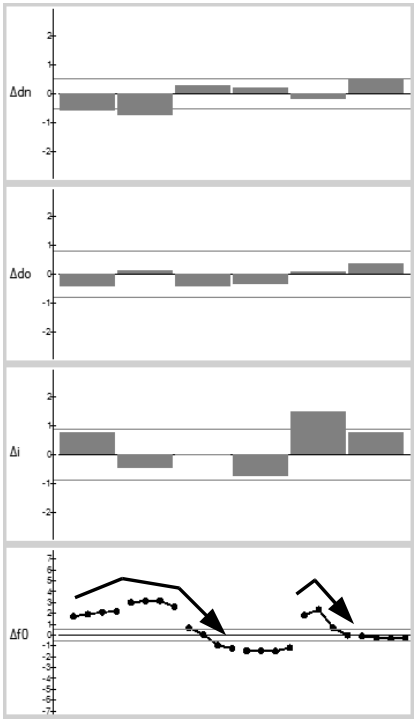
Questão SN Retórica (*rhetoric yes-no question*)

MORAES [5] observa que o padrão desta atitude apresenta um contorno de pitch da palavra inicial muito similar à Questão SN. Porém, a sílaba tônica final é caracterizada por um movimento de descida, ao contrário do movimento de subida da Questão SN. Além disso, o nível do pitch na postônica é reduzido.

Ao compararmos os modelos das figuras 5.9b e 5.9a, nota-se a diferença no movimento de descida na sílaba tônica ao invés de subida.



(a) Atitude de Questão SN com uma subida no contorno de pitch da primeira sílaba tônica, seguido por uma descida até a pretônica. Após, uma subida na sílaba tônica final, descendo novamente na postônica.



(b) Atitude de Questão SN Retórica com movimento inicial similar à Questão SN, porém com um movimento de descida na tônica.

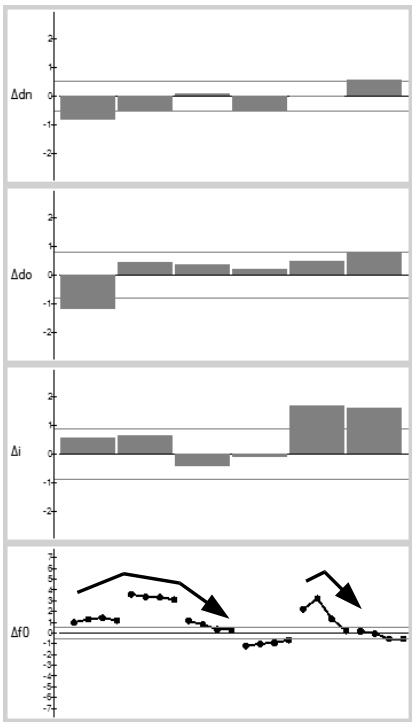
Figura 5.9: Parâmetros dos modelos das atitudes Questão SN e Questão SN Retórica.

Pedido (*request*)

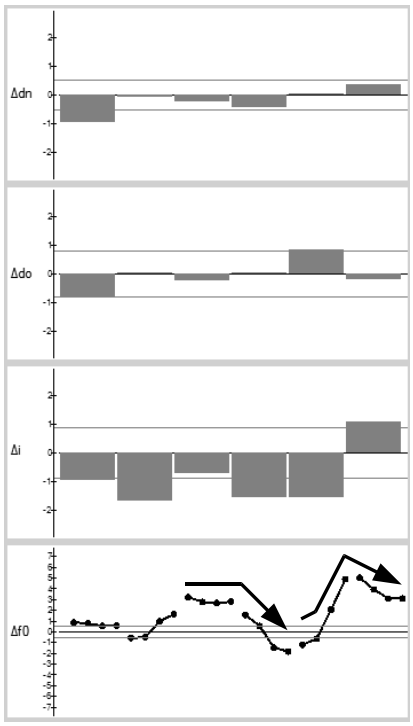
Esta atitude foi produzida na sentença "repete de novo". MORAES [5] indica que o contorno de pitch desta atitude, mostrado na figura 5.10a, apresenta um padrão similar à Questão SN Retórica, como podemos observar em comparação com a figura 5.9b.

Questão SN Incrédula (*incredulous yes-no question*)

MORAES [5] descreve que a característica predominante para o reconhecimento desta atitude é o vale no contorno de pitch formado com o movimento de descida na sílaba pretônica e a subida na tônica final. Na figura 5.10b, é mostrado o modelo desta atitude, quando produzida tal como a sentença original utilizada por MORAES. Nesta caso, nota-se o vale citado pelo autor como predominate e ainda uma alteração global nas intensidades. Constata-se que esta alteração global na intensidade não indica uma diferenciação desta atitude, mas sim uma gravação de baixa amplitude.



(a) Atitude de Pedido, cujos parâmetros são similares aos da Questão SN Retórica.



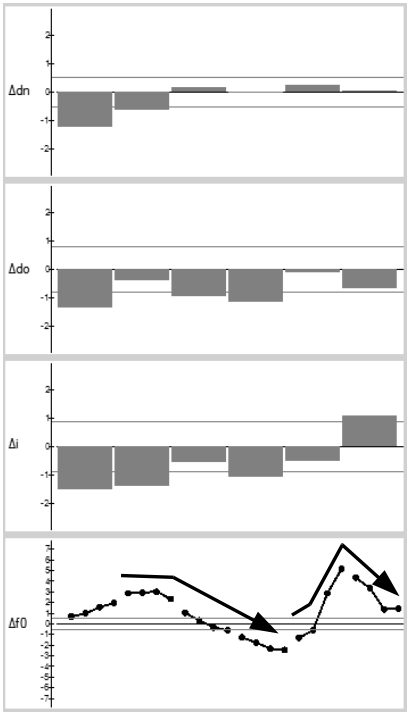
(b) Atitude Questão SN Incrédula, caracterizada por um vale no contorno de pitch entre a sílaba pretônica e tônica.

Figura 5.10: Parâmetros dos modelos das atitudes de Pedido e da Questão SN Incrédula.

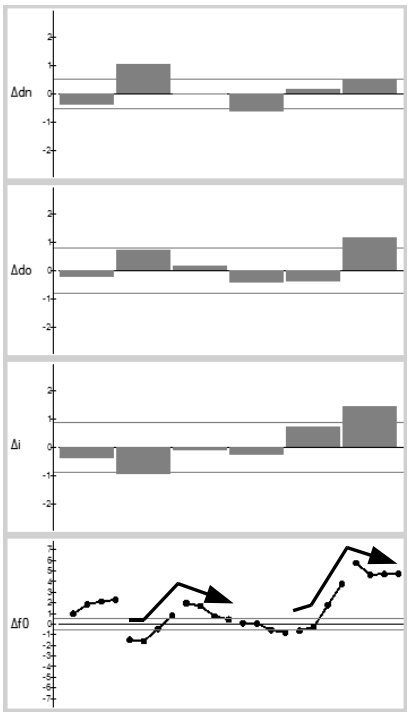
Na produção desta atitude, foi vista a existência de algumas variantes. Por exemplo, na figura 5.11a é mostrado um exemplo de uma produção muito semelhante à Questão SN. O que diferencia esta produção da Questão SN é basicamente o formato de vale no contorno de pitch entre as sílabas pretônica e tônica finais, o que MORAES descreve ser o fator predominante.

Nos casos apresentados, o foco da descrença está na palavra "jogava". Se na produção desta atitude, a descrença também se aplicar à palavra "Renata", curiosamente o movimento de subida no contorno de pitch encontrado na sílaba tônica

final também surge na sílaba tônica da primeira palavra, tal como mostra a figura 5.11b.



(a) Atitude Questão SN Incrédula semelhante à Questão SN.



(b) Atitude de Questão SN Incrédula, onde surge na sílaba tônica da primeira palavra o mesmo movimento no contorno de pitch.

Figura 5.11: Parâmetros do modelo da atitude Questão SN Incrédula.

5.4.3 Atitudes com alteração nas durações

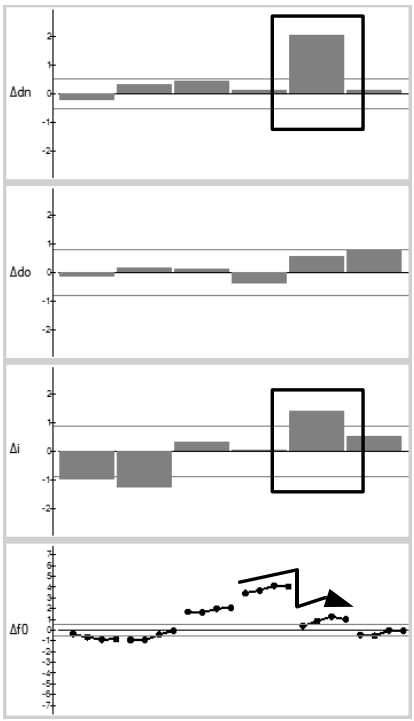
Observando o modelo das atitudes que possuem a duração da sílaba alongada, este alongamento pode decorrer tanto do aumento das vogais da sílaba quando das consoantes em *onset*, tal como citado anteriormente na proposta de parametrização das variáveis prosódicas, no Capítulo 4. Neste grupo serão vistas a seguir as atitudes de Aviso, Irônica, Incrédula e de Ênfase Intensiva.

Aviso (*warning*)

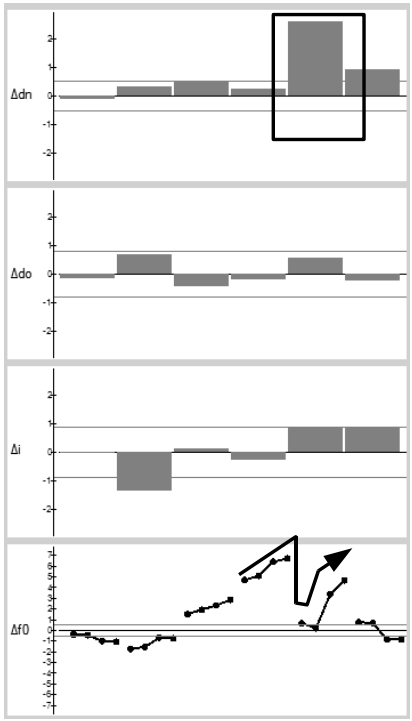
Na atitude de Aviso, MORAES [5] descreve que as características cruciais para a identificação desta atitude é o aumento da duração da vogal juntamente com o alto nível do pitch na sílaba pretônica. Na figura 5.12a é mostrado o modelo de uma atitude de Aviso onde se nota o aumento típico do parâmetro k_{dn} de duração da vogal do *núcleo* da sílaba. Além disso, observa-se o movimento de subida do contorno de pitch, terminando em um alto nível na sílaba pretônica.

Irônica (*ironic assertion*)

MORAES [5] identifica que o alongamento da vogal da sílaba tônica, assim como o alto nível do pitch na pretônica, ocorrem nesta atitude tal como na atitude de Aviso. No entanto, o contorno de pitch da sílaba tônica apresenta um formato particular, que inicia por um nível constante até a metade da vogal quando ocorre um movimento de subida. Na figura 5.12b é mostrado o padrão desta atitude onde se nota o mesmo aumento em k_{dn} exibido pela atitude de Aviso e o formato particular do contorno de pitch da sílaba tônica.



(a) Atitude de Aviso com um alongamento da vogal da sílaba tônica e um pitch de nível alto na pretônica, seguido por um declínio na sílaba tônica para um nível pouco maior do que a neutra e com um formato convexo.



(b) Atitude Irônica, parecida com a atitude de Aviso, no entanto com um formato diferente no pitch na sílaba tônica, contendo um leve declínio, seguido de uma subida na metade até o fim.

Figura 5.12: Parâmetros dos modelos das atitudes de Aviso e Irônica.

Incrédula (*incredulous assertion*)

MORAES [5] indica que a atitude Incrédula é caracterizada por um contorno de pitch com poucas modulações e um leve declínio, praticamente plano. Este fator é apontado como preponderante para o reconhecimento desta atitude.

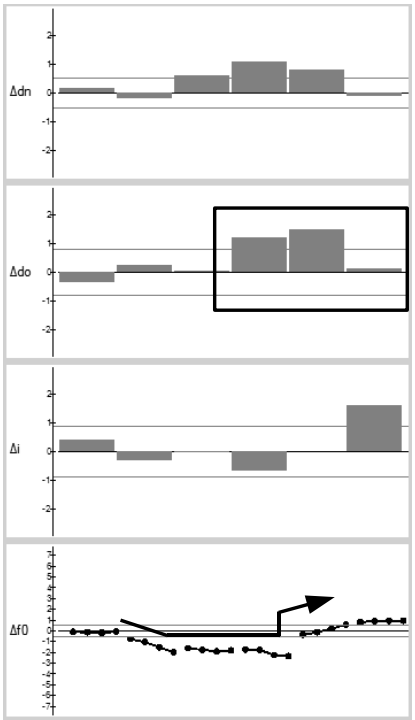
Neste caso, verifica-se que o contorno obtido pela diferença com a sentença neutra, como mostra a figura 5.13a, exibe um formato praticamente simétrico à atitude neutra, o que dá origem a um formato plano após a superposição.

Por ressíntese, nota-se que pequenas diferenças entre a prosódia neutra da sentença de referência e o modelo da prosódia neutra criam modulações no contorno de pitch que são suficientes para descaracterizar esta atitude. Neste caso, o transplante por cópia apresenta um resultado melhor do que o transplante por superposição.

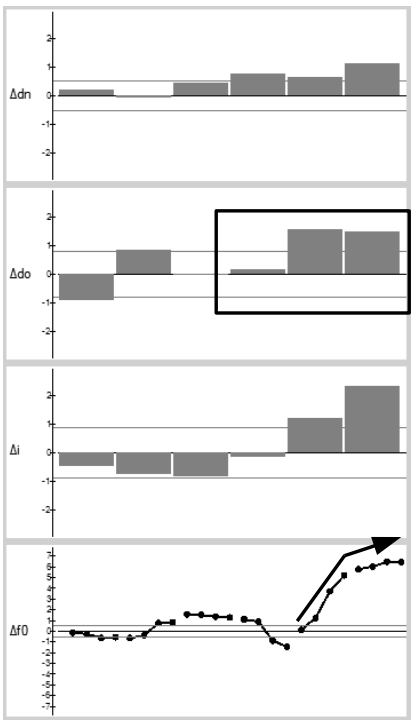
O aumento na duração das consoantes, principalmente na sílaba tônica, é citado por MORAES como um segundo fator marcante desta atitude. Verifica-se na figura 5.13a um aumento no parâmetro Δk_{do} que representa esta alteração.

Ênfase Intensiva (*intensive emphasis*)

Esta atitude se aplica somente em nomes e verbos quantificáveis [5], com o propósito de indicar uma qualidade ou ação acima da média. MORAES [5] identifica que os fatores preponderantes para o reconhecimento desta atitude é o alto nível do pitch no fim da sílaba tônica, mantendo-se alto na sílaba postônica, e o aumento da duração dos segmentos consonantais nestas sílabas. Na figura 5.13 observamos o aumento em k_{do} nas últimas sílabas, refletindo o aumento nas durações das consoantes em *onset*, e o alto nível no na sílaba postônica, tal como indicado por MORAES. Nota-se que esta atitude possui um comportamento de uma sentença não terminada, o que explica o aumento da intensidade na última sílaba.



(a) Atitude Incrédula contendo um aumento k_{do} e um contorno de pitch simétrico à atitude neutra, resultando em um contorno de pitch plano.



(b) Atitude de Ênfase Intensiva com uma subida no contorno de pitch da sílaba tônica mantendo-se em nível alto, e um alongamento das consoantes expresso por k_{do}.

Figura 5.13: Parâmetros dos modelos das atitudes Incrédula e de Ênfase Intensiva.

Observa-se que o aumento na duração das consoantes no *onset* da sílaba acontece geralmente em casos onde se deseja certo grau de ênfase. Supõe-se que este alongamento produz o efeito de atrasar a emissão da vogal, de maneira a enfatizá-la. O alongamento da duração das vogais, ao invés de enfatizar a palavra, provoca a sensação de uma fala mais lenta.

Ao observarmos os parâmetros de intensidade ao longo dos modelos de prosódia estudados, nota-se que são poucas as ocasiões em que o parâmetro de intensidade é significante para distinguir uma atitude, apesar do modelo apresentar uma forte alteração desta característica intensidade. Isto se deve ao fato da intensidade ser extremamente sensível às condições de gravação. Além disso, outra explicação para este fato é a correlação do pitch com a intensidade [69].

Ao longo do estudo das atitudes, em diferentes ocasiões o modelo prosódico de uma atitude foi comparado ao de outra, ao invés de ser comparado ao da prosódia neutra. Isto induz a proposta de executar a transformação de uma prosódia em outra, que será apresentada a seguir.

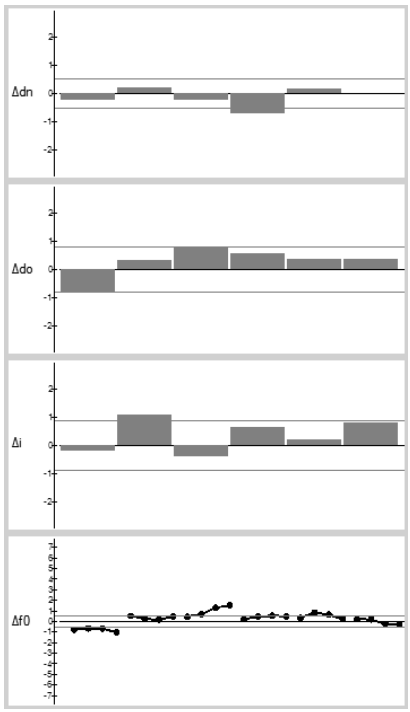
5.5 A combinação de atitudes

A abordagem usada no transplante por superposição pode ser usada para comparar diferentes atitudes entre si. Além desta comparação, o modelo da diferença entre as atitudes pode ser usado para transformar uma atitude em outra, sem o intermédio de um modelo de prosódia neutra. Assim, sejam duas atitudes att_1 e att_2, o modelo que transforma a atitude att_2 na atitude att_1 é dado por $\Delta K_{att_2 \rightarrow att_1} = K_{att_2} - K_{att_1}$.

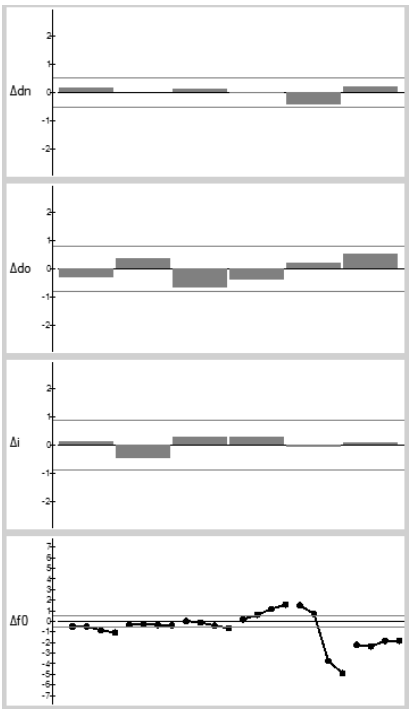
Por exemplo, no estudo das atitude de Pedido e Questão SN Retórica foi visto que o padrão prosódico da Questão SN Retórica, quando aplicado a uma sentença imperativa, faz com que esta sentença seja percebida como uma atitude de Pedido. Na figura 5.14a é dado um exemplo de obtenção do modelo de transformação entre estas atitudes. Nota-se que os parâmetros se encontram praticamente dentro dos limiares, indicando serem atitudes semelhantes.

Em um outro exemplo, MORAES [5] compara as atitudes de Questão SN com a Questão SN Retórica, observando que a diferença entre estas atitudes é a mudança da inclinação do contorno de pitch da sílaba tônica de um movimento de subida para um movimento de descida. Na figura 5.14b é mostrado um exemplo do modelo de transformação entre estas atitudes, onde a diferença observada por MORAES [5] surge como, praticamente, a única caracteristica distintiva.

O modelo de transformação entre atitudes pode fornecer uma medida de distância entre as atitudes. Se tomarmos a norma euclidiana acumulada ao longo das sílabas do modelo de transformação entre as atitudes, temos um indicativo da distância



(a) Modelo de transformação da atitude de Pedido para a Questão SN Retórica, mostrando a semelhança entre as atitudes.



(b) Modelo de transformação da atitude Questão SN para uma Questão SN Retórica.

Figura 5.14: Exemplo do modelo de transformação entre atitudes.

entre estas atitudes. Nota-se que a diferença no contorno de pitch tem maior peso na norma, pois são quatro parâmetros representativos do pitch.

Na figura 5.15 é mostrada a matriz de distância entre as atitudes. Neste caso, os tons mais escuros indicam atitudes mais próximas, enquanto os tons mais claros indicam atitudes mais distantes. A matriz representada na figura 5.15 é simétrica, pois o cálculo da norma independe do sentido da transformação, e a diagonal é igual a zero, quando uma atitude é comparada a ela mesma. Verifica-se que os pares de atitudes de Aviso e Irônica, de Comando e Questão QU, de Pedido e Questão SN Retórica apresentam os menores valores de distância, em torno de 1,7. Esta indicação coincide com as observações de MORAES [5] sobre a semelhança entre estes padrões.

Em [70], com enfoque na classificação de atitudes, os autores executam um procedimento de *cluster* entre atitudes. Porém, a medida de distância entre as atitudes é dada por variáveis globais da sentença, tais como F0, duração e energia médios.

Utilizando os parâmetros prosódicos e a medida de distância utilizados nesta tese, podemos visualizar como as atitudes se agrupam, tal como em [70]. Na figura 5.16 é mostrada a figura de um dendograma, formado conforme a matriz de distância entre as atitudes.

A abordagem da prosódia das atitudes por vetores de parâmetros permite explo-

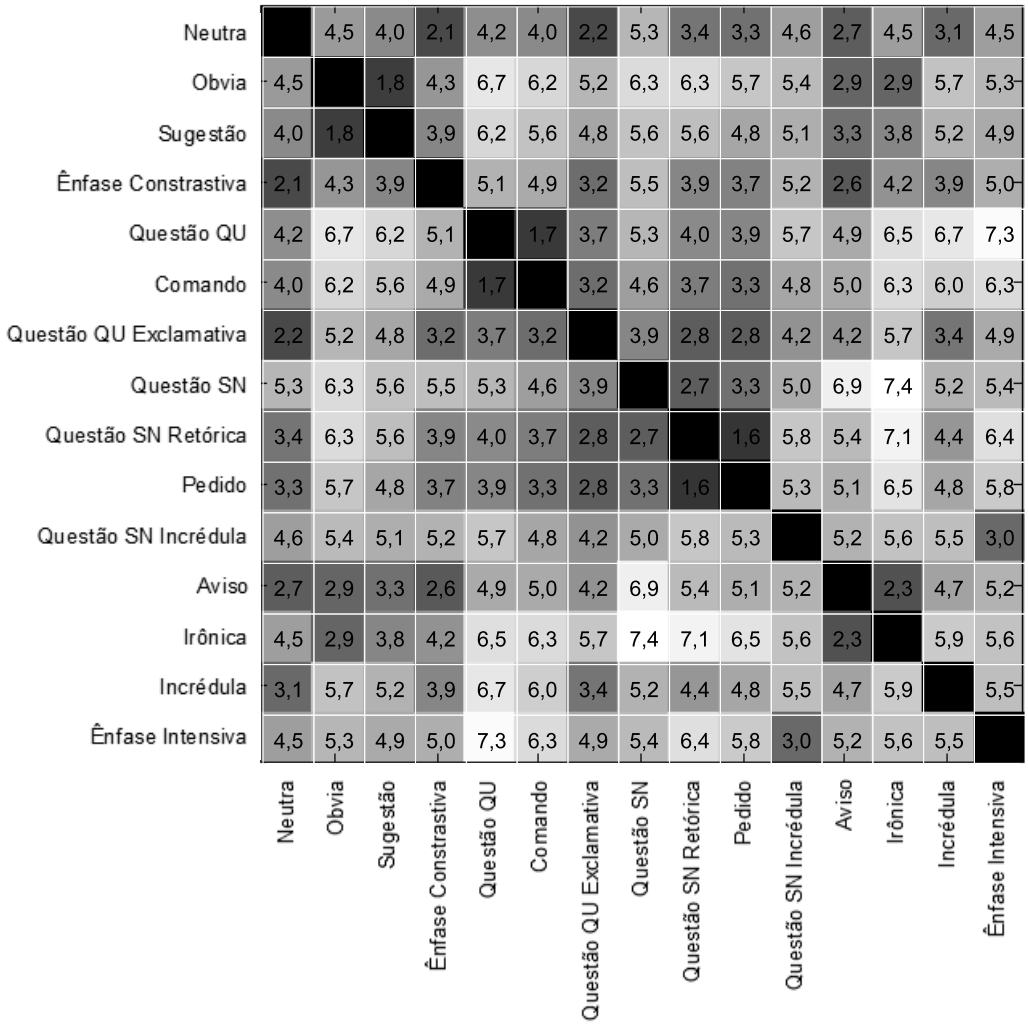


Figura 5.15: Esquema do transplante paramétrico de prosódia por cópia.

rar algumas possibilidades interessantes de manipulação das atitudes. Por exemplo, a sobreposição de uma atitude sobre a prosódia neutra poderia ser ponderada por uma constante, fazendo $\mathbf{K}_{att} = \mathbf{K}_{neutra} + \alpha_{att}\Delta\mathbf{K}_{att}$. Esta possibilidade foi verificada e em alguns casos, como nas atitudes Óbvia, Questão SN e Questão SN Retórica, esta ponderação reflete um aumento ou diminuição da expressividade da atitude. No entanto, em outros casos a atitude é completamente descaracterizada. A ponderação dos parâmetros por uma constante é uma simplificação do caso mais geral, em que a correlação entre os parâmeros seria considerada, ou seja, quando α_{att} seria substituído por uma matriz de transformação. Futuramente, com a gravação de novas produções das atitudes, esta simplificação poderá ser generalizada.

Uma outra possibilidade interessante de manipulação das atitudes é estender o conceito de tomar a prosódia neutra como base. Podemos supor que cada atitude seja formada pela combinação linear de padrões básicos. A comparação entre as atitudes fornece um indício desta possibilidade. Seria ainda muito interessante verificar se esta decomposição se reflete na decomposição da expressividade do falante.

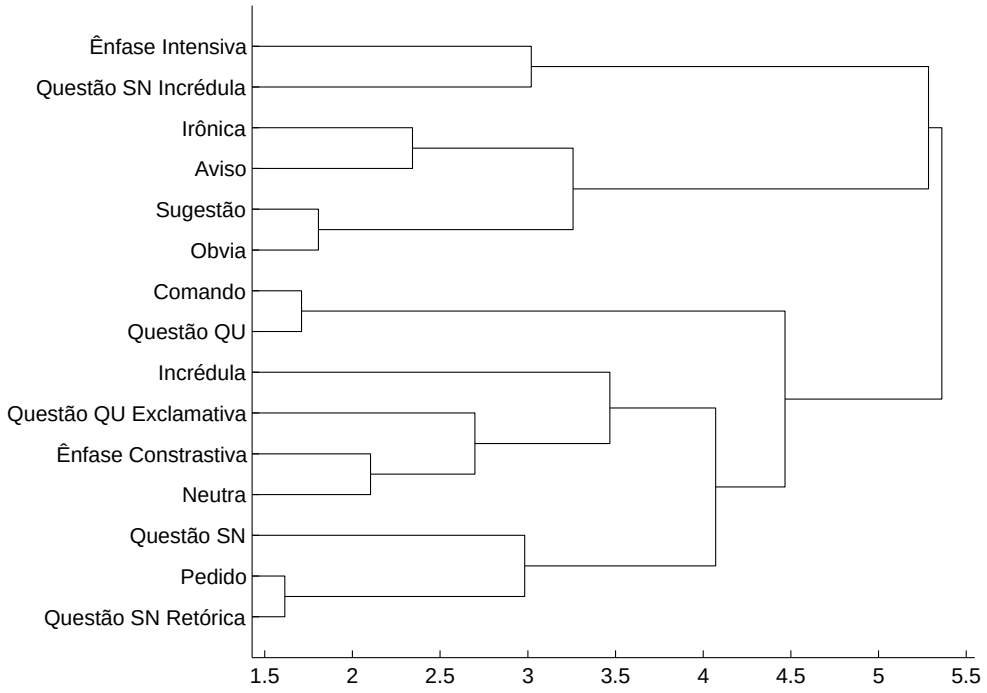


Figura 5.16: Agrupamento das atitudes conforme a medida de distância entre atitudes.

5.6 Conclusões

Neste capítulo foi apresentado o transplante paramétrico da prosódia baseado na codificação dos parâmetros prosódicos da sílaba. Esta codificação permitiu realizar o transplante de prosódia entre sentenças de mesma estrutura silábica, porém, com qualquer conteúdo segmental.

Além do transplante por cópia, foi apresentado o transplante por superposição onde o transplante é realizado sobrepondo o modelo prosódico de uma atitude a uma sentença de prosódia neutra. Por meio deste modo de transplante foi possível observar as alterações nos parâmetros prosódicos das diferentes atitudes quando comparadas a uma prosódia neutra média.

Para validar o esquema proposto, as alterações observadas nos modelos das atitudes foram comparadas com as alterações nas variáveis prosódicas encontradas por MORAES [5] como significativas para o reconhecimento das atitudes. Todas as alterações apontadas por MORAES [5] como contrastivas entre as atitudes foram encontradas nos parâmetros dos modelos. Portanto, conclui-se que o modelo é capaz de indicar as diferenças que são significativas entre as atitudes.

MORAES [5] conclui que decidir o que é contrastivo entre as atitudes, por inspeção visual do contorno melódico, não é uma tarefa simples. Por este ponto de

vista, a parametrização das variáveis prosódicas de duração, intensidade e pitch, no domínio da sílaba oferecem um meio simples de observar e caracterizar novas atitudes.

A implementação do transplante de prosódia de forma paramétrica, onde o conteúdo da sentença pode ser alterado, permite ainda que um determinado padrão prosódico, produzido em uma sentença específica, possa ser aplicado a uma sentença de conteúdo diferente. Por exemplo, a aplicação do padrão de uma atitude de uma questão para uma sentença imperativa. Isto permite que, futuramente, seja verificada a relação entre o significado da sentença e a prosódia.

Utilizando o transplante por superposição foi possível observar as características que diferem uma atitude de outra, e ainda transformar uma atitude diretamente em outra. A partir dos parâmetros desta transformação obteve-se uma medida indicativa das semelhanças entre as atitudes, que coincidem com as observações de MORAES [5].

No próximo capítulo será apresentado o *software* contruído tanto para a análise e síntese da prosódia, incluindo as ferramentas de transplante descritas ao longo deste capítulo.

Capítulo 6

O sistema SASPRO

6.1 Introdução

O objetivo deste capítulo é apresentar as ferramentas de *software* implementadas no sistema de análise e síntese da prosódia, nomeado de SASPRO. Estas ferramentas foram construídas com o propósito de permitir ao usuário observar e interagir com as diferentes etapas do processo de análise e síntese da prosódia, e foram fundamentais para a concretização desta pesquisa.

Todas os componentes citados foram implementadas na linguagem C++, na plataforma Microsoft Visual Studio © 2010, sob o paradigma de orientação por objetos. A Engenharia de Software oferece procedimentos bem estabelecidos para a documentação e descrição do desenvolvimento de um *software*. Porém, neste capítulo não se pretende seguir tal rigor, mas sim apresentar sob um ponto de vista mais amplo algumas das decisões tomadas, que contribuem para possíveis abordagens futuras. Ao longo do capítulo serão apresentados diagramas de classes simplificados. Por questão de clareza, os diagramas de classes completos, incluindo os atributos e métodos de cada classe são apresentados no Anexo D.

Na Seção 6.2 será feita a apresentação inicial das funcionalidades básicas disponíveis pelo sistema, e na Seção 6.3 serão descritas as ferramentas de edição, que executam as funções de edição e visualização dos sinais, das marcas de pitch e das etiquetas de fronteira entre segmentos.

Na Seção 6.4 serão apresentadas as funcionalidades básicas do processamento de texto, descrito anteriormente no Capítulo 2, assim como a implementação da estrutura hierárquica de dados.

Na Seção 6.5 será descrita a implementação das ferramentas de edição de prosódia, que incluem as etapas de análise e síntese, propostas no Capítulo 4.

Na Seção 6.6 será descrito o sistema TTS protótipo, como a junção das funcionalidades descritas nas seções anteriores. Além disso, será descrita a utilização do

sistema para a construção do banco de unidades por concatenação.

Por fim, na Seção 6.7 serão apresentadas as ferramentas de transplante de prosódia, que executam o transplante da prosódia por alinhamento, como descrito no Capítulo 3, e da forma paramétrica, como discutido no Capítulo 5, que podem ser aplicados tanto em sentenças naturais quanto em sentenças geradas pelo sintetizador.

6.2 Apresentação do sistema

Na figura 6.1 é mostrado um recorte da tela principal do sistema SASPRO. No menu de funções, a função Iniciar é usada para indicar ao sistema a localização e a descrição das sentenças que serão analisadas. As três funções seguintes no menu exibem telas dedicadas ao processamento do texto, à conversão texto-fala e à execução do transplante de prosódia. A função seguinte é usada para o gerenciamento de janelas e a última função executa rotinas para a exportação de variáveis e geração dos resultados apresentados nesta tese.

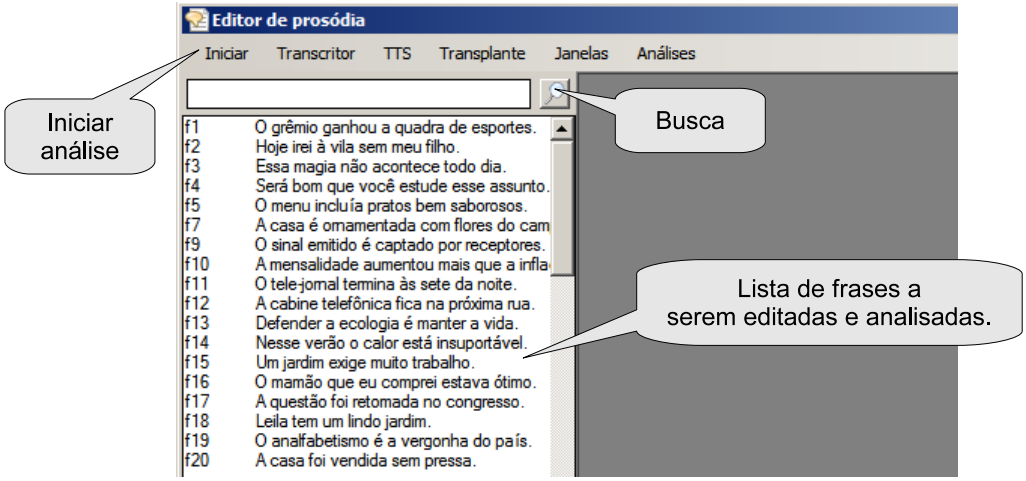


Figura 6.1: Recorte da tela do aplicativo de manipulação prosódica.

No lado esquerdo da tela é mostrada uma lista de sentenças. Esta lista é lida pelo sistema a partir de um arquivo texto contendo na primeira linha o diretório de armazenamento dos arquivos e nas linhas seguintes as sentenças, iniciadas pelo nome que será atribuído aos arquivos. Deste modo, as sentenças a serem editadas e analisadas podem ser organizadas previamente pelo usuário. Na caixa de entrada de texto disposta acima da lista de sentenças, com a indicação de busca, é possível inserir um texto a ser buscado ao longo da lista de sentenças.

Para cada sentença da lista, são gravados quatro arquivos: um arquivo (.wav) contendo os sinais de fala e do EGG (eletroglotógrafo) ou do microfone de contato [10]; um arquivo (.mrk) contendo os instantes de tempo das marcas de pitch; um

arquivo (.lmk) contendo os instantes de tempo das etiquetas de fronteira dos segmentos; e um arquivo (.phn) contendo o resultado do processamento de texto. Este quarteto de arquivos contém toda a informação necessária para a análise subsequente das sentenças.

Ao selecionar uma sentença da lista uma tela de edição é aberta. Se os sinais da sentença selecionada já foram gravados, editados e escritos em disco, então os arquivos são automaticamente lidos e as informações são mostradas nesta tela de edição. Porém, se os sinais ainda não foram gravados e editados, a tela de edição é exibida para que o usuário inicie a gravação e edição das sentenças, por meio das ferramentas de edição, descritas a seguir.

6.3 Ferramentas de edição

As ferramentas de edição foram aprimoradas a partir do *software* iniciado na tese de mestrado [10] e executam as funções de visualização e edição dos sinais, das marcas de pitch e das etiquetas de fronteira entre segmentos.

6.3.1 Funcionalidades

Na figura 6.2 é mostrada a tela de edição da sentença "Os pesquisadores acreditam nesta teoria." já gravada e editada. Na parte superior da tela é mostrada a

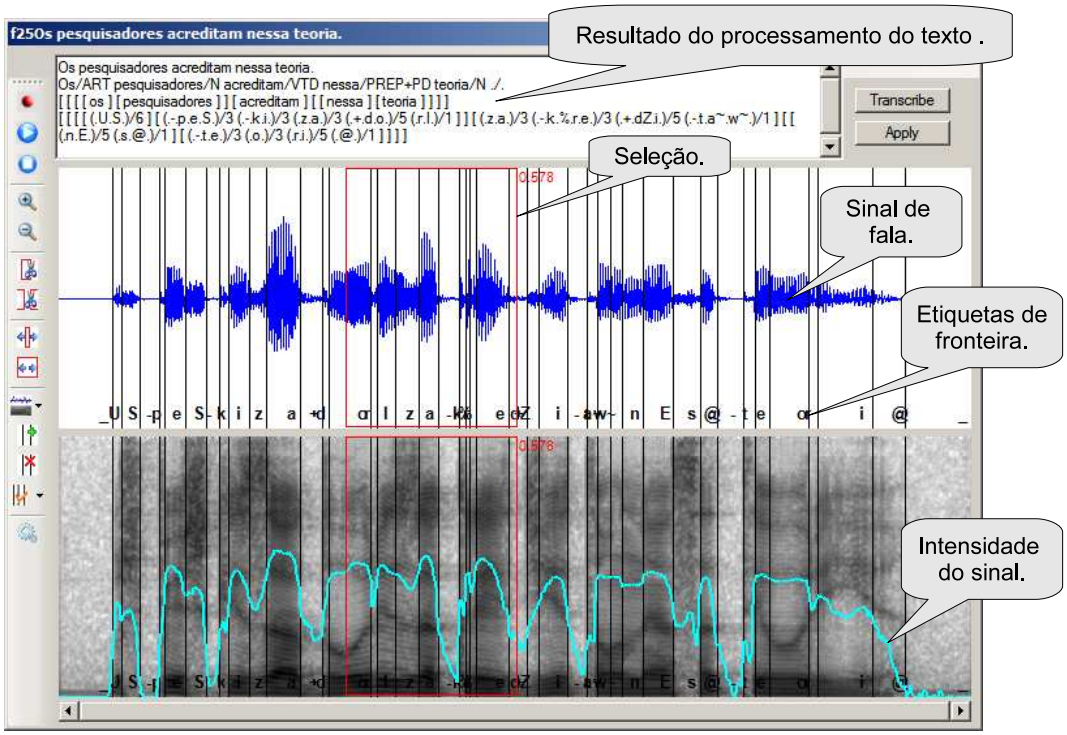


Figura 6.2: Tela de edição.

representação do resultado do processamento do texto da sentença, cujos detalhes

serão vistos na Seção 6.4. Nos gráficos abaixo são mostrados o sinal de fala e o espectrograma do sinal, juntamente com as respectivas etiquetas fonéticas. Sobre o espectrograma também está traçada a curva de intensidade, para auxiliar a inserção das etiquetas de fronteira entre os segmentos.

No lado esquerdo da tela estão disponíveis as funcionalidades das ferramentas de edição, dispostas em um menu vertical e descritas a seguir na figura 6.3.

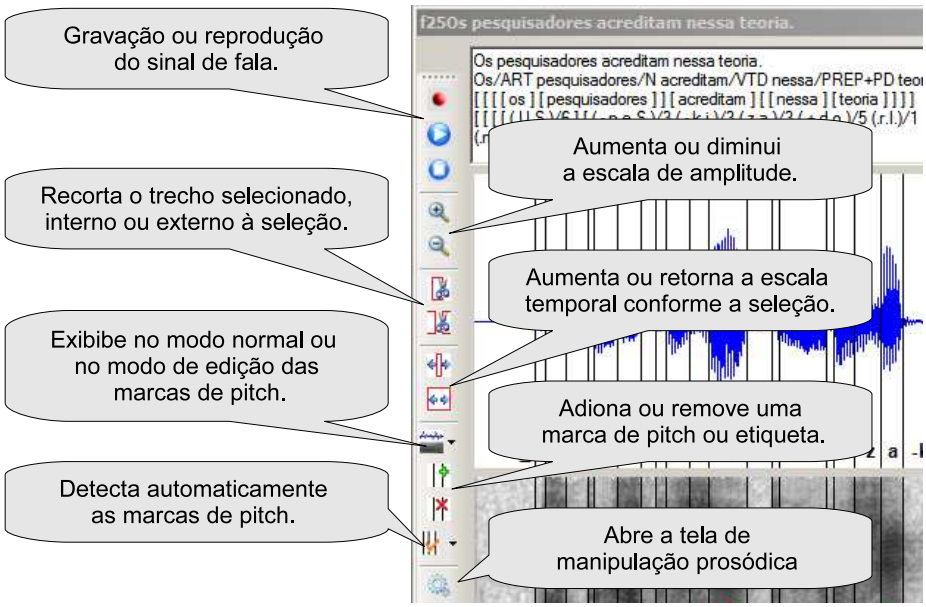


Figura 6.3: Descrição das funcionalidades da tela de Edição.

A função de reprodução pode ser feita somente em um trecho do sinal. Ao arrastar o *mouse* sobre os gráficos, uma área retangular é destacada, tal como mostra a figura 6.2, selecionando um trecho do sinal. As funções de recorte ou de aumento das escala temporal também fazem referência a este trecho do sinal selecionado.

Após a gravação do sinal e recorte do trecho útil do sinal, o usuário precisa inserir as marcas de fronteira dos segmentos e executar a detecção das marcas de pitch.

As etiquetas de fronteira dos segmentos são inseridas manualmente, com um clique do *mouse* sobre o sinal ou sobre o espectrograma. Estas marcações definem somente a posição temporal das fronteiras, que são identificadas por um símbolo temporário "?". A identificação correta das etiquetas é feita somente após a transcrição automática, representada na caixa de texto no topo da figura. Deste modo, mantém-se a correspondência unívoca entre a transcrição automática e as etiquetas, sincronizando toda a estrutura do texto ao sinal.

Para a edição das marcas de pitch existe a opção na barra de ferramentas de visualizar, no espaço onde é mostrado o espectrograma, o sinal auxiliar proveniente do EEG ou do microfone de contato [10]. Neste modo de visualização também são exibidas as marcas de pitch, ao invés das etiquetas de fronteira dos segmentos. Na figura 6.4 é mostrada a tela de edição quando este modo de visualização é selecio-

nado. No gráfico superior é mostrado o sinal de fala e no gráfico inferior o sinal do EGG. Neste caso, um trecho do sinal foi selecionado e a escala temporal dos gráficos foi ampliada, para melhor visualização dos sinais e das marcas de pitch. A barra de rolagem na parte inferior da tela permite que esta janela de visualização do sinal seja deslocada para a esquerda ou para a direita.

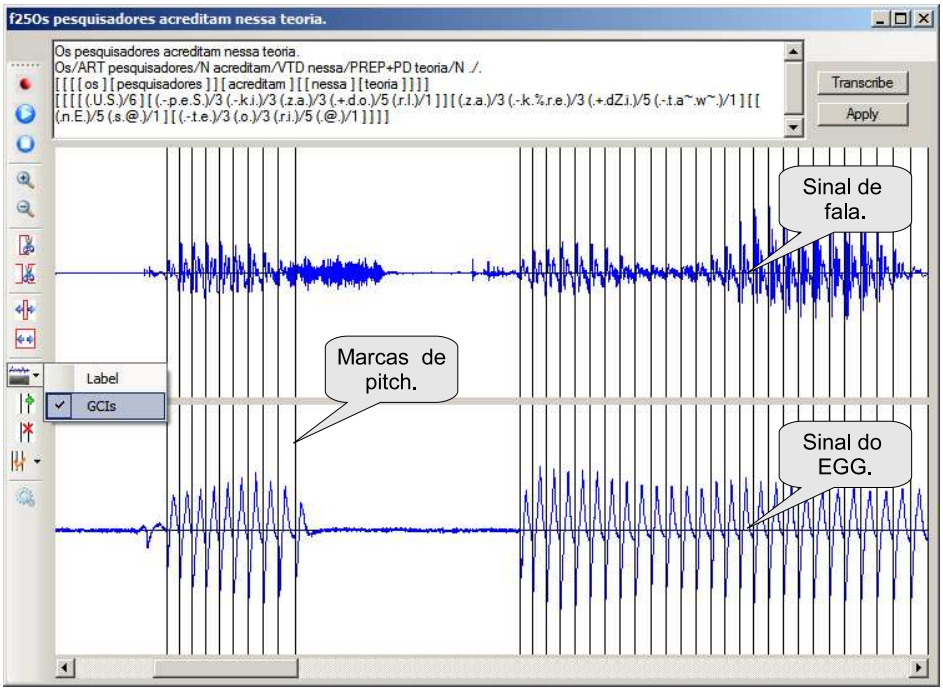


Figura 6.4: Tela de edição no modo de edição das marcas de pitch.

A inserção das marcas de pitch é feita automaticamente a partir do sinal EGG ou do sinal do microfone de contato. No entanto, as marcas inseridas erroneamente podem ser apagadas e novas marcas podem ser inseridas manualmente, com o *mouse* ou por teclas de atalho.

6.3.2 Implementação

As funções de visualização são implementadas por duas classes, nomeadas CPlot e CSPlot, que são utilizadas para desenhar todas as figuras, e traçar todos os gráficos, mostrados ao longo do trabalho. No caso de uma implementação futura multiplataforma, as classes CPlot e CSPlot deverão ser adaptadas por utilizarem funções gráficas nativas do Windows.

Na figura 6.5 é mostrado um diagrama simplificado das classes envolvidas na implementação das ferramentas de edição. Os métodos e atributos completos são mostrados na figura D.1 no Apêndice D. As quatro classes CWave, CMarks, CLabelMarks e CSentence encapsulam, respectivamente, os sinais de áudio, as marcas de pitch, as etiquetas de fronteira dos segmentos e a transcrição do texto.

Na seção 6.3 foi descrito que para cada sentença da lista de sentenças eram gravados e lidos quatro arquivos que contêm toda a informação necessária para a análise da sentença. Os dados contidos nestes arquivos correspondem aos dados encapsulados pelas classes CWave, CMarks, CLabelMarks e CSentence.

Associada à classe CWave está a classe CAudio, responsável por executar a interface entre os aplicativos e a placa de áudio. Esta classe realiza as funções de gravação, reprodução, leitura e escrita em disco dos arquivos de áudio.

A classe CPitchMarker implementa a função de detectar as marcas de pitch no sinal do EGG ou do microfone de contato, e repassá-las para a classe CMarks.

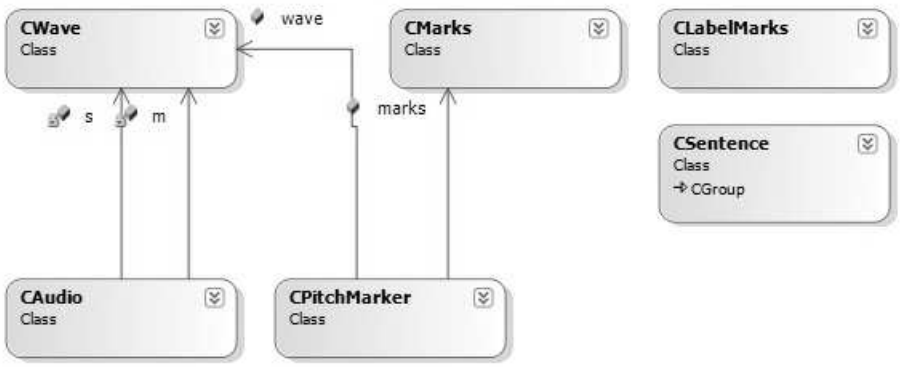


Figura 6.5: Diagrama simplificado das Classes que implementam as ferramentas de edição.

Após a etiquetagem da sentença a classe CLabelMarks contém todos os instantes de tempo das etiquetas, porém não contém a identidade das etiquetas. A classe CSentence, que será vista a seguir na seção 6.4, é a classe que contém a estrutura hierárquica do texto e a transcrição fonética da sentença. Assim, é por um método de associação entre as classes CLabelMarks e CSentence que a identidade das etiquetas são determinadas conforme a transcrição fonética.

6.4 O processamento do texto

O processamento de texto foi incluído em um aplicativo a parte para que as regras e os algoritmos aplicados ao longo do processo pudessem ser verificados em diferentes casos e não só para as sentenças listadas para análise. Entretanto, as funcionalidades do processamento do texto são usadas em diferentes telas com o usuário e no sistema protótipo TTS.

6.4.1 Funcionalidades

Na figura 6.6 é mostrada a tela que executa as funções do processamento do texto. O espaço de texto mostrado no topo da figura contém o texto de entrada, que pode

ser escrito ou importado de um arquivo texto com caracteres no formato UTF-8. No outro espaço de texto da figura está representada a estrutura de dados resultante, em um formato textual.

Em [71] os autores apresentam um módulo de processamento do texto para o português brasileiro adaptado para o sistema MARY (*modular architecture for research on speech synthesis*) [72], onde a representação da estrutura de dados é feita por arquivos XML. Nesta tese, a construção da estrutura hierárquica de dados permite facilmente a representação no formato XML, pela forma de organização dos dados em árvore. No entanto, os arquivos XML são adequados para a interface entre sistemas, mas não são formas muito agradáveis de visualização, por serem demasiadamente extensos. Por esta razão, optou-se por representar a estrutura hierárquica na forma textual, por ser mais simples de ser observada.

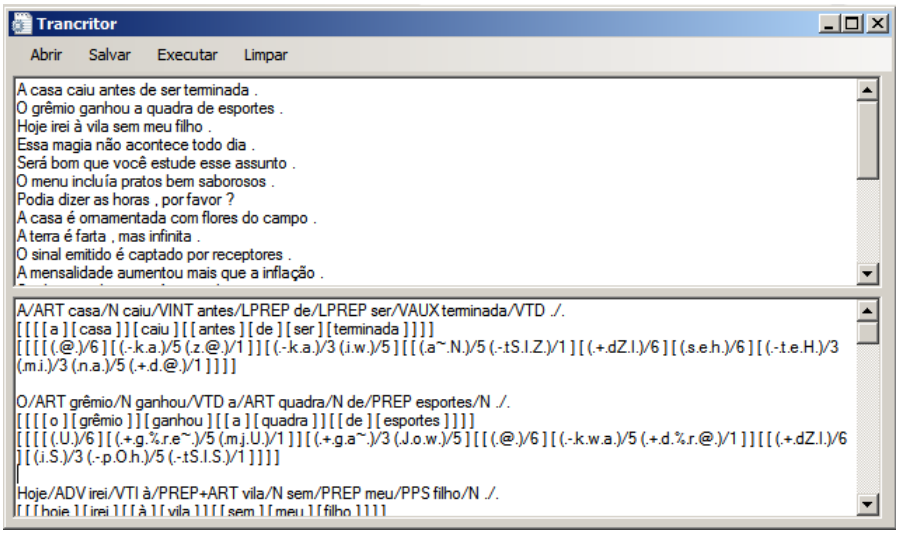


Figura 6.6: Tela do aplicativo de processamento do texto.

Nesta forma de representação textual, a estrutura hierárquica é representada em três linhas de texto. Na primeira linha é mostrado o resultado da classificação morfossintática, onde as *tags* POS são inseridas após a palavras. Na segunda linha, é mostrado o resultado do *parser* da sentença e do agrupamento de palavras. Na última linha é representada a transcrição fonética, onde os símbolos que identificam os fonemas aparecem agrupados por sílabas seguidas da indicação de tonicidade.

Pela importância do tópico de processamento do texto ao longo desta tese, os conceitos envolvidos na aplicação dos algoritmos que executam as etapas do processamento de texto foram apresentados isoladamente no Capítulo 2. Assim, na Seção 6.4.2 a seguir, será descrito como a estrutura hierárquica de dados foi concebida. Em seguida, na Seção 6.4.3, será apresentado brevemente como os algoritmos do processamento de texto são aplicados a esta estrutura.

6.4.2 Implementação da estrutura hierárquica de dados

A implementação da estrutura hierárquica de dados foi feita por listas duplamente encadeadas em diferentes níveis. A estruturação por níveis significa que uma lista de elementos de um determinado nível é também um elemento da lista do nível superior. A vantagem desta forma de implementação é permitir a estruturação dos dados em um formato de árvore, com vários ramos, e com uma quantidade de níveis indeterminada. A desvantagem está na necessidade de um procedimento de busca mais complexo em termos computacionais.

A implementação das listas é feita conforme o diagrama de classes simplificado mostrado na figura 6.7. O diagrama completo das classes é mostrado na figura D.2 no Apêndice D. Todos os elementos da estrutura são modelados na classe CElement e a classe CList implementa a lista encadeada de instâncias da classe CElement. Além disso, para que uma lista seja um elemento de uma lista de nível superior, a classe CList herda as propriedades da classe CElement.

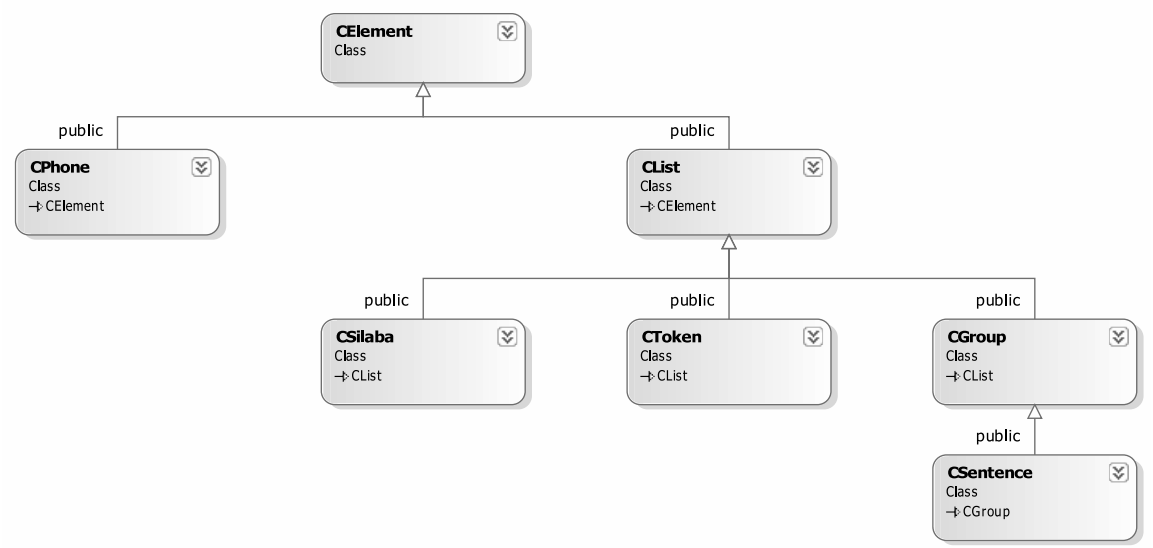


Figura 6.7: Diagrama de classes da estrutura hierárquica montada por listas encadeadas em diferentes níveis.

A classe CPhone representa o nível mais baixo da hierarquia, herdando somente as características de um elemento CElement. As classes referentes às palavras, sílabas, grupos e sentenças foram nomeadas respectivamente de CToken, CSilaba, CGroup e CSentence, e herdam as funcionalidades de CList. A classe CSentence é a lista de mais alto nível, representando o topo da árvore, enquanto que a classe CPhone representa os nós terminais.

As listas são implementadas conforme os princípios básicos de listas duplamente encadeadas, permitindo remover ou inserir elementos, buscar à esquerda ou à direita, entre outras funções típicas. Na figura 6.8 é mostrada uma ilustração da montagem

hierárquica das listas, onde cada bloco representa um elemento da lista e os diferentes tons de cinza indicam níveis diferentes.

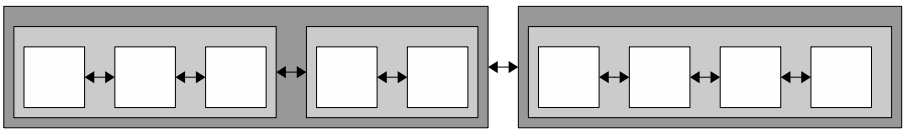


Figura 6.8: Exemplo das listas dispostas em níveis hierárquicos.

Além das funções básicas, outras funções foram implementadas para permitir a modificação da estrutura ao longo dos algoritmos do processamento do texto. A seguir serão apresentados os conceitos básicos das funções de obtenção da referência de início e fim de um determinado nível, de busca por um elemento, de mover um elemento para uma lista vizinha, de divisão e de agrupamento de listas.

Referência de início e fim

No formato adotado de listas inseridas em listas, a sentença é a única referência disponível imediatamente. A consulta por qualquer outro elemento da estrutura precisa ser feita por busca, tendo a referência da sentença como ponto de partida. Considere o exemplo em que se deseja consultar a primeira sílaba de uma sentença. Neste caso, primeiramente obtém-se acesso à primeira palavra e em seguida o acesso à primeira sílaba desta palavra. Do mesmo modo, o acesso à última sílaba é obtido pelo acesso à última palavra e em seguida o acesso à última sílaba da palavra. A partir de um elemento de início ou fim, é possível percorrer todos os elemento de um mesmo nível pelo procedimento de busca descrito a seguir.

Busca

Considere agora que se deseja percorrer todas as sílabas da sentença, da esquerda para a direita. Neste caso, é obtida a referência para a primeira sílaba, tal como descrito acima, e move-se para a próxima sílaba da lista. Porém, ao percorrer as sílabas da primeira palavra, chega-se naturalmente à última sílaba desta palavra e não há um próximo elemento nesta lista de sílabas, pois as listas são terminadas. Assim, foi preciso criar um procedimento de busca onde um elemento terminal de uma lista verifica se seu nível superior possui um elemento seguinte. No caso do exemplo, equivale a buscar a sílaba inicial da palavra seguinte.

Na figura 6.9 é ilustrado um exemplo de busca pelos elementos de mais baixo nível da hierarquia. A seta em escuro indica o caminho percorrido entre os elementos, sendo necessária a transição entre níveis, tal como descrito. A figura 6.9(a) ilustra a busca para frente (*forward*), da esquerda para direita, enquanto a

figura 6.9(b) ilustra a busca para trás (*backward*).

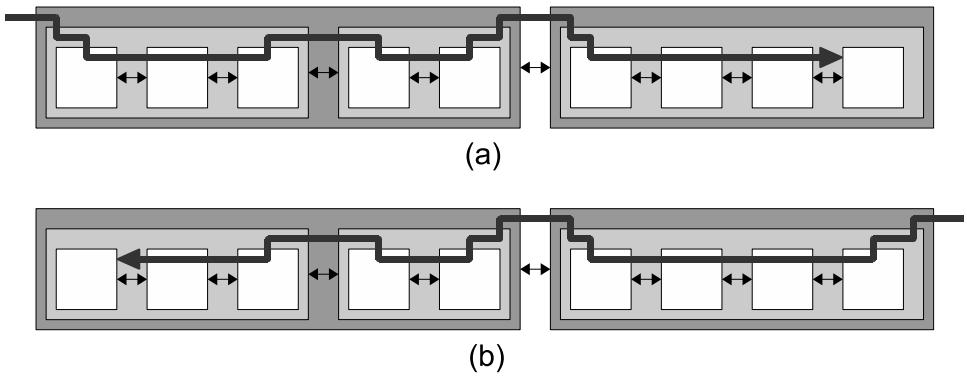


Figura 6.9: Ilustração do procedimento de busca pelos elementos de nível mais baixo na hierarquia. Em (a) a busca para frente e em (b) a busca para trás.

Mover à direita e à esquerda

Nas etapas de processamento de texto, principalmente na junção de palavras, é recorrente a necessidade de deslocar um elemento de uma lista para outra, no mesmo nível. Por exemplo, quando uma palavra termina por consoante e a palavra seguinte inicia por vogal, a consoante se junta com a vogal da palavra seguinte formando uma única sílaba. Assim, foi necessário implementar rotinas que movessem os elementos entre listas de mesmo nível. Na figura 6.10 são ilustrados os procedimento de mover à direita 6.10(a) e mover à direita 6.10(b).

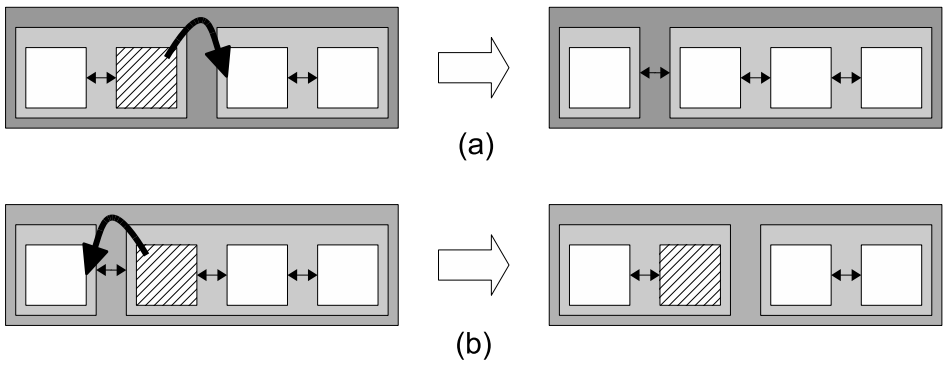


Figura 6.10: Ilustração do procedimento de mover a direita (a) ou à esquerda (b).

Divisão

Outra funcionalidade recorrente é a divisão de uma lista, usada principalmente na separação sílabica. Neste caso, um elemento da lista é indicado como ponto de corte, que dará origem a duas listas conectadas, de mesmo nível. Na figura 6.11 é ilustrado um exemplo desta divisão.

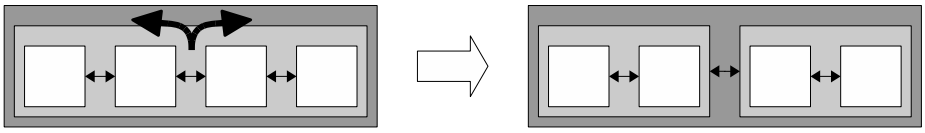


Figura 6.11: Ilustração da divisão da lista.

Agrupamento

Um último caso que merece atenção é quando elementos são agrupados formando um elemento de nível superior. Esta funcionalidade é usada fundamentalmente quando palavras são agrupadas formando um grupo, em um nível superior. Este é o caso, por exemplo, de palavras compostas que embora sejam palavras separadas por hífen, possuem o comportamento de um grupo coeso. Na figura 6.12 é ilustrado um exemplo.

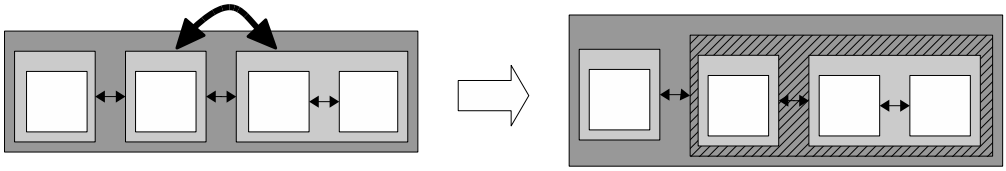


Figura 6.12: Ilustração do agrupamento de elementos de um nível para um nível superior.

6.4.3 Implementação do processamento do texto

Dentre as etapas do processamento de texto, a adaptação do classificador morfos

Após o *parser* da sentença a classe CToken ocupa o nível mais baixo na estrutura. Então, inicia-se a etapa de transcrição onde cada elemento CToken é invocado por CSentence a obter a transcrição fonética. Após a conversão de todos os elemento da classe CToken, os níveis hierárquicos abaixo de CToken são preenchidos, ou seja, CToken é preenchido por elementos da classes CSilaba, que por sua vez são preenchidas por elementos da classe CPhone. Assim, os elemento de nível mais baixo da estrutura passam a ser da classe CPhone.

Por fim, a estrutura hierárquica se apresenta totalmente preenchida com as informações obtidas do texto. Na seção 6.3 foi visto que a classe CSentence é associada com a classe CLabelMarks para fornecer a identidade das etiquetas. Neste caso, cada etiqueta da classe CLabelMarks recebe a identidade de um elemento CPhone da classe CSentence. Ao mesmo tempo, um elemento CPhone recebe a referência temporal da etiqueta, vinculando a estrutura de dados com o sinal. Na seção a seguir será visto que com este sincronismo estabelecido, os elementos da estrutura hierárquica passam a ser preenchidos com as informações obtidas do sinal.

6.5 Ferramentas de edição da prosódia

As ferramentas de edição de prosódia codificam os valores das variáveis prosódicas de duração, pitch e intensidade, obtidas no domínio do tempo, para o domínio dos parâmetros. De forma inversa, estes parâmetros são também decodificados para determinar o comportamento destas variáveis no domínio do tempo. Estes dois procedimentos de codificação e decodificação foram nomeados no Capítulo 4 de análise e síntese das variáveis prosódicas. Ao longo do Capítulo 4 as figuras utilizadas como exemplo foram geradas por estas ferramentas.

6.5.1 Funcionalidades

Na figura 6.13 é mostrada a tela de ferramentas de edição de prosódia. Esta tela é usada tanto para análise quanto para a síntese.

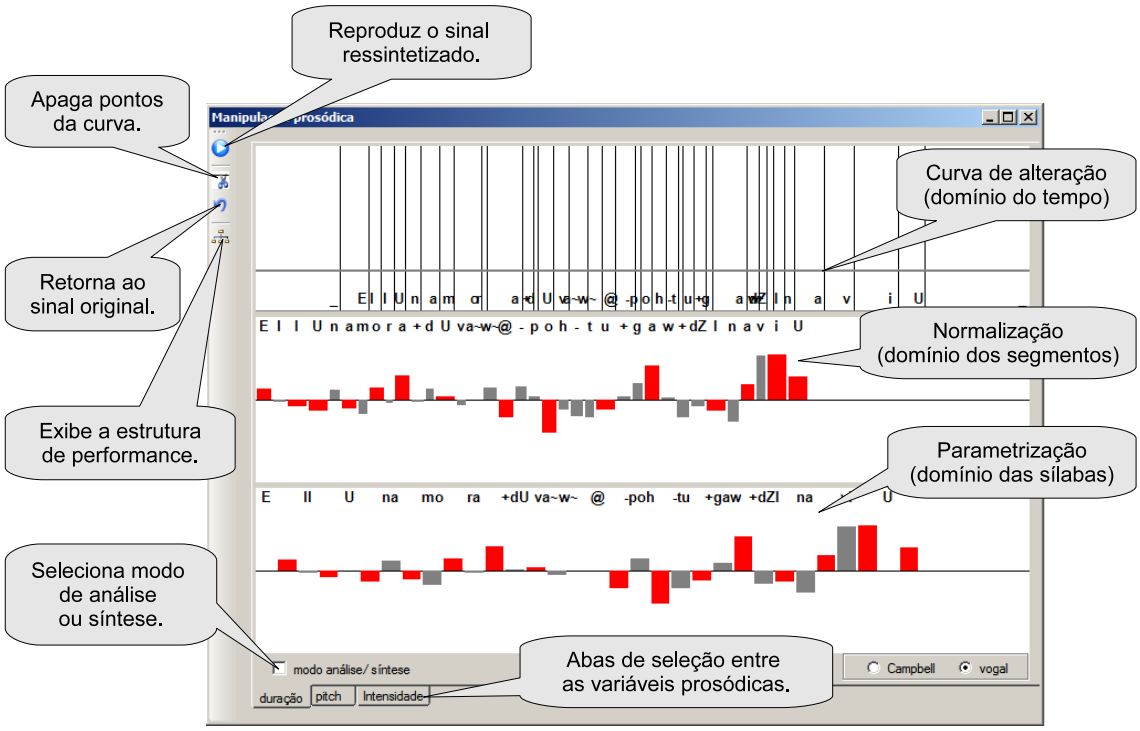


Figura 6.13: Tela de ferramentas de edição de prosódia.

Quando a tela é aberta, a etapa de análise é feita automaticamente. Nesta etapa, as curvas de alteração das variáveis no domínio do tempo, mostradas no primeiro gráfico, podem ser modificadas manualmente, o que irá modificar automaticamente os valores dos parâmetros dos segmentos e das sílabas. Do mesmo modo o valor dos parâmetros dos segmentos, mostrados no gráfico do meio, podem ser alterados com um clique do *mouse* sobre a representação dos parâmetros.

Ao selecionar o modo de síntese no *checkbox* na parte inferior da tela, as variáveis prosódicas são sintetizadas a partir dos parâmetros da sílaba. Neste modo de

síntese, somente os parâmetros das sílabas, mostrados no gráfico inferior, podem ser alterados manualmente, o que irá alterar automaticamente os parâmetros dos segmentos e a curva de alteração das variáveis no domínio do tempo.

Nesta mesma tela são exibidas as variáveis de duração, pitch e intensidade, que podem ser escolhidas por meio das abas de seleção no inferior da tela, conforme indicado na figura 6.13.

6.5.2 Implementação

Após o processamento do texto, as informações extraídas do texto estão contidas na forma de atributos das classes que formam a estrutura hierárquica. Os parâmetros prosódicos dos segmentos e das sílabas também são representados como atributos das classes. Assim, as classes reúnem atributos que são relativos tanto à informações do texto quando informações do sinal.

Foi visto nas seções anteriores, que um método de associação entre as classes CLabelMarks e CSentence, sincroniza a estrutura hierárquica com as referências de tempo das etiquetas de fronteira dos segmentos. Assim, é por meio destas referências temporais que o atributo de duração nominal dos elementos da classe CPhone é determinado. Do mesmo modo, as classes CWave e CMarks também possuem métodos de associação com a classe CSentence, que determina, respectivamente, os atributos de intensidade e contorno de pitch aos elementos da classe CSentence.

De posse destes valores nominais, as classes são autônomas no procedimento de análise, ou seja, quando os atributos dos valores nominais são atribuídos às classes, estas classes automaticamente codificam estes valores nominais para atributos na forma de parâmetros. Os valores mostrados nos gráficos da figura 6.13 nada mais são do que representações dos atributos das classes CPhone e CSilaba.

Na etapa de síntese, a classe CSilaba determina os parâmetros dos elementos da classe CPhone contidos por ela. É desta maneira que a prosódia da sentença pode ser especificada somente pelos parâmetros prosódicos das sílabas. Conforme descrito no capítulo 4, as alterações nos parâmetros prosódicos das sílabas resultam em curvas de alteração das variáveis prosódicas no domínio do tempo, que serão usadas pelo PSOLA para manipulação dos sinais. Estas curvas são externas aos elementos da classe CSentence.

As curvas de alteração das durações, intensidade e pitch são encapsuladas pela classe CRealPoint, onde um dos atributos desta classe indica ser uma curva do tipo DUR, PITCH ou GAIN. Esta classe CRealPoint é um vetor de pontos formados por um valor e uma referência temporal. Na etapa de síntese, a classe CRealPoint possui métodos que lêem a estrutura hierárquica da classe CSentence e obtém os pontos das curvas, tal como descrito no capítulo 4.

A modificação do sinal é feita pela aplicação das curvas de alteração das durações, pitch e ganho, pelo algoritmo TD-PSOLA. Na figura 6.14 são mostradas as classes envolvidas neste processo. A classe CPsola implementa o algoritmo TD-PSOLA. As classes CWave, CMarks e CLabelMarks são associadas à classe CPsola para fornecerem, respectivamente, o sinal a ser modificado, as marcas de pitch do sinal, e as etiquetas de fronteira dos segmentos para serem corrigidas conforme o acerto de durações. Um método de síntese da classe CPsola recebe as três curvas encapsuladas pela classe CRealPoint e modifica o sinal.

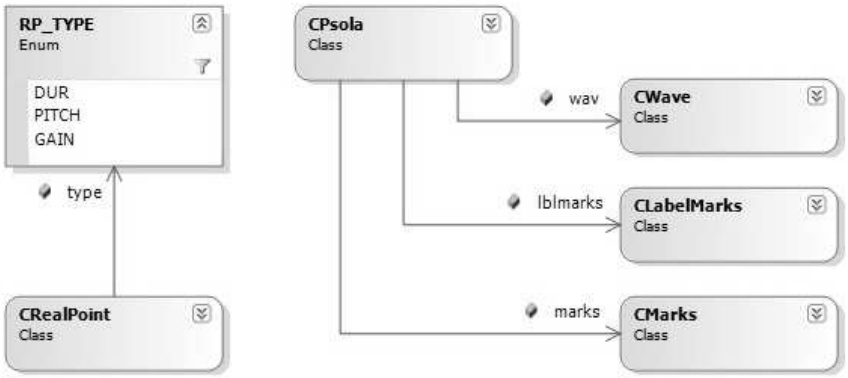


Figura 6.14: Classes envolvidas na alteração das variáveis prosódicas do sinal.

Na seção a seguir será apresentado o protótipo do sistema TTS que utiliza das funcionalidades descritas até o momento.

6.6 O protótipo do sistema TTS

Conforme [1] podemos modelar um sistema TTS em duas etapas principais: uma etapa de processamento do texto, que codifica o texto para uma forma intermediária; e uma etapa de síntese de voz, que decodifica esta forma intermediária para o sinal de voz. O protótipo do sistema TTS desenvolvido nesta tese segue este conceito, onde o processamento de texto é tal como descrito na seção 6.4 e a forma intermediária em questão é a estrutura hierárquica de dados.

Na etapa de síntese de fala, a técnica implementada utiliza um banco reduzido de unidades de concatenação e o método TD-PSOLA para a manipulação do contorno de pitch, durações e intensidade.

Um banco de unidades de concatenação vem sendo construído com o auxílio de ferramentas que foram aprimoradas a partir do trabalho de mestrado [10] e serão descritas a seguir.

6.6.1 Funcionalidades

A principal funcionalidade do sistema protótipo TTS é permitir que a conversão texto-fala seja observada em duas etapas. A tela referente ao sistema protótipo TTS é muito similar à tela de edição dos sinais, porém, somente com os recursos de visualização. Na figura 6.15 é mostrado um exemplo da conversão texto-fala da sentença "Renata jogava". A sentença foi escrita na entrada de texto e em seguida foi executado o processamento de texto.

A montagem da estrutura hierárquica de dados, resultante do processamento do texto, está representada na forma textual, tal como descrita na seção 6.4.

Em seguida, foi executada a síntese a partir da estrutura hierárquica. Os gráficos da figura 6.15 mostram, respectivamente, o sinal proveniente da concatenação das unidades e o espectrograma deste sinal. Além disso, no primeiro gráfico estão representadas as marcas de pitch e no segundo gráfico são mostradas as etiquetas de fronteiras.

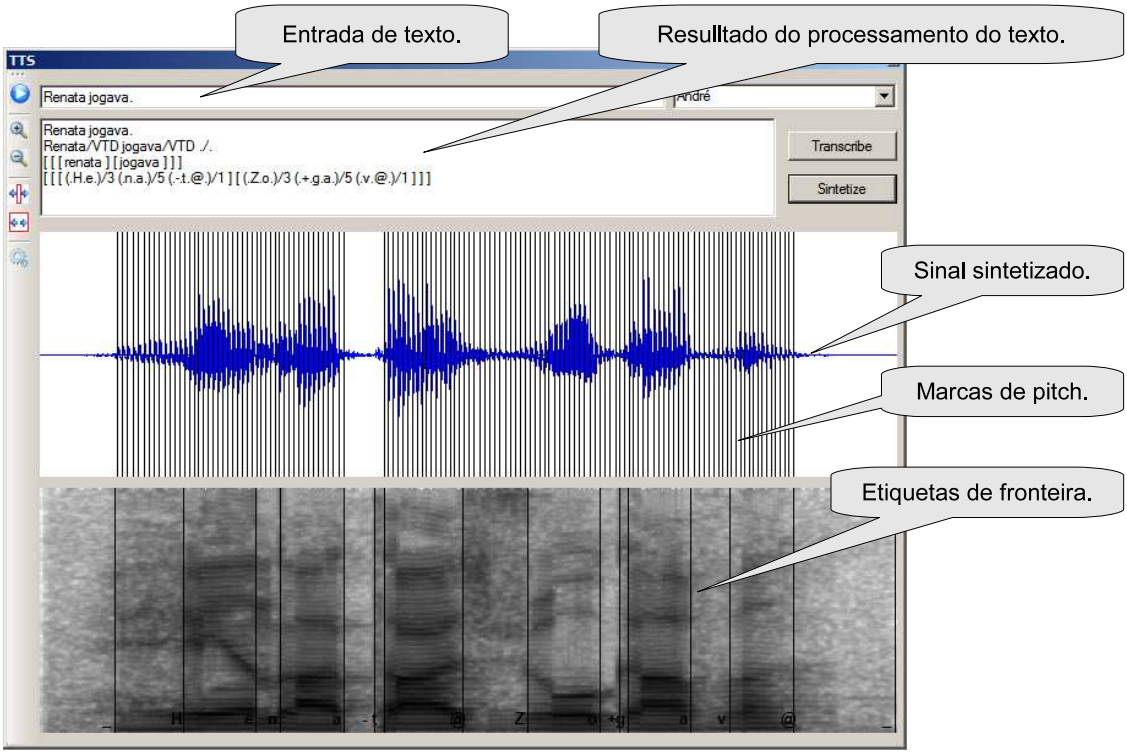


Figura 6.15: Tela do sistema protótipo TTS.

O último botão do menu de ferramentas, à esquerda da tela, tem a função de abrir a tela de manipulação prosódica, permitindo que o usuário aplique ao sinal alterações na prosódia.

6.6.2 A construção do banco de unidades

Para a construção do banco de unidades, o sistema SAPRO é utilizado em um modo alternativo. Neste modo alternativo, as sentenças ou logotomas de onde serão recortadas as unidades são organizados previamente em uma lista, tal como anteriormente. Porém, a leitura desta lista é feita por outro ítem do menu, para indicar que as funcionalidades adicionais de construção do banco de unidades deverão estar disponíveis.

A construção do banco de unidades é feita automaticamente a partir de um arquivo de definições das unidades. Neste arquivo de definições é gravado em cada linha um código de identificação da unidade, o nome do arquivo onde a unidade se encontra e os instantes de tempo que delimitam a unidade no sinal.

Na figura 6.16 é mostrada a tela do sistema SASPRO, onde o sub-ítem Database do menu foi expandido. Ao abrir a lista de sentenças ou logotomas, o arquivo de definições também é aberto automaticamente, e a lista de sentenças do lado esquerdo aparece dividida em duas. Na lista superior estão as sentenças, tal como anteriormente. Porém, na lista inferior são mostrados as unidades que já foram definidas. Ao selecionar uma unidade da lista, a tela de edição é aberta conforme o nome do arquivo indicado na definição da unidade.

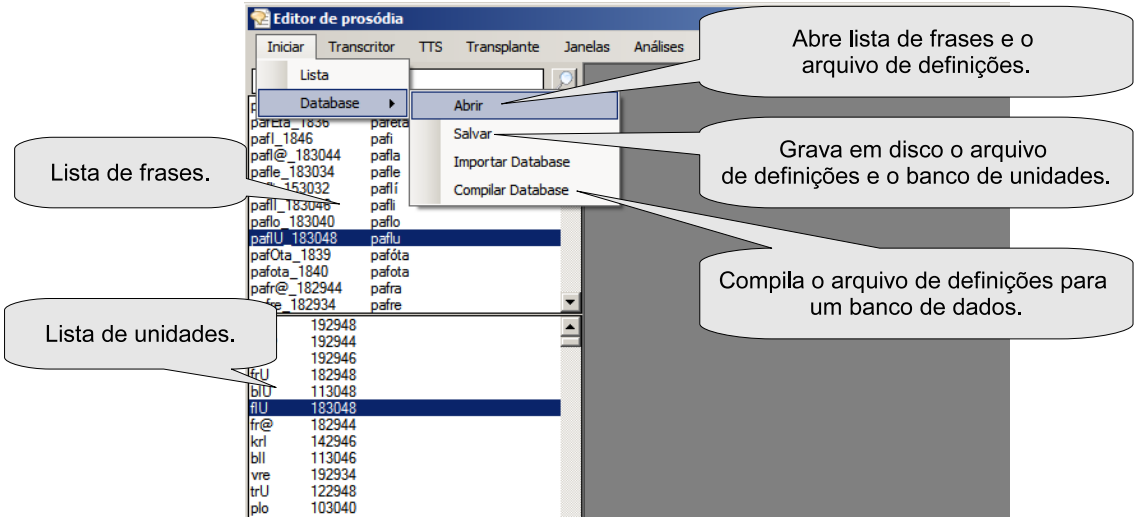


Figura 6.16: Tela da sistema quando em modo de Edição das unidades.

Para a definição de uma nova unidade ainda não definida, abre-se a tela de edição de uma sentença. A gravação e edição dos sinais, assim como a inserção das marcas de pitch e a etiquetagem decorrem normalmente. Então, para definir uma unidade o usuário seleciona um trecho do sinal que equivale à unidade, e requisita que a unidade seja adicionada ao arquivo de definições.

Após a definição manual das unidades, o banco de dados é construído pela compilação deste arquivo de definições. Esta função de compilação abre os arquivos

indicados e recorta a unidade conforme os limites definidos. Assim, as amostras do sinal, as marcas de pitch e as etiquetas incluídas nestes limites são inseridas no banco de unidades.

6.6.3 implementação

Na figura 6.17 é mostrado um diagrama simplificado das classes envolvidas na preparação e concatenação das unidades. As unidades são encapsuladas pela classe CUnit que são inseridas, removidas e buscadas no banco de unidades pela classe CUnitData. A classe CUnitData é responsável pela leitura e escrita em disco do banco de unidades, assim como pela compilação do arquivo de definições. A estrutura TUnitDef contém as informações que definem a unidade, ou seja, um código de identificação, o arquivo onde a unidade é localizada e os limites da unidade no sinal.

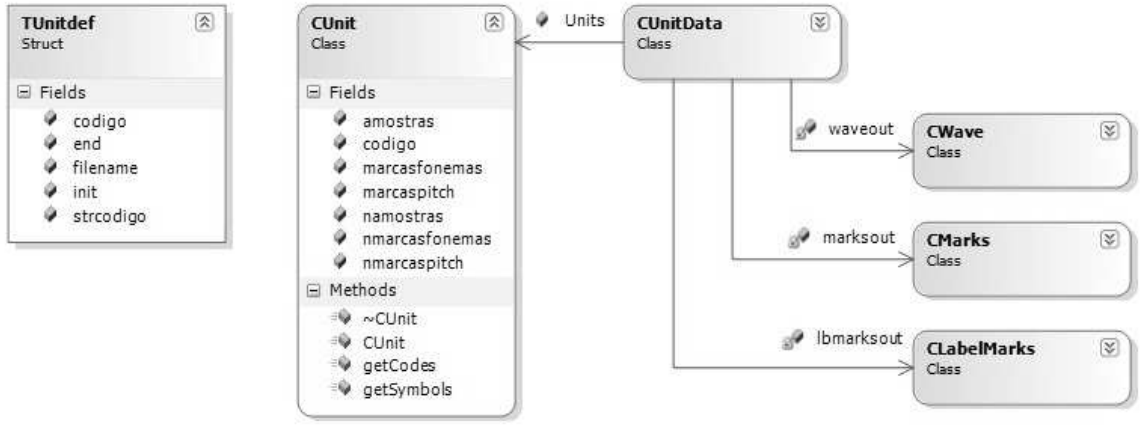


Figura 6.17: Diagrama das classes envolvidas na preparação e concatenação das unidades.

Atualmente, o inventário de unidades usado neste trabalho, baseado em [10], é composto principalmente por difones, contendo alguns trifones para segmentos que são muito sensíveis ao contexto. No total tem-se cerca de 500 unidades adicionadas ao banco de unidades de 3 MB.

A classe CUnitData possui ainda a função de executar a concatenação das unidades a partir da estrutura de dados contida na classe CSentence. A busca pelas unidades a serem concatenadas é feita de modo que se uma unidade desejada não existir são aplicadas regras de substituição para uma unidade mais próxima. Por exemplo, na concatenação da palavra “peixe” [pej.SI], os difones [ej] [jS] são substituídos por [eI] [iS]. Ocorrem casos em que a substituição resulta em uma concatenação ruim, porém é mais razoável do que não utilizar nenhuma unidade. Este problema vem sendo minimizado com a inserção de novas unidades.

Como resultado da concatenação, são retornadas as amostras do sinal concatenado, as marcas de pitch e as marcas de fronteiras fonéticas. Este trio de elementos são encapsulados, respectivamente, pelas classes CWave, CMarks e CLabelMarks, como mostrado na figura 6.17. Assim, a concatenação do sinal resulta nos mesmo componentes de um sinal obtido por gravação, e portanto, o sinal sintetizado pode ser tratado tal como um sinal obtido por gravação.

Por fim, o sistema protótipo TTS é na verdade a conjunção das classes utilizadas até o momento. Conforme o conceito adotado de modelar um sistema TTS em duas etapas, a etapa de processamento de texto, incluindo a montagem da estrutura hierárquica, é feita tal como descrita na seção 6.4. Na etapa de síntese da fala, a concatenação das unidades é feita conforme descrito acima e a manipulação prosódica do sinal concatenado é feita tal como apresentado na seção 6.5 anterior.

6.7 Ferramentas de transplante de prosódia

As ferramentas de transplantes executam o transplante de prosódia por alinhamento temporal, descrito no capítulo 3, assim como as modalidades do transplante paramétrico, descritas no capítulo 5. Na figura 6.18 é mostrada a tela usada para o transplante. No gráfico superior é desenhado o sinal de teste e no gráfico inferior o sinal de referência. O sinal de teste é selecionado na lista de sentenças do sistema SASPRO, e o sinal de referência pode ser gerado pelo TTS protótipo ou lido também da lista de sentenças, no espaço indicado na figura por "seleção do sinal de referência".

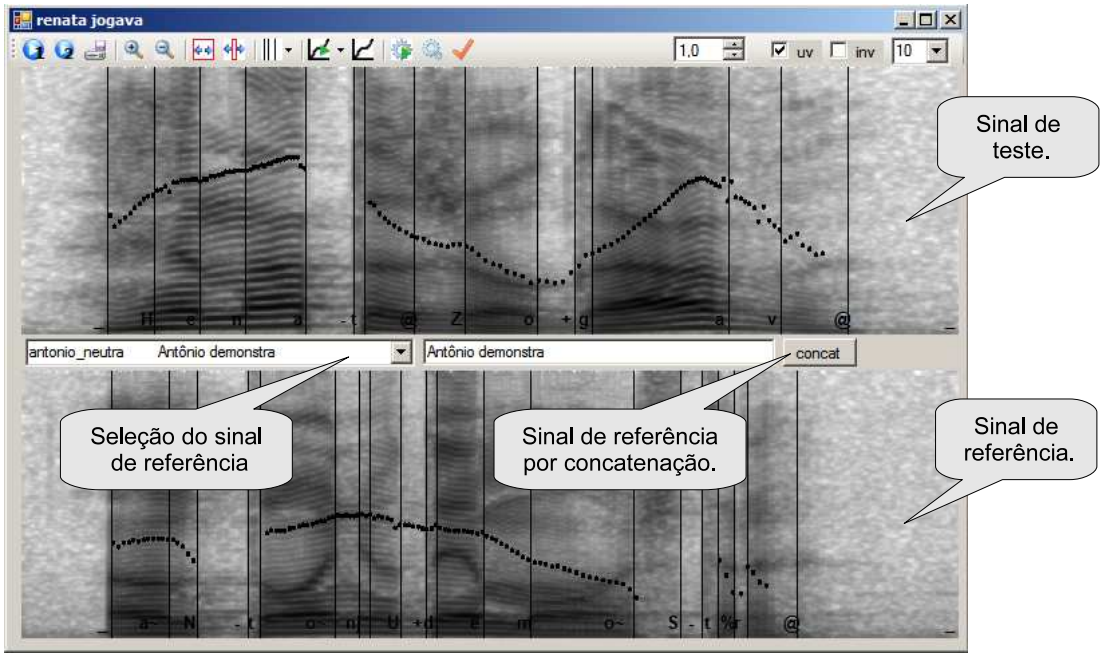


Figura 6.18: Tela de ferramentas de transplante.

Na figura 6.19 são mostradas as funcionalidades disponíveis nesta tela. Dentre as funcionalidades descritas na figura, nas opções de transplante, são mostrados quatro possibilidades de transplante:

- No item DTW o transplante é feito por alinhamento temporal, na forma convencional.
- No item PSDTW o transplante é executado de maneira síncrona com o pitch, tal como descrito no Capítulo 3.
- No item "por Cópia" é executado o transplante paramétrico por cópia. Neste caso os parâmetros das sílabas do sinal de teste são copiados para a sentença de referência.
- No item "por Superposição", o transplante é feito por superposição, onde antes da execução do transplante é preciso primeiro definir o modelo, que pode ser feito de duas formas:
  - "por diferença", quando o modelo é estabelecido por comparação dos parâmetros da sentença de teste com os parâmetros da sentença de referência.
  - "por média", quando o modelo é estabelecido com referência aos parâmetros do modelo de prosódia neutra médio, determinado previamente. Neste caso, não há necessidade do sinal de referência.

Na opção de transplante por superposição, após determinado o modelo, qualquer sentença de mesma estrutura silábica pode ser selecionada ou gerada como referência para a aplicação do modelo.

Figura 6.19: Funcionalidades da tela de transplante.

Estas funções de transplante são executadas por uma única classe CAligner. O diagrama de classes envolvidas no transplante está representado na figura D.4.

Somente na modalidade de transplante por alinhamento temporal, esta classe se associa a uma classe CFeatures, que é responsável por obter, dos sinais encapsulados na classe CWave, os vetores de parâmetros representativos do espectro. A partir deste vetor de parâmetros, a classe CAligner executa o alinhamento por DTW determinando o caminho de alinhamento entre os sinais. No caso do transplante simples, por DTW, as curvas de alteração das durações, ganho e pitch, representadas pela classe CRealPoint, são determinadas através do melhor caminho de alinhamento. Dentre as funções mostradas na figura 6.19, estas curvas podem ser vistas na tela de manipulação prosódica. No caso do transplante pelo método PSDTW, o procedimento de *Overlap and Add* é feito diretamente, e portanto, não são geradas as curvas de alteração.

No transplante paramétrico, o procedimento é realizado somente entre duas instâncias da classe CSentence. Neste caso os parâmetros são transferidos entre sílabas: das sílabas da sentença de teste para as sílabas da sentença de referência. Em seguida, basta a execução do procedimento de síntese da classe CSentence que irá gerar as curvas de alteração. Tal como mostrado no procedimento de síntese da prosódia descrito na seção 6.5.

6.8 Conclusão

Neste capítulo foram apresentadas as ferramentas realizadas ao longo deste trabalho, incluindo a descrição das funcionalidades e dos conceitos usados para a implementação. Este capítulo foi escrito na intenção de auxiliar o uso futuro destas ferramentas e também orientar a continuidade deste trabalho. Neste caso, a construção dos aplicativos de forma orientada a objetos é um grande facilitador para futuras modificações.

Uma grande parcela do trabalho dedicado a esta tese precisou ser dirigida à implementação do sistema SASPRO. O objetivo de construir aplicativos que não só executam procedimentos, mas que permitem a iteração do usuário nas diferentes etapas do processo, foi um dos maiores responsáveis por esta dedicação.

Em particular, foram descritas as ferramentas de edição e visualização dos sinais de áudio, das marcas de pitch e da etiquetagem das fronteiras fonéticas dos segmentos. As ferramentas de processamento de texto foram também apresentadas, onde foi dado enfoque à concepção da estrutura hierárquica de dados. Por meio de funções específicas desta estrutura hierárquica foi possível basear os algoritmos de processamento de texto na movimentação e transformação dos elementos desta estrutura. Além disso, a sincronização da estrutura hierárquica de dados com o sinal de fala, permitiu reunir as informações a respeito da estrutura do texto com as variáveis prosódicas dos sinal.

Foram descritas ainda as ferramentas de manipulação prosódica e o sistema TTS protótipo, onde foi apresentado o procedimento adotado de construção do banco de unidades.

Por fim, foram apresentadas as ferramentas usadas para a aplicação das diferentes formas de transplante de prosódia apresentadas neste trabalho.

Capítulo 7

Conclusão

7.1 Considerações finais

Ao longo desta tese foram apresentadas as etapas de desenvolvimento de um sistema TTS, cujo esquema geral é visto na figura 7.1.

Mais especificamente, a etapa de processamento de texto, descrita no Capítulo 2, inclui um classificador morfossintático e a transcrição fonética automática, obtida pela montagem de uma estrutura hierárquica de dados. As regras de obtenção da transcrição fonética, descritas em trabalhos anteriores, foram modificadas para serem aplicadas em diferentes etapas. Com este procedimento foram apresentados algoritmos alternativos para a separação silábica e determinação de tonicidade das sílabas, basedas no domínio dos fonemas. Além disso, foram implementadas regras pós-lexicas que consideram os processos envolvidos na junção de palavras.

Dando inicio ao estudo da prosódia, no Capítulo 3 foi descrito o procedimento de transplante de prosódia, onde foi proposta a conjugação dos algoritmos DTW e PSOLA, resultando num esquema nomeado de PSDTW-OLA. Este algoritmo proposto otimiza a segmentação inerente ao método DTW utilizando o janelamento típico do PSOLA, centrado nas marcas de pitch de um sinal de voz. O resultado é um melhor casamento das etapas de alinhamento temporal e mapeamento de pitch de um sinal de análise para o sinal de síntese, como desejado no problema de transplante de prosódia.

Com o propósito de estudar o comportamento das variáveis prosódicas no domínio das sílabas e abstrair-se do conteúdo segmental das sílabas, no Capítulo 4, as variáveis de duração e intensidade dos segmentos foram normalizadas a partir da distribuição estatística destas variáveis ao longo de um *corpus* de análise construído para este propósito. Na etiquetagem das sentenças deste *corpus*, e de todas as sentenças estudadas nesta tese, foi imposta a associação direta das etiquetas com a transcrição fonética automática do texto das sentenças. Isto permitiu vincular as

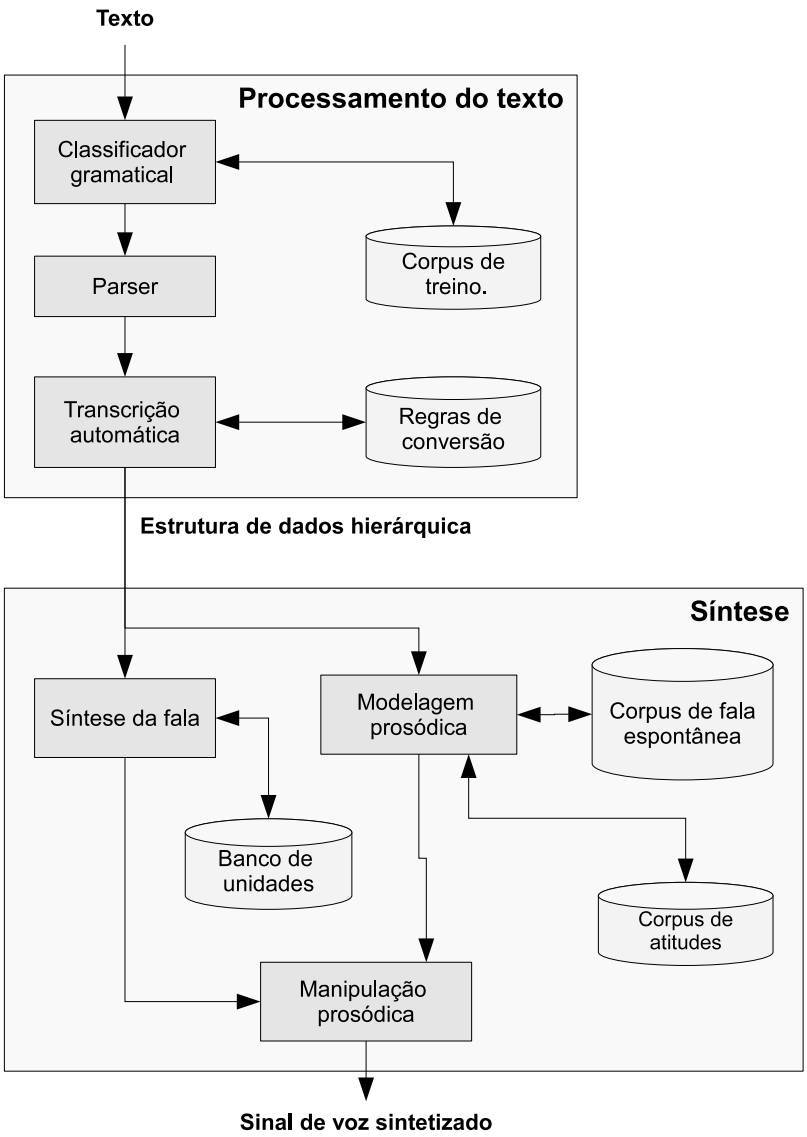


Figura 7.1: Ilustração do sistema TTS.

diferentes estruturas linguísticas do texto com o sinal de fala.

Na etapa de síntese mostrada na figura 7.1, a síntese da fala executa a concatenação das unidades a partir de um banco de unidades que vem sendo construído. Além do sinal concatenado, a síntese da fala retorna as marcas de pitch do sinal assim como as etiquetas de fronteira dos segmentos. Deste modo, a manipulação prosódica do sinal concatenado é feita tal como num sinal natural gravado e etiquetado. Ao longo do Capítulo 4 a modelagem das variáveis prosódicas no domínio da sílaba foi proposta tendo em vista a síntese das atitudes. Neste caso foi necessário adaptar o modelo de CAMPBELL para capturar separadamente os efeitos de alongamento das consoantes em *onset* e das vogais da sílabas. Foi ainda proposto um método para a codificação das intensidades.

Dando prosseguimento ao estudo da prosódia, no Capítulo 5 foram apresentados

mecanismos de transplante das atitudes a partir da codificação das variáveis prosódicas. Além disso, foi demonstrado ao longo do capítulo que a parametrização das variáveis no domínio da sílaba oferecem um meio simples de observar e caracterizar novas atitudes.

Conclui-se portanto que dentre as diversas contribuições realizadas ao longo desta tese a contribuição geral consiste na descrição das ferramentas de desenvolvimento do sistema de conversão texto-fala, incluindo as ferramentas de análise e síntese da prosódia, como descrito ao longo do Capítulo 6.

De fato, enquanto os estudo descritivos da prosódia, no campo da linguística, observam normalmente as variáveis prosódicas na sua forma básica, no domínio do tempo, os sistemas TTS atuais buscam suprimir a etapa de descrição da prosódia, utilizando procedimentos de aprendizagem automática. Assim, o sistema de desenvolvimento realizado nesta tese busca aproximar estes extremos, permitindo ao mesmo tempo observar e manipular as variáveis prosódicas em diferentes domínios e abordar a descrição da prosódia sob um ponto de vista de um vetor de parâmetros vinculados à organização silábica do texto.

Algumas das tarefas realizadas neste trabalho, tal como a etiquetagem do *corpus*, exigem certo grau de experiência prática que ultrapassa o que é possível ser transmitido por artigos ou livros. No entanto, a experiência adquirida ao longo desta tese tornou possível realizar tais tarefas de maneira a atingir os objetivos propostos. Alguns dos assuntos abordados, exigiu incursões em diversas áreas, desde a linguística e processamento da linguagem natural à engenharia de software. Em alguns casos, para o entendimento de um determinado assunto, foi necessário começar de conceitos simples por livros didáticos básicos até se conseguir alcançar o nível mínimo de entendimento para a adordagem do assunto. Esperamos que esta tese possa abreviar este processo de aprendizagem para os futuros especialistas brasileiros em sistemas TTS.

7.2 Trabalhos futuros

Esta tese não tem a pretensão de esgotar o tema de sistemas TTS, que por si só é uma área extremamente fértil em diversos aspectos. Dentro das linhas principais aqui apresentadas, podemos considerar as seguintes possibilidades de continuidade deste trabalho:

- O aprimoramento do banco de unidades por concatenção: Este processo poderia se dar de duas formas. Inicialmente, complementando o *corpus* atual com a devida gravação das unidades que faltam, o que pode ser determinado por inspeção. Posteriormente, após a validação do *corpus* completado, o trabalho poderia ser estendido pela gravação de um *corpus* de um locutor profissional.

- A geração de modelos prosódicos: Ao longo do Capítulo 3 as variáveis prosódicas foram codificadas no domínio da sílaba de forma independente do conteúdo. No entanto, os parâmetros prosódicos da sílaba dependem de informações a respeito da função e da posição que a sílaba ocupa na estrutura silábica da sentença. Assim, tendo em mãos a estrutura hierárquica que pode fornecer facilmente informações a respeito da organização das sílabas, e de posse dos parâmetros prosódicos de cada sílaba, um prodedimento de aprendizagem automática, por redes neurais por exemplo, pode mapear os dois tipo de informação permitindo gerar modelos de prosódia. Para a estimativa da prosódia neutra, o *corpus* de fala de análise construído neste trabalho poderá ser usado. O mesmo se aplica às atitudes, pois se forem gravadas produções das atitudes em sentenças com diferentes estruturas silábicas, uma rede neural poderá ser usada como um gerador de modelos de atitudes.

- O estudo da decomposição das atitudes em padrões básicos: A parametrização das atitudes permite um novo entendimento deste tipo de informação. Esta nova forma de análise ao mesmo tempo que aumenta o nível de abstração permite uma quantificação de diversos aspectos que antes eram considerados extremamente subjetivos. Desta forma, foi possível se realizar uma análise do grau de similaridade entre diferentes atitudes, como observado na parte final do Capítulo 5. Uma outra extensão desta metodologia de análise seria, assim, a possível identificação de atitudes básicas, que permitiriam a descrição de todas as demais atitudes como uma combinação (possivelmente linear) destas atidudes-padrão.

Referências Bibliográficas

[1] TAYLOR, P. *Text-to-Speech Synthesis*. Cambridge University Press, 2009.

[2] RABINER, L. R., SCHAFER, R. W. *Introduction to digital speech processing. Foundations and trends in signal processing*, v. 1. Boston, Mass: Now, 2007.

[3] DUTOIT, T. *An introduction to text-to-speech synthesis*. Norwell, MA, USA, Kluwer Academic Publishers, 1997.

[4] SAGISAKA, Y., YAMASHITA, T., KOKENAWA, Y. "Speech Synthesis with Attitude". In: *Speech Prosody 2004*, pp. 401–404, 2004.

[5] DE MORAES, J. A. "The Pitch Accents in Brazilian Portuguese: analysis by synthesis". In: *Proceedings of the Fourth Conference on Speech Prosody*, pp. 389–398, 2008.

[6] CAMPBELL, N. "Timing in speech: A multi-level Process." In: *In Prosody: Theory and Experiments*, pp. 281–334, Kluwer Academic Publishers, 2000.

[7] TAYLOR, P., BLACK, A. W., CALEY, R. "Heterogeneous Relation Graphs as a mechanism for representing linguistic information", *Speech Communications*, v. 33, pp. 153–174, 2001.

[8] HUNT, A. J., BLACK, A. "Unit Selection in a Concatenative Speech Synthesis system using Large Speech Database". In: *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing - ICASSP*, pp. 373–376, 1996.

[9] VAN SANTEN, J., KAIN, A., KLABBERS, E., et al. "Synthesis of Prosody using Multi-level Unit Sequences", *Speech Communications*, v. 46, pp. 365–375, 2005.

[10] LATSCH, V. L. *Construção de Banco de Unidades para Síntese da Fala por Concatenação no Domínio Temporal*. Tese de Mestrado, COPPE/UFRJ, 2006.

[11] AUBERGÉ, V., CATHIARD, M. "Can we hear the prosody of smile?" *Speech Communication*, v. 40, pp. 87–97, Abril 2003. ISSN: 0167-6393.

[12] MORLEC, Y., AUBERGÉ, G. B. V. "Generating the prosody of attitudes". In: *Intonation: Theory, Models, and Applications*, pp. 251–254, Athens, Greece, 1997.

[13] BACHENKO, J., FITZPATRIC, E. "A computational grammar of discourse-neutral prosodic phrasing in English", *Computational Linguistics*, v. 16, pp. 155–170, 1990.

[14] MONNIN, P., GROSJEAN, F. "Les structures de performance en français : caractérisation et prédiction". In: *L'année psychologique*, v. 93, pp. 9–30, 1993.

[15] SEARA, I. C., PACHECO, F. S., KAFKA, S. G., et al. "Morphosyntactic Parser for Brazilian Portuguese: Methodology for Development and Assessment." In: *9th International Conference on Computational Processing of Portuguese Language (PROPOR 2010).*, pp. 1–6, 2010.

[16] AIRES, R. V. R. *Implementação, adaptação, combinação e avaliação de etiquetadores para o português do Brasil*. Tese de Mestrado, Instituto de Ciências Matemáticas de São Carlos, Universidade de São Paulo, 2000.

[17] BRILL, E. D. "Some Advances in Transformation-based Part of Speech Tagging." In: *Proceedings of the Twelfth National Conference on Artificial Intelligence (AAAI-94)*, v. 1, 1994.

[18] BARRY, W. J., FOURCIN, A. J. "Levels of labelling", *Computer Speech & Language*, v. 6, n. 1, pp. 1 – 14, 1992.

[19] WILLIAMS, B. *The Segmentation and Labelling of Speech Databases*. Eopl edinburgh occasional papers in linguistics, University of Edinburgh, 1995.

[20] TEIXEIRA, J. P. *A prosody Model to TTS Systems*. Tese de Doutorado, Faculdade de Engenharia da Universidade do Porto, may 2004.

[21] CASEIRO, D. "Grapheme-to-Phone using Finite-State Transducers". In: *IEEE Workshop on Specch Synthesis*, Santa Monica, California, USA, 2002.

[22] OLIVEIRA, C. A. M., MOUTINHO, L. C., TEIXEIRA, A. "Um novo sistema de conversão grafema-fone para o PE baseado em transdutores". In: *Actas do II Congresso Internacional de Fonética e Fonologia*, Maranhão, Brasil, 2004.

[23] BRAGA, D. *Algorítmos de processamento da Linguagem Natural para Sistemas de Conversão Texto-Fala em Português*. Tese de Doutorado, Departamento de Galego-Portugués, Francés e Lingüística, Faculdade de Filoloxía da universidade da Coruña, Coruña, 2008.

[24] SILVA, D. C., LIMA, A. A., MAIA, R., et al. "A rule-based grapheme-phone converter and stress determination for Brazilian Portuguese natural language processing". In: *VI International Telecommunications Symposium (ITS2006)*, Fortaleza, Ceará, Brasil, 2006.

[25] ALBANO, E. C., MOREIRA, A. A. "Archisegment-based Letter-to-Phone Conversion for Concatenative Speech Synthesis in Portuguese". In: *Proceedings of International Conference on Speech and Language Procesing*, v. 3, pp. 1708–1711, 1996.

[26] SILVA, T. C., ALMEIDA, S. L. "ASPA: a formulação de um banco de dados de referência da estrutura sonora do português contemporâneo." In: *XXV Congresso da Sociedade Brasileira de Computação*, v. 1, pp. 2268–2277, São Leopoldo, Rio Grande do Sul, Brasil, 2005.

[27] VASILÉVSKI, V. *Construção de um Sistema Computacional para Suporte à Pesquisa em Fonologia do Português do Brasil*. Tese de Doutorado, Universidade Federal de Santa Catarina, Florianópolis, 2008.

[28] CHBANE, D. T. *Desenvolvimento de sistema para conversão de textos em fonemas no idioma português*. Tese de Mestrado, Escola Politécnica, Universidade de São Paulo, São Paulo, 1994.

[29] JR., R. S., SEARA, I. C., KAFKA, S. G., et al. "Parâmetros Lingüísticos Utilizados para a Geração Automática de Prosódia em Sistemas de Síntese de Fala". In: *XXI Simpósio Brasileiro de Telecomunicações (SBrT 2004)*, pp. 1–6, Belém, PA, 2004.

[30] SILVA, T. C. *Fonética e Fonologia do Português - Roteiro de Estudos e Guia de Exercícios*. 9 ed. São Paulo, Editora Contexto, 2007.

[31] BISOL, L. *Introdução a estudos de fonologia do português brasileiro*. Porto Alegre - Rio Grande do Sul, EDIPUCRS, 2005.

[32] BOUMA, G. "A finite state and data-oriented method for grapheme to phoneme conversion". In: *Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference*, pp. 303–310, Seattle, USA, 2000. Morgan Kaufmann Publishers Inc.

[33] TONELI, P. M. *A palavra prosódica em Português Brasileiro: o estatuto prosódico das palavras funcionais*. Tese de Mestrado, Universidade Estadual de Campinas, Instituto de Estudos da Linguagem, 2009.

[34] CAGLIARI, L. C. *Elementos de Fonética do Português Brasileiro*. Tese de Livre-Docência, UNICAMP, 1982.

[35] TENANI, L. E. *Domínios Prosódicos no Português do Brasil: Implicações para a Prosódia e para Aplicações de Processos Fonológicos*. Tese de Doutorado, Universidade Estadual de Campinas, Instituto de Estudos da Linguagem, São Paulo, 2002.

[36] NOGUEIRA, M. V. *Aspectos segmentais dos processos de Sândi vocálico externo no Falar de São Paulo*. Tese de Mestrado, Faculdade de Filosofia, Letras e Ciências Humanas, Universidade de São Paulo, São Paulo, 2007.

[37] HAMON, C., MOULINE, E., CHARPENTIER, F. "A diphone synthesis system based on time-domain prosodic modifications of speech". In: *International Conference on Acoustics, Speech, and Signal Processing, 1989. ICASSP-89*, v. 1, pp. 238–241, May 1989.

[38] MOULINES, E., CHARPENTIER, F. "Pitch-Synchronous Waveform Processing Techniques for Text-to-Speech Synthesis Using Diphones", *Speech Communications*, v. 9, pp. 453–467, 1990.

[39] HÖHNE, H., COKER, C., LEVINSON, S. E., et al. "On temporal alignment of sentences of natural and synthetic speech". In: *IEEE Transactions on Acoustics, Speech, and Signal Processing*, v. 31, pp. 807 – 813, 1983.

[40] MALFRERE, F., DUTOIT, T. "Speech Synthesis for Text-To-Speech Alignment and Prosodic Feature Extraction". In: *Proceedings of IEEE International Symposium on Circuits and Systems ISCAS 97*, v. 4, pp. 2637–2640, Hong-Kong, 1997.

[41] MALFRERE, F., DUTOIT, T. "High-quality speech synthesis for phonetic speech segmentation". In: *Proceedings of the European Conference on Speech Communication and Technology*, pp. 2631–2634, Rhodes, Greece, 1997.

[42] LENZO, K., BLACK, A. "Diphone collection and synthesis". In: *Proceedings of the International Conference on Speech and Language Processing - ICSLP*, pp. –, 2000.

[43] SERGIO, P., OLIVEIRA, L. C. "DTW-based phonetic alignment using multiple acoustic features". In: *Proceedings of the EUROSPEECH*, pp. 309–312, 2003.

[44] YNOGUTI, C. A., VIOLARO, F. "A Brazilian Portuguese Speech Database". In: *XXVI SIMPÓSIO BRASILEIRO DE TELECOMUNICAÇÕES SBrT 08*, pp. 00–00, 2008.

[45] MALFRERE, F., DEROO, O., DUTOIT, T., et al. "Phonetic alignment: speech synthesis-based vs. viterbi-based", *Speech Communications*, v. 40, n. 4, pp. 503–515, 2003.

[46] RABINER, L., JUANG, B.-H. *Fundamentals of speech recognition*. Upper Saddle River, NJ, USA, Prentice-Hall, Inc., 1993.

[47] VERHELST, W., BROUCKXON, H. "Rejection phenomena in inter-signal voice transplantations". In: *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 165–168, 2003.

[48] BLACK, A. W., LENZO, K. A. *Building Synthetic Voices*. Festvox, jan 2003. Disponível em: <`http://festvox.org/bsv/`>. Published on the festvox website.

[49] LATSCH, V. L., NETTO, S. L. "Pitch-synchronous time alignment of speech signals for prosody transplantation." In: *ISCAS*, pp. 2405–2408, 2011.

[50] BARBOSA, P. A., BAILLY, G. "Generation of pauses within the z-score model." In: *Progress in Speech Synthesis*, pp. 365–381. Springer Verlag, 1997.

[51] ALCAIM, A., SOLEWICZ, J. A., MORAES, J. A. "Frequência de ocorrência dos fones e lista de frases foneticamente balanceadas no português falado no Rio de Janeiro", *Revista da Sociedade Brasileira de Telecomunicações*, v. 7, pp. 23–41, 1992.

[52] TURK, A., NAKAI, S., SUGAHARA, M. "Acoustic segment durations in prosodic research: a practical guide". In: *Methods in Empirical Prosody Research*, v. 3, *Language, Context, and Cognition*, Berlin, De Gruyter Language, Context, and Cognition, 2006.

[53] LANDER, T. "The CSLU labeling guide". may 1997.

[54] SEARA, I. C. *Estudo acústico-perceptual da nasalidade das vogais do português brasileiro*. Tese de Doutorado, Universidade Federal de Santa Catarina, Florianópolis, 2000.

[55] CAMPOS, H. O. V. *Duração dos segmentos vocálicos orais, nasais e nasalizados do português brasileiro*. Tese de Mestrado, Faculdade de Letras, Universidade Federal de Minas Gerais, Belo Horizonte, 2009.

[56] SOUZA, E. M. G. *Para a caracterização fonético-acústica da nasalidade no português do Brasil*. Tese de Mestrado, Instituto de Estudos da Linguagem Universidade Estadual de Campinas, São Paulo, 1994.

[57] SILVEIRA, F. *Vogal epentética no português Brasileiro: um estudo acústico em encontros consonantais*. Tese de Mestrado, Centro de comunicação e expressão, Universidade Federal de Santa Catarina, Florianópolis, 2007.

[58] CAMPBELL, W. N., ISARD, S. D. "Segment durations in a syllable frame", *Journal of phonetics*, v. 19, pp. 37–47, 1991.

[59] SANTEN, J. P. H. V. "Assignment of segmental duration in text-to-speech synthesis", *Computer Speech and Language*, v. 8, n. 2, pp. 95 – 128, 1994.

[60] BOERSMA, P., WEENINK, D. "Praat: doing phonetics by computer (Version 5.1.04)". April 2009. Disponível em: <`http://www.praat.org/`>.

[61] CAMPBELL, N. "Syllable-Based Segmental Duration". In: *Talking Machines*, Amsterdam, Elsevier Science Publishers, 1992.

[62] HIRST, D. J., ESPESSER, R. "Automatic modelling of fundamental frequency curves using a quadratic spline function". In: *Travaux de l'Institut de Phonétique d'Aix*, v. 15, 1993.

[63] HIRST, D. "A praat plugin for MOMEL and INTSINT with improved algorithms for modelling and coding intonation". In: *ICPhS XVI*, pp. 1233–1236, Saarbrücken, 2007.

[64] CELESTE, L. C. *MOMEL e INTSINT: uma Contribuição à Metodologia do Estudo Prosódico do Português Brasileiro*. Tese de Doutorado, Universidade Federal de Minas Gerais, Faculdade de Letras, Belo Horizonte, MG, 2007.

[65] TAYLOR, P. "Analysis and synthesis of intonation using the Tilt model." *The Journal of the Acoustical Society of America*, v. 107, pp. 1697–1714, 2000.

[66] FUJISAKI, H. "Information, prosody, and modeling - with emphasis on tonal features of speech -". In: *Speech Prosody 2004*, pp. 00–00, 2004.

[67] D'ALESANDRO, C., MERTENS, P. "Automatic pitch contour stylization using a model of tonal perception", *Computer Speech and Language*, v. 9, pp. 257–288, 1995.

[68] BAILLY, G., HOLM, B. "SFC: A trainable prosodic model", *Speech Communication*, v. 46, pp. 348–364, 2005.

[69] ROSENBERG, A., HIRSCHBERG, J. "On the Correlation between Energy and Pitch Accent in Read English Speech". In: *INTERSPEECH-2006*, Pittsburgh, PA, USA, 2006.

[70] SYRDAL, A. K., KIM, Y.-J. "Dialog speech acts and prosody: considerations for TTS". In: *Speech Prosody*, pp. 661–665, 2008.

[71] COUTO, I., NETO, N., TADAIESKY, V., et al. "An Open Source HMM-based Text-to-Speech System for Brazilian Portuguese". In: *7th International Telecommunications Symposium (ITS)*, Manaus, 2010.

[72] SCHRÖDER, M., BREUER, S. "XML Representation Languages as a Way of Interconnecting TTS Modules". In: *International Conference on Speech an Language Processing - ICSLP'04*, Jeju, Korea, 2004.

[73] MEDEIROS, B. R. "Vogais nasais do português Brasileiro: reflexões preliminares de uma revisita", *Revista Letras*, pp. 165–188, 2007.

[74] SILVEIRA, F., SEARA, I. C. "Vogal de apoio em grupos consonantais CCV no Português Brasileiro", *Revista da ABRALIN*, v. 7, pp. 27–47, 2008.

Apêndice A

Regras de conversão g2p.

Tabela A.1: Tabela de regras de conversão grafema-fonema.

grafema	fone	anterior	posterior	fones	inc
a	a				
		!	m#	a w	2
		!	s#	a	1
		!	n#	a	2
		!	*#	'a	1
		!	**#	'a	1
		!	%%	a	1
		!	%*	a	1
		!	u#	'a w	2
		!	us#	'a w	2
		!	i#	'a j	2
		!	is#	'a j	2
		!	iu#	a	1
á	'a				
		!	s#	'a	1
		!	%*	'a	1
		!	%#	'a	1
ã	'a				
		!	e#	'a j	2
		!	es#	'a j	2
		!	o#	'a w	2
		!	os#	'a w	2
		!	e	a j	2
		!	o	a w	2
â	'a				
		!	%*	'a	1
		!	i	'a j	2
à	'a				

grafema	fone	anterior	posterior	fones	inc
b	b				
c	k				
		!	#	k i	1
		!	h	S	2
		!	e	s	1
		!	ê	s	1
		!	é	s	1
		!	i	s	1
		!	y	s	1
		!	í	s	1
		!	$	k	1
		!	ç	k s	2
ç	s				
d	d				
		!	#	dZ i	1
		!	i	dZ	1
		!	e#	dZ	1
		!	es#	dZ	1
		!	l	d	1
		!	r	d	1
		!	*	dZ	1
e	e				
		!	%s#	e j	2
		!	%#	e j	2
		!	s#	e	1
		!	*#	’e	1
		!	**#	’e	1
		!	u#	’e w	2
		!	i#	’e j	2
		!	us#	’e w	2
		!	is#	’e j	2
		!	%*	e	1
		#a	r	E	1
		a	*	j	1

grafema	fone	anterior	posterior	fones	inc
é	'E				
		!	%s#	'e j	2
		!	%#	'e j	2
		!	s#	'E j	1
		!	i	'E j	2
		!	u	'E w	2
		!	o	'E w	2
ê	'e				
		!	%#	'e j	2
		!	%*	'e	2
		!	s#	'e j	1
		!	z#	'e j	1
f	f				
g	g				
		!	ua	g w	2
		!	uã	g w	2
		!	ue	g	2
		!	ué	g	2
		!	ui	g	2
		!	uo	g w	2
		!	uõ	g w	2
		!	e	Z	1
		!	ê	Z	1
		!	é	Z	1
		!	i	Z	1
		!	í	Z	1
		!	ü	g w	2
h	!				
		#	!	!	1
		!	#	!	1
		$	$	H	1

grafema	fone	anterior	posterior	fones	inc
i	i				
		!	%#	’i	1
		!	ns#	’i	1
		*	#	’i	1
		gu	#	’i	1
		qu	#	’i	1
		$	*#	’i	1
		*	*#	’i	1
		*	**#	’i	1
		!	u#	’i w	2
		!	on#	j o	3
		!	on$	j o	2
		!	%*	i	1
		qu	!	i	1
		qü	!	i	1
		gu	!	i	1
		$	**	i	1
		$	!	j	1
í	’i				
		!	%*	’i	1
		!	%#	’i	1
j	Z				
k	k				
l	l				
		!	$	l	1
		!	he#	l	2
		!	hi	l	2
		!	*	l	1
m	m				
		$	%	m	1
		$	*		1
		$	#		1
n	n				
		!	h	J	2
		$	*		1
		$	s#		1
		$	#		1

grafema	fone	anterior	posterior	fones	inc
o	o				
		!	sa#	’O	1
		!	sas#	’O	1
		!	sos#	’O	1
		a	s#	w	1
		a	#	w	1
		!	m#	’o	2
		!	ns#	’o	2
		!	n#	o	2
		!	s#	u	1
		!	*#	’o	1
		!	**#	’o	1
		!	u#	’o w	2
		!	i#	’o j	2
		!	us#	’o w	2
		!	is#	’o j	2
		!	%*	o	1
		é	!	w	1
ô	’o				
		!	%*	’o	1
		!	o	’o	1
õ	’o				
		!	e#	’o j	2
		!	es#	’o j	2
		!	e%#	’o	1
		!	e	o j	2
ó	’O				
		!	%	’o	1
		!	i	’O j	2
p	p				
		!	h	f	2
q	k				
		!	#	k i	1
		!	ua	k w	2
		!	uá	k w	2
		!	uã	k w	2
		!	ue	k	2
		!	ui	k	2
		!	ué	k	2
		!	uê	k	2
		!	uí	k	2
		!	u$	k	2
		!	ü	k w	2

grafema	fone	anterior	posterior	fones	inc
r	r				
		#	!	H	1
		!	r	!	1
		r	!	H	1
		!	%	H	1
		!	b	H	1
		!	d	H	1
		!	g	H	1
		!	j	H	1
		!	l	H	1
		!	v	H	1
		!	z	H	1
		!	*	h	1
		!	#	h	1
		l	$	H	1
s	s				
		!	s	s	2
		!	ce	s	2
		!	ci	s	2
		!	ç	s	2
		p	#	s	1
		!	#	S	1
		$	$	z	1
		tran	$	z	1
		trân	$	z	1
		#ex	$	z	1
		!	%	Z	1
		!	d	Z	1
		!	g	Z	1
		!	j	Z	1
		!	l	Z	1
		!	v	Z	1
		!	z	Z	1
		!	*	S	1
		*	$	s	1
t	t				
		!	#	tS i	1
		!	h#	tS i	2
		!	h	t	2
		!	i	tS	1
		!	í	tS	1
		!	r	t	1
		!	l	t	1
		!	*	tS	1

grafema	fone	anterior	posterior	fones	inc
u	u				
		!	m#	’u	1
		!	ns#	’u	1
		!	n#	u	2
		*	#	’u	1
		*	*#	’u	1
		!	%*	u	1
		*	**#	’u	1
		*	is#	’u j	2
		!	i#	’u j	2
		m	it	u j	2
		$i	*	u	1
		$	!	w	1
ú	’u				
		!	%*	’u	1
		!	%#	’u	1
v	v				
w	v				
x	S				
		#	!	S	1
		!	#	k s	1
		!	ce	s	2
		!	cê	s	2
		!	ci	s	2
		!	cé	s	2
		!	cí	s	2
		#e	$	z	1
		e	$	k s	1
y	i				
		*	#	’i	1
z	z				
		#	!	z	1
		$	#	S	1
		!	m	Z	1

Apêndice B

Critério de etiquetagem

A seguir, são apresentados exemplos do critério de posicionamento das etiquetas correspondentes ao grupo das Oclusivas e Africadas [p b t d k g tS dZ], das Fricativas [f v s z S Z h H], das Consoantes Nasais [m n J], das Líquidas [r l L], das Vogais Orais [a e i o u O E] e das Vogais Nasais [ã ẽ ĩ õ ũ]. Os símbolos usados para identificar as etiquetas são apresentados no Capítulo 2.

Oclusivas e Africadas [p b t d k g tS dZ]

Foi adotaddo etiquetar as oclusivas por duas fases: uma fase de oclusão e uma fase seguinte que termina com o início da vogal posterior, tal como [44]. Apesar de tais fases não possuírem *status* fonológico, estes são casos em que símbolos foram acrescidos à cadeia de fones da transcrição ampla.

As etiquetas [+] e [-] foram atribuídas à fase oclusiva das oclusivas sonoras e surdas, respectivamente. Às etiquetas da região que correspondem à fase que vai da soltura da oclusão ao inicio do vozeamento da vogal contígua à oclusiva, foram associadas aos próprios símbolos [p], [t], [k], [b], [d] e [g].

A fase de oclusão é identificada no sinal pela intensidade reduzida, seguida de um aumento brusco na amplitude que indica o inicio da explosão, como pode ser observado na figura B.1. O final da fase de desbloqueio das oclusivas surdas é marcado antes do primeiro pulso glotal da vogal seguinte, enquanto o fim da fase oclusiva das sonoras é marcado quando a intensidade atinge o nível da vogal, podendo vir a incluir 1 ou 2 ciclos glotais da vogal seguinte, tal como mostrado no exemplo da figura B.1.

As africadas [dZ] e [tS] também apresentam a divisão em uma fase de oclusão e uma fase de fricção, por essa razão adotou-se o mesmo critério de divisão em duas fases e os mesmos símbolos [+] e [-], respectivamente.

O posicionamento das fronteiras da fase de oclusão, segue o mesmo procedimento das oclusivas. No entanto a fase de fricção é marcada pela presença acentuada

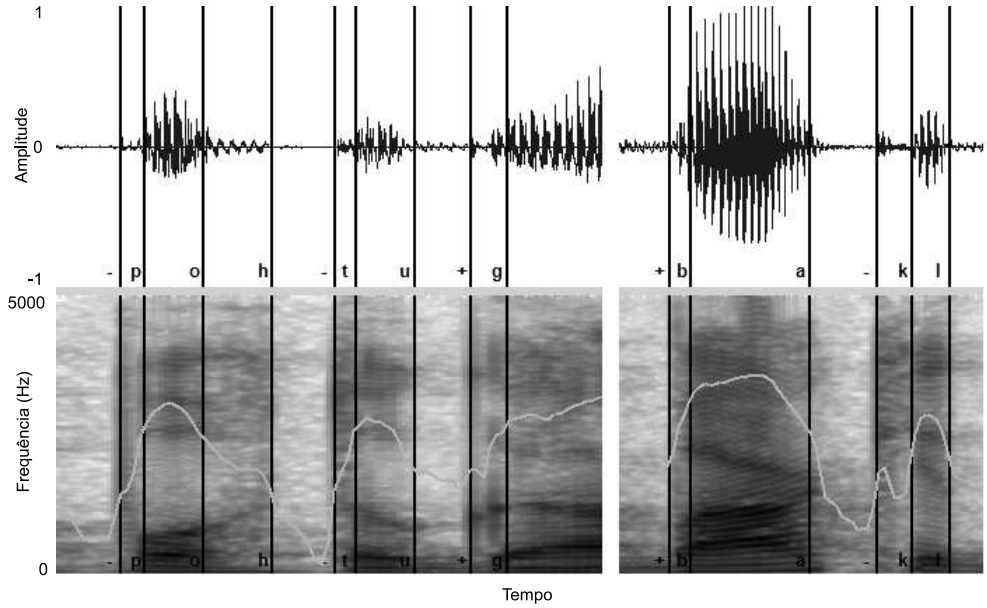


Figura B.1: Exemplo de etiquetagem em duas fases das oclusivas [p], [t], [g], [b] e [k]

de ruído aleatório com a concentração da energia em alta frequência. O início da primeira formante da vogal seguinte é ainda um outro indicativo da fronteira posterior, tal como é mostrado no exemplo da figura B.2.

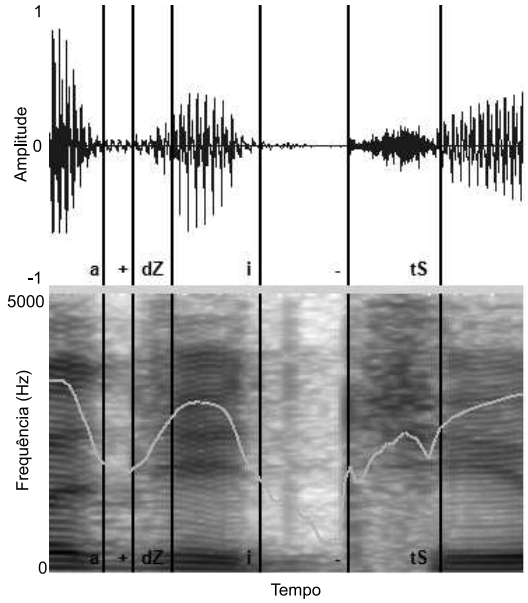


Figura B.2: Exemplo de etiquetagem em duas fases das fricadas [dZ] e [tS].

Fricativas [f v s z S Z h H]

As fricativas são caracterizadas primeiramente pela concentração da energia em altas frequências e pela baixa amplitude no sinal. As fricativas surdas apresentam menor

amplitude do que as sonoras, sendo caracterizadas basicamente por ruído aleatório.

A energia em alta frequência mostrada no espectrograma, assim como o ponto médio do decaimento e subida da curva de intensidade, foram os melhores indícios de fronteira das fricativas. Na figura B.3 são mostrados exemplos de etiquetagem das fricativas [Z], [s], [z] e [S], e na figura B.4 exemplos das fricativas [v], [f], [H] e [h], todas entre vogais.

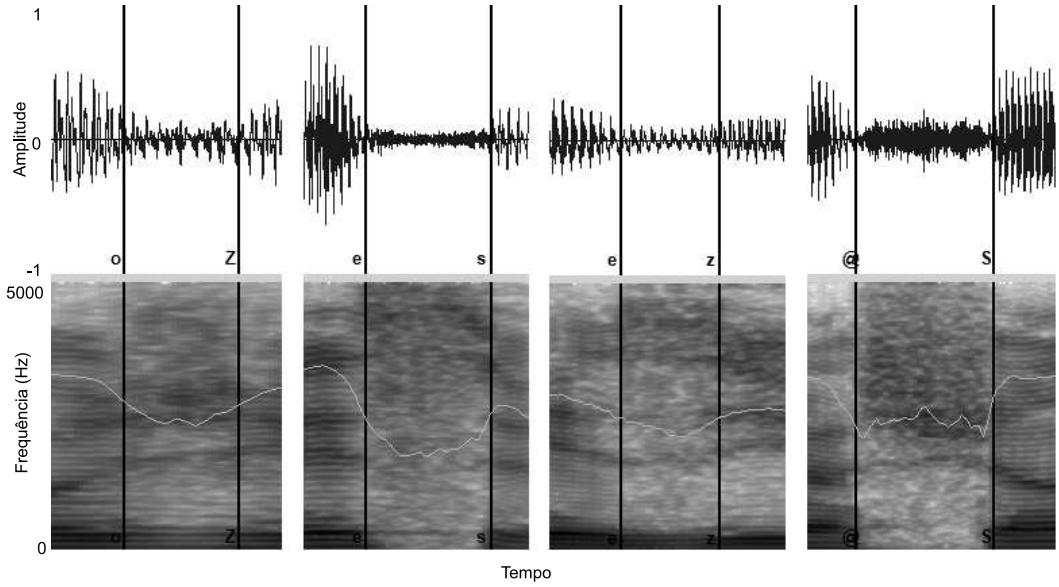


Figura B.3: Exemplo de etiquetagem das fricativas [Z],[s],[z] e [S].

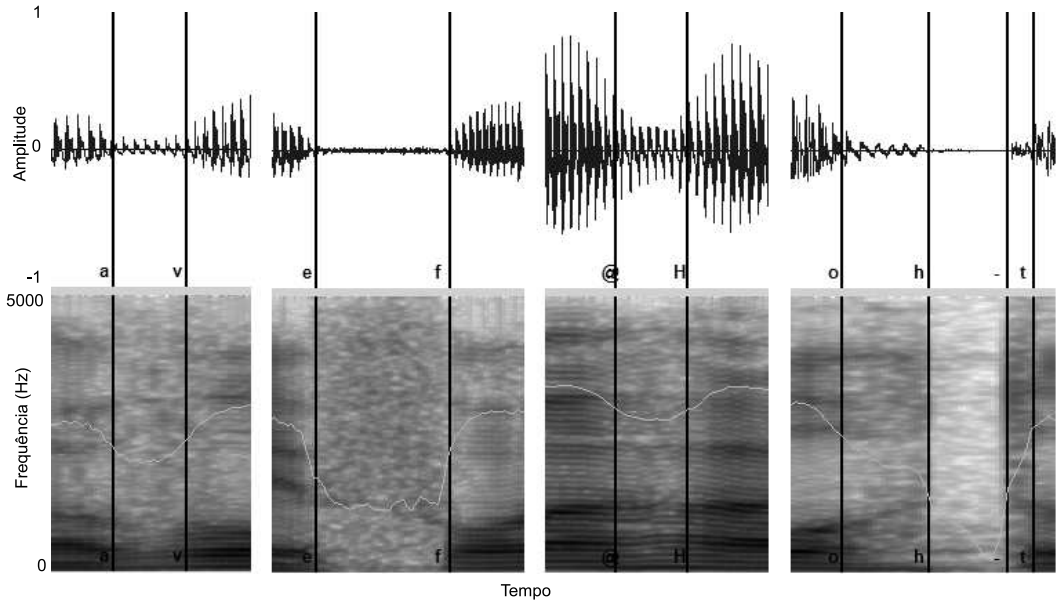


Figura B.4: Exemplo de etiquetagem das fricativas [v], [f], [H] e [h].

Consoantes Nasais [m n J]

As consoantes nasais [m] [n] são facilmente identificáveis, com uma interrupção nas formantes superiores e um nítido decaimento na amplitude e intensidade do sinal. Na figura B.5 são dados exemplos da etiquetagem destas consoantes.

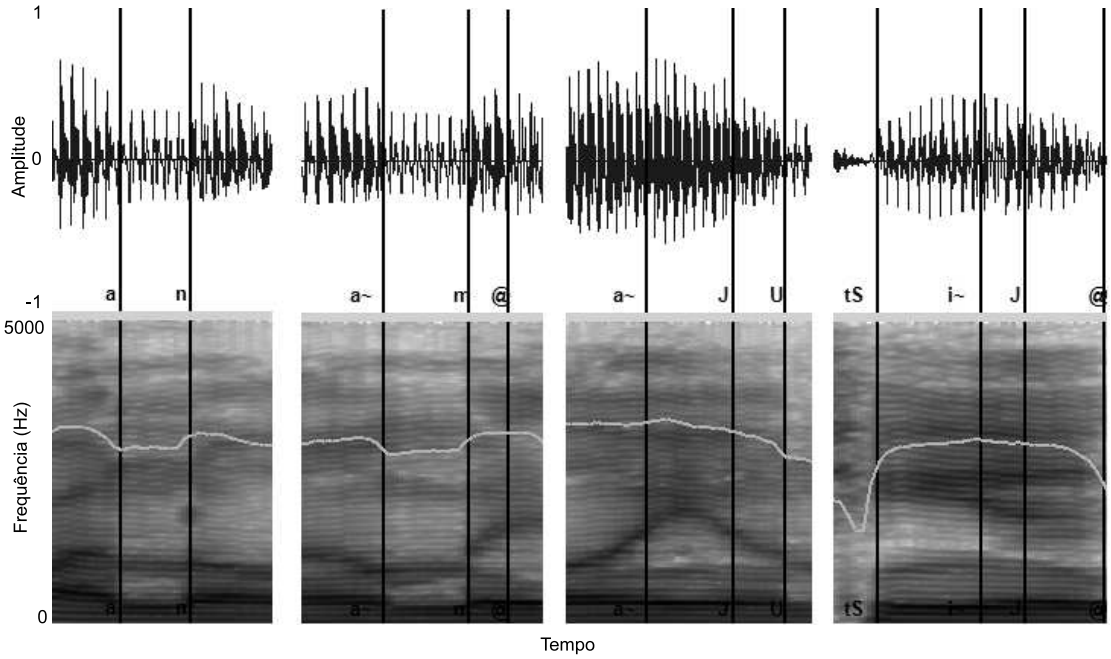


Figura B.5: Exemplo de etiquetagem das consoantes nasais [n], [m], [ãJ] e [ĩJ].

O fone [J] entretanto não é tão facilmente identificável, pois na maioria dos dialetos do português brasileiro é pronunciado como [j̃], por exemplo, na palavra "banha" [bãJa] que pode ser pronunciada como [bãj̃a] [30]. Apesar deste fato, foi mantida a notação [J] e o recorte é feito conforme a transição da segunda formante. No caso da vogal precedente ser um [ĩ], a transição se torna ainda menos definida (se é que existe) e neste caso a marcação é feita como indicada na figura B.5, isolando a transição do [ĩ] para a vogal seguinte. Esta ocorrência surge também na junção de palavras terminadas em [ĩ] com palavras iniciadas por vogal, como em "capim amarelo", que pode ser transcrito por [kapĩamarElu] ou [kapĩJamarElu], onde o primeiro caso foi preferido.

Líquidas [r l L]

As consoantes líquidas [r l L] são identificáveis principalmente pela amplitude do sinal e pela curva de intensidade. O [r] é caracterizado por um curta duração e é encontrado entre vogais, ou precedido por uma oclusiva ou por uma vogal de apoio. Na figura B.6 nota-se nitidamente o declínio abrupto na amplitude e a descontinuidade no padrão das formantes. As consoantes [l] e [L] demonstram um comportamento na intensidade similar às nasais, porém com um rompimento mais nítido nas formantes

mais baixas. Nota-se que a consoante [L], presente na palavra “julho” [ZuLU], pode ter a mesma emissão que na palavra “Júlio” [ZuljU], como observado em [30], porém foi mantida a representação de [L].

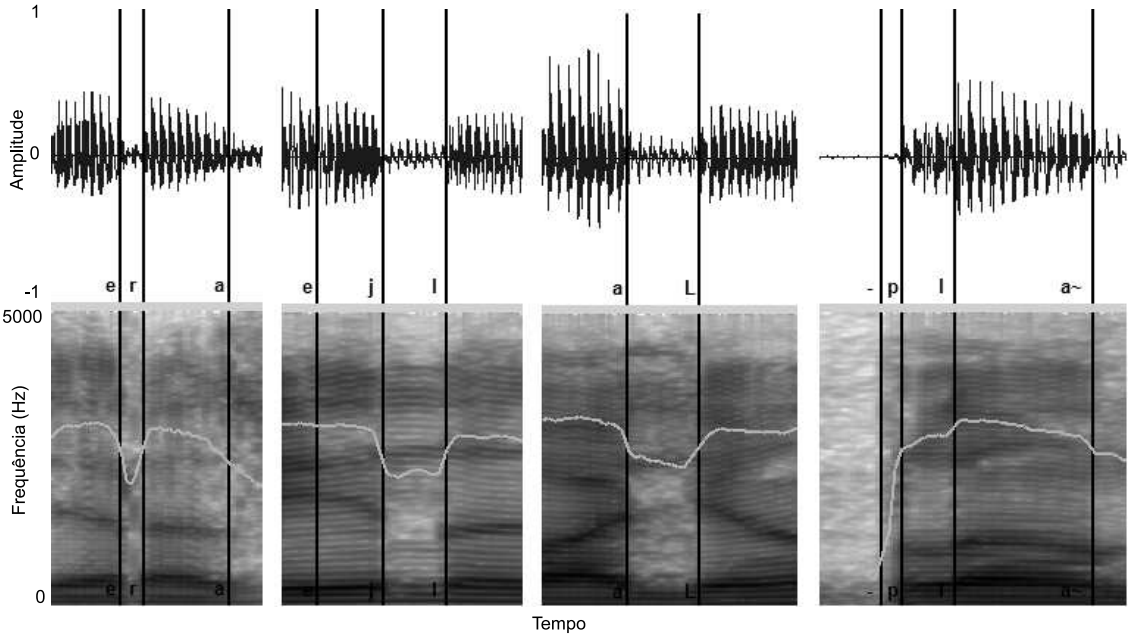


Figura B.6: Exemplo de etiquetagem das consoantes [r], [l], [L] e [pl]

Vogais orais [a e i o u O E]

As vogais são identificadas principalmente pela frequência das primeiras formantes (F1, F2 e F3). Na figura B.7 é mostrado o triângulo das vogais, construído com base nas duas primeiras formantes F1 e F2.

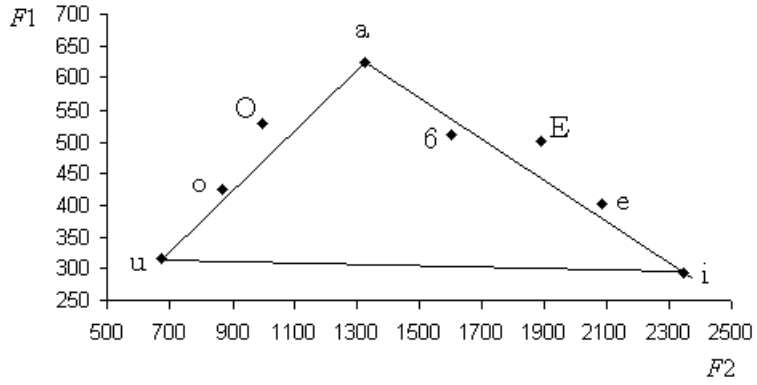


Figura B.7: Caracterização das vogais a partir das formantes F1 e F2 [20].

No caso do encontro entre consoantes e vogais foi mais conveniente observar a fronteira a partir das características da consoante.

Na separação dos encontros vocálicos, existe uma grande parcela de arbitrariedade, como nota [19]. Porém alguns fatores podem ser observados como indicativos

para a etiquetagem. A transição das formantes é um primeiro indicativo, tal como em [6], onde o ponto médio da transição foi tomado como critério.

No caso em que as formantes F1 ou F2 das vogais a serem separadas estão distantes, por exemplo [au] ou [ui], a transição nas formantes é clara. No entanto, quando a vogais possuem formantes muito próximas, esta transição não é tão evidente, e outros eventos podem ser indicativos de recorte, tais como: o surgimento de formantes superiores, instantes de mínimos e máximos na transição das formantes; e ainda mudanças de amplitude.

Na figura B.8 são mostrados exemplos de casos críticos de etiquetagem em encontros vocálicos.

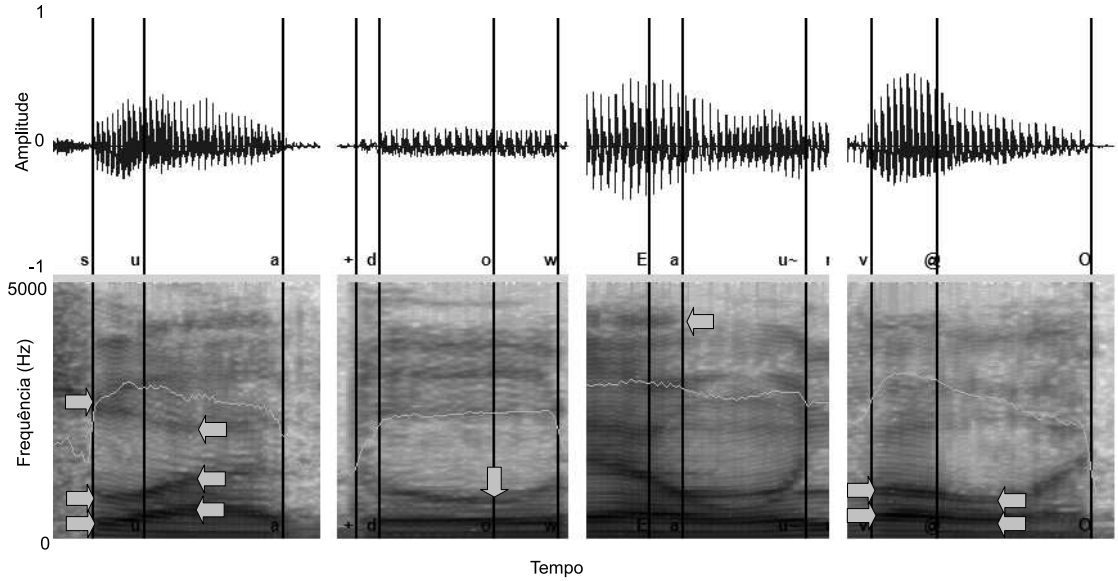


Figura B.8: Exemplo de etiquetagem dos encontros vocálicos [ua], [ow], [Eaũ] e [aO].

No caso das vogais finais, CAMPBELL [6] cita que esta é a maior fonte de divergência na etiquetagem. O fim da atividade glotal, indicado pelo eletrolotógrafo, não é um bom indicativo para o recorte final, pois a ressonância do trato vocal permanece e é percebida após o fim do vozeamento. No espectograma se observa esta energia remanescente e a continuação das formantes. Assim, foi tomado como critério observar a intensidade, utilizando como corte um ponto onde a energia caia razoavelmente e o espectrograma demonstre o desaparecimento do segmento. Este critério se aplica também às consoantes finais [h] e [S] principalmente.

vogais Nasais [ã ẽ ĩ õ ũ]

Diferentes autores diferenciam as vogais nasalizadas das vogais nasais. Em [56] a autora define que as vogais nasalizadas têm origem na coarticulação com uma consoante nasal [m n J], e as vogais nasais são independentes deste contexto. Em [30] a autora cita que para a maioria dos dialetos do Português Brasileiro (PB), as

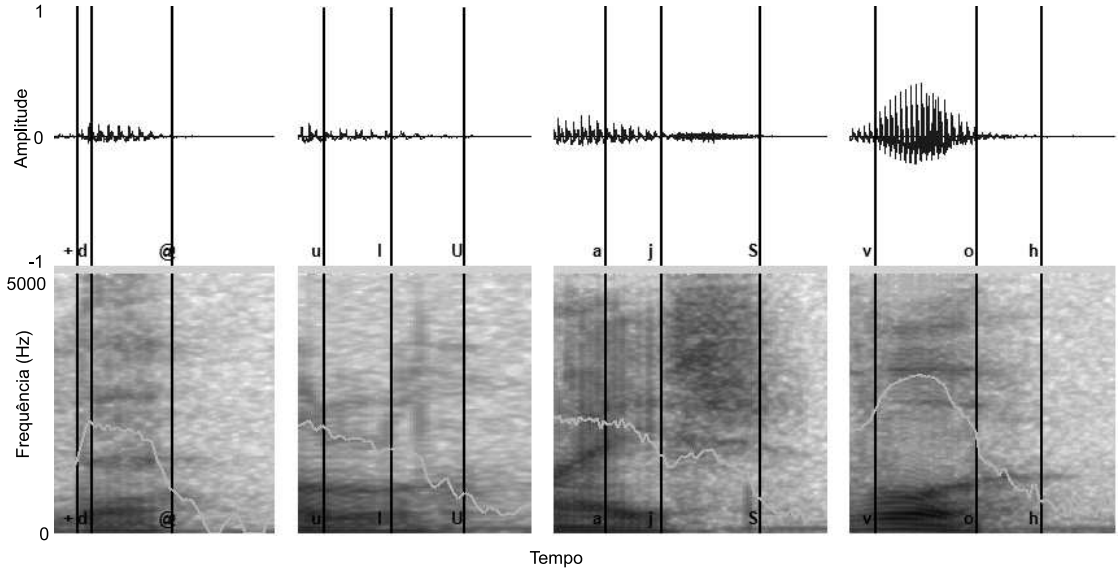


Figura B.9: Exemplo de etiquetagem dos segmentos finais [@], [U], [S] e [h].

vogais tônicas quando seguidas de [m n J] são nasalizadas, como em "fome" [fõmI]. Porém, as vogais nasais, como em "santo" [sãtU], são obrigatórias em qualquer dialeto do PB.

SEARA [54] (apud [55]) e SOUZA [56], concordam que foneticamente as vogais nasais configuram pelo menos duas fases: uma porção da vogal e o término em um "murmúrio nasal". SOUZA [56] descreve que este murmúrio nasal é passível de ser isolado e parece ocupar grande parte da duração da vogal nasal, podendo ser o responsável pela maior duração das vogais nasais em comparação com as orais.

MEDEIROS [73] observa que o mumúrio nasal das vogais nasais está mais presente antes das oclusivas. Neste caso, SOUZA [56] observa que o tempo da oclusão é inegavelmente menor quando precedido por vogais nasais, o que parece que a vogal está "roubando" o tempo da oclusiva. Pela observação desta influência do murmúrio nasal na duração das oclusivas, foi então decidido dividir a vogal nasal em duas fases, isolando o murmúrio nasal, quando houver uma consoante oclusiva posterior.

O murmúrio nasal é sonoro, apresentando um declínio na intensidade da vogal nasal. No espectrograma pode ser identificado por uma mudança no padrão das formantes da vogal. A duração do murmúrio nasal, quando seguido por uma oclusiva sonora, pode ocupar quase completamente o período de oclusão, tal como pode ser observado na figura B.10, onde foi utilizado o símbolo [N] para a etiquetagem.

Nos ditongos nasais [ẽj̃ ãj̃ õj̃ ũj̃ ãw̃ õw̃], a fronteira entre a vogal e a semivogal (glide), ambos nasais, foi etiquetada. O critério usado segue o mesmo utilizado nas vogais orais, de observar a transição das formantes, embora o padrão das formantes nasais seja um pouco diferentes das orais.

Quanto sucedidos por uma oclusiva, o murmúrio nasal também foi etiquetado ao

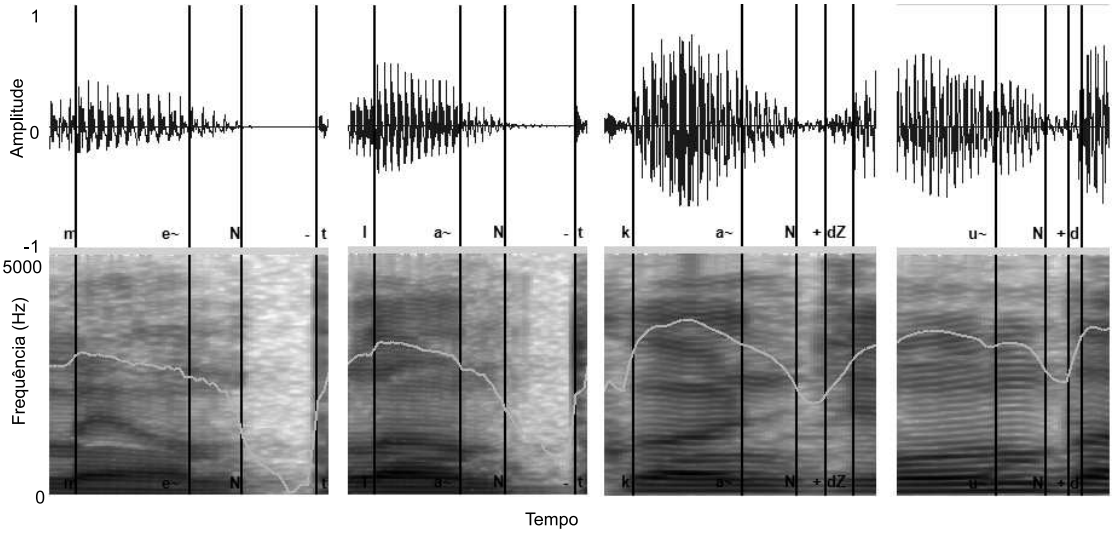


Figura B.10: Exemplo de etiquetagem das vogais nasais seguidas do murmúrio nasal [N].

fim dos ditongos nasais, como pode ser observado na figura B.11.

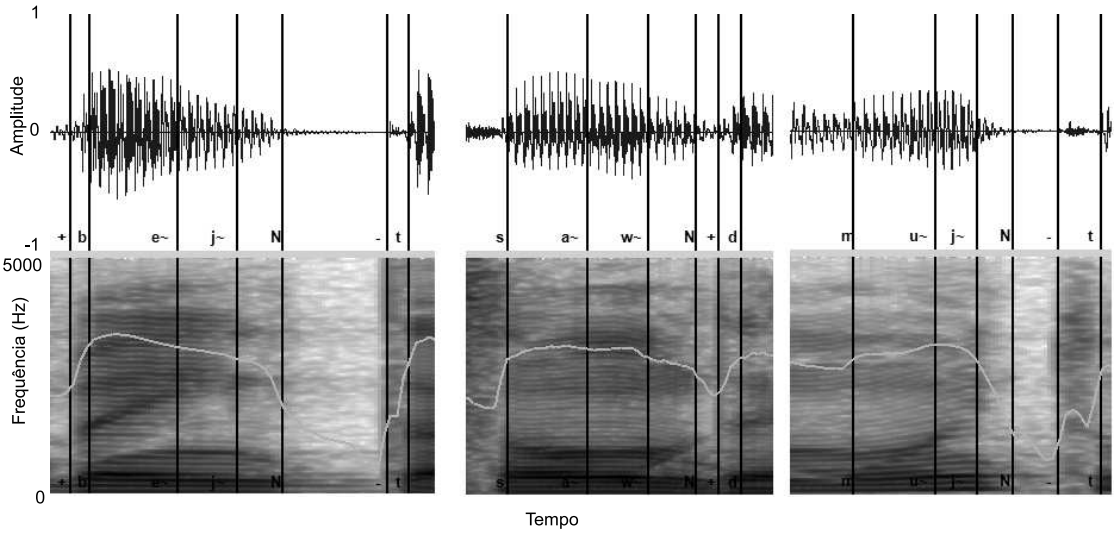


Figura B.11: Exemplo de etiquetagem de ditongos nasais seguidas do murmúrio nasal [N].

As vogais de apoio e epentética

Em [57] a autora estuda o segmento vocálico que surge entre encontros consonantais, usando o termo "vogal epentética" nas situações onde há a inserção de uma vogal nos encontros consonantais heterossilábicos (consoantes em sílabas diferentes), como nas palavras "advogado" e "amnésia". Nas situações em que a inserção da vogal é feita nos encontros consonantais tautossilábicos (consoantes na mesma sílaba), tal como "prato" e "brisa", é usado o termo "vogal de apoio". Ao longo deste trabalho

adotou-se a mesma nomenclatura.

No caso da vogal de apoio, em [74] a autora conclui que esta vogal aparece diante de [r] em 96% dos casos observados. Ao longo da etiquetagem do corpus deste trabalho, foi observada a presença da vogal de apoio diante de [r] em todos os casos, enquanto que diante de [l] a vogal de apoio foi raramente observada.

A vogal epentética, segundo alguns autores, pode dar origem a uma nova sílaba. Por exemplo, se na transcrição da palavra "objeto" [ob.ZE.tU], fosse considerada a existência da vogal epentética, a palavra seria transcrita como [o.bi.'ZE.tU]. Porém este remanejamento da estrutura silábica não é consensual entre diferentes autores. Neste contexto, ao longo da etiquetagem do corpus observou-se que a não etiquetagem da vogal epentética em alguns casos, implicaria em uma duração errônea do segmento adjacente, principalmente no caso das oclusivas, cuja duração é pequena.

Portanto, foi decidido etiquetar a vogal de apoio quando diante de [r] e a vogal epentética sempre que a consoante posterior for uma nasal, ou estiver entre oclusivas, ou entre uma oclusiva e fricativa, tal como as observações de SILVEIRA [57]. Na figura B.12 são dados exemplos de etiquetagem das vogais de apoio e epentética, com o símbolo [%], quando consideradas conforme o critério acima. Os exemplos foram obtidos das palavras "adstringente" [adZS.t%rĩ.gẽ.tSi], "psicólogo" [psi.kO.lo.gu] e "obter" [ob%.teh].

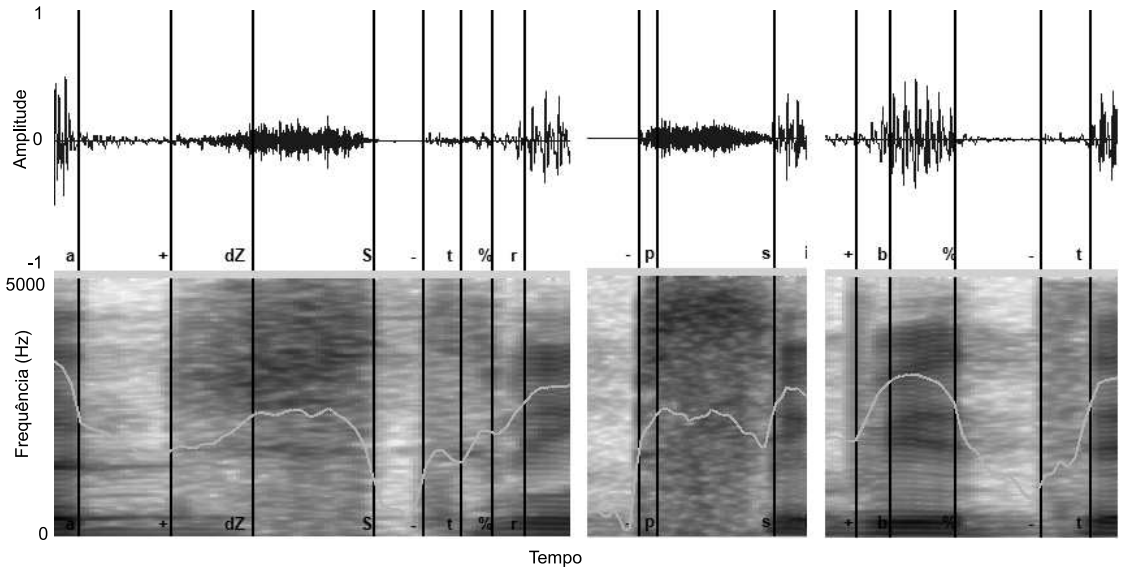


Figura B.12: Exemplo de etiquetagem da vogal epentética [%] quando considerada.

Casos particulares

As regras aplicadas ao longo da obtenção da transcrição fonética automática foram aperfeiçoadas no intuito de maximizar a associação direta entre a transcrição automática e as ocorrências fonéticas no corpus. Porém, nem todas as ocorrências no

corpus foram passíveis de serem generalizadas por regras ou viáveis de serem tratadas como exceção. Por isso, foram criados mecanismos de escape para estes casos particulares.

Um primeiro caso particular refere-se ao encontro de vogais na junção de palavras. Na fala contínua, este encontro vocálico resulta em fenômenos onde certas vogais são suprimidas, modificadas ou assimiladas por outra. Na obtenção da transcrição fonética automática, foram formuladas regras para a ocorrência destes fenômenos em contextos mais favoráveis, descritas no Capítulo 2. Porém, devido a dificuldade de generalização destes fenômenos, muitas vezes ocorre que o sinal de fala indica a presença ou ausência de vogais que a transcrição automática não considera. Um destes casos é a junção de vogais de mesma natureza, em que a observação do espectograma indica uma única vogal e a transcrição automática não considera a junção. Neste caso, é etiquetada a separação entre as vogais, mesmo que não esteja claro que tal separação exista. Normalmente, a duração excessiva da vogal justifica a separação [33].

Outro caso particular é a assimilação de uma vogal pela consoante. Por exemplo, quando o [i] e [u] aparece em posição postônica final, precedido por [S] ou [s]. Neste caso, ocorre geralmente que não há indicativos da presença da vogal, nem no espectrograma, nem no sinal no domínio do tempo. A vogal é perceptível pelo som, porém misturada à consoante, não sendo possível definir as fronteiras ou duração a da vogal. No entanto, a presença da vogal não pode ser ignorada, pois resultaria em uma sílaba sem vogal e portanto optou-se por etiquetar a vogal com uma duração desprezível ($< 10ms$), tal como é mostrado na figura B.13, de maneira que este caso possa ser tratado posteriormente.

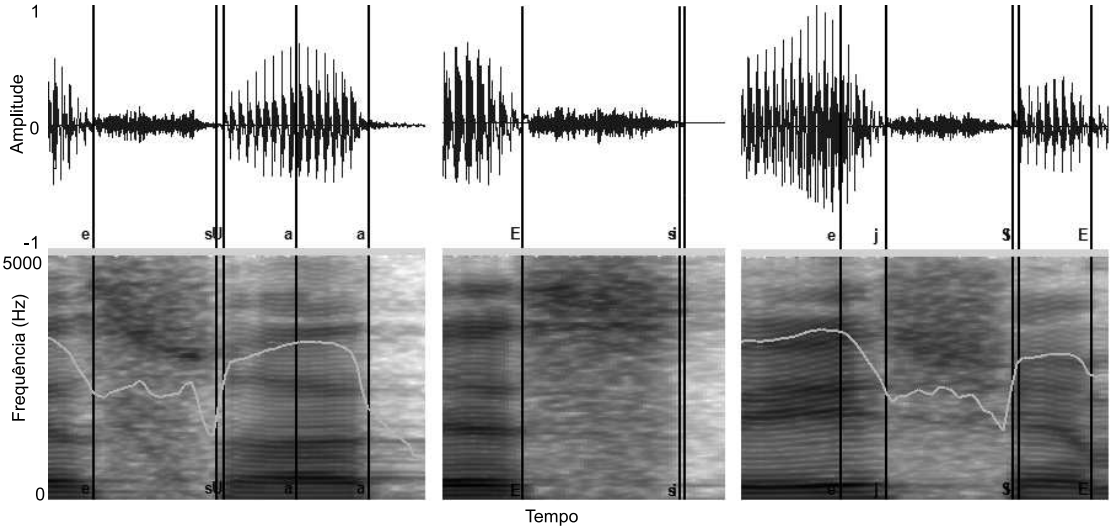


Figura B.13: Exemplo de etiquetagem nos casos de degeminação e assimilação da vogal pela fricativa.

Nos casos descritos acima, são considerados os segmentos que são indicados na transcrição mas a existência no sinal não é evidente, e vice versa, segmentos evidentes no sinal mas não indicados na transcrição. Há ainda os casos em que a existência do segmento não é questionável, mas a etiqueta atribuída ao segmento não corresponde exatamente ao segmento presente. Estes casos são: a redução vocálica do [e] e [o] para [i] e [u], em posição pretônica ou postônica medial [30], que não foram implementadas em regras. Assim, nas palavras "menino" [mininu], "pérola" [pErula] e "bonito" [bunitu], por exemplo, mantêve-se o [e] e [o], sendo portanto transcritas por [meninu], [pErola] e [bonitu]; e também os casos de formação de ditongos, onde as vogais [u] e [i] são visivelmente transformadas em glides, apresentando no espectograma uma duração reduzida. Em ambos os casos, a etiqueta atribuída será aquela indicada pela transcrição fonética automática.

Apêndice C

Tabelas de distribuição das durações e intensidades

Tabela C.1: Média e desvio padrão da intesidade dos segmentos e correlação com a intensidade da vogal.

segmentos	média (dB SPL)	desvio (dB SPL)	corr
p t k	47,93	4,71	0,25
b d g	55,69	3,87	0,63
dZ tS	54,94	2,97	0,65
f	42,43	2,76	0,00
v	51,66	3,39	0,37
s z S	54,53	3,49	0,00
Z	56,84	2,69	0,34
h	51,66	6,30	0,66
H	61,58	3,68	0,58
m n	61,12	3,13	0,72
J	64,97	4,51	0,78
r	54,64	4,28	0,72
l	61,37	3,45	0,71
L	57,20	4,20	0,72
i u ĩ ũ	64,77	3,64	1,00
e o ẽ õ ã	67,16	3,45	1,00
a E O	68,55	3,53	1,00
I U	60,04	7,54	1,00
@	64,21	7,19	1,00
w j w̃ j̃	65,34	4,73	0,82
-	37,98	4,42	0,16
+	49,98	3,40	0,40
%	65,31	6,28	0,72
N	58,80	4,17	0,63

Tabela C.2: Média e desvio padrão da duração dos segmentos e correlação com a duração da vogal.

segmentos	posição	anterior	posterior	média	desvio
[e E a O o]		glide oral		4,39	0,38
[e E a O o]			glide oral	4,39	0,38
[a e o]			glide nasal	4,39	0,38
[e E a O o a e o]				4,60	0,40
[i u i u]				4,19	0,52
@				4,64	0,21
I U				4,26	0,40
j w			vogal	3,83	0,51
j w j w				3,94	0,38
-		N		3,93	0,41
-				4,30	0,34
+		N		3,15	0,29
+				3,88	0,36
N			+	3,72	0,36
N				3,62	0,40
%				3,15	0,34
S	coda		_	5,15	0,17
S	coda			4,35	0,20
Z	coda			4,11	0,29
h	coda		_	4,79	0,23
h	coda			4,24	0,29
H	coda			4,23	0,36
f	onset			4,86	0,26
v	onset			4,33	0,26
s	onset			4,89	0,23
z	onset			4,42	0,27
H h				4,17	0,37
f s S				5,00	0,18
v z Z				4,43	0,29
p				2,99	0,31
b				2,97	0,32
t				3,15	0,28
d				2,88	0,27
k				3,44	0,31
g				3,21	0,36
tS				3,83	0,50
dZ				3,55	0,36
L l				4,08	0,33
r				3,16	0,21
n				4,10	0,32
m J				4,28	0,27

Apêndice D

Diagramas de Classes

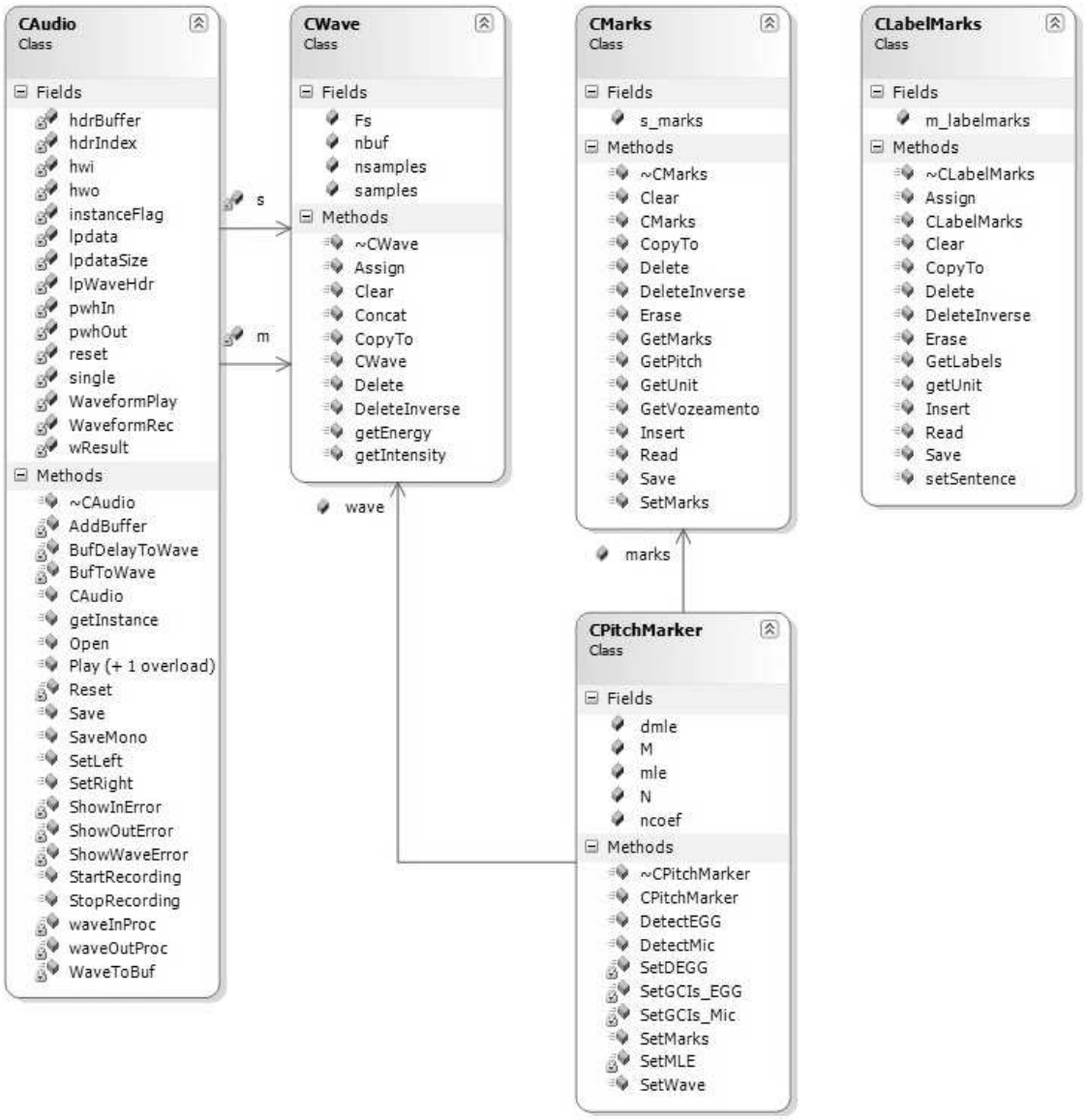


Figura D.1: Classes que implementam as ferramentas de edição.

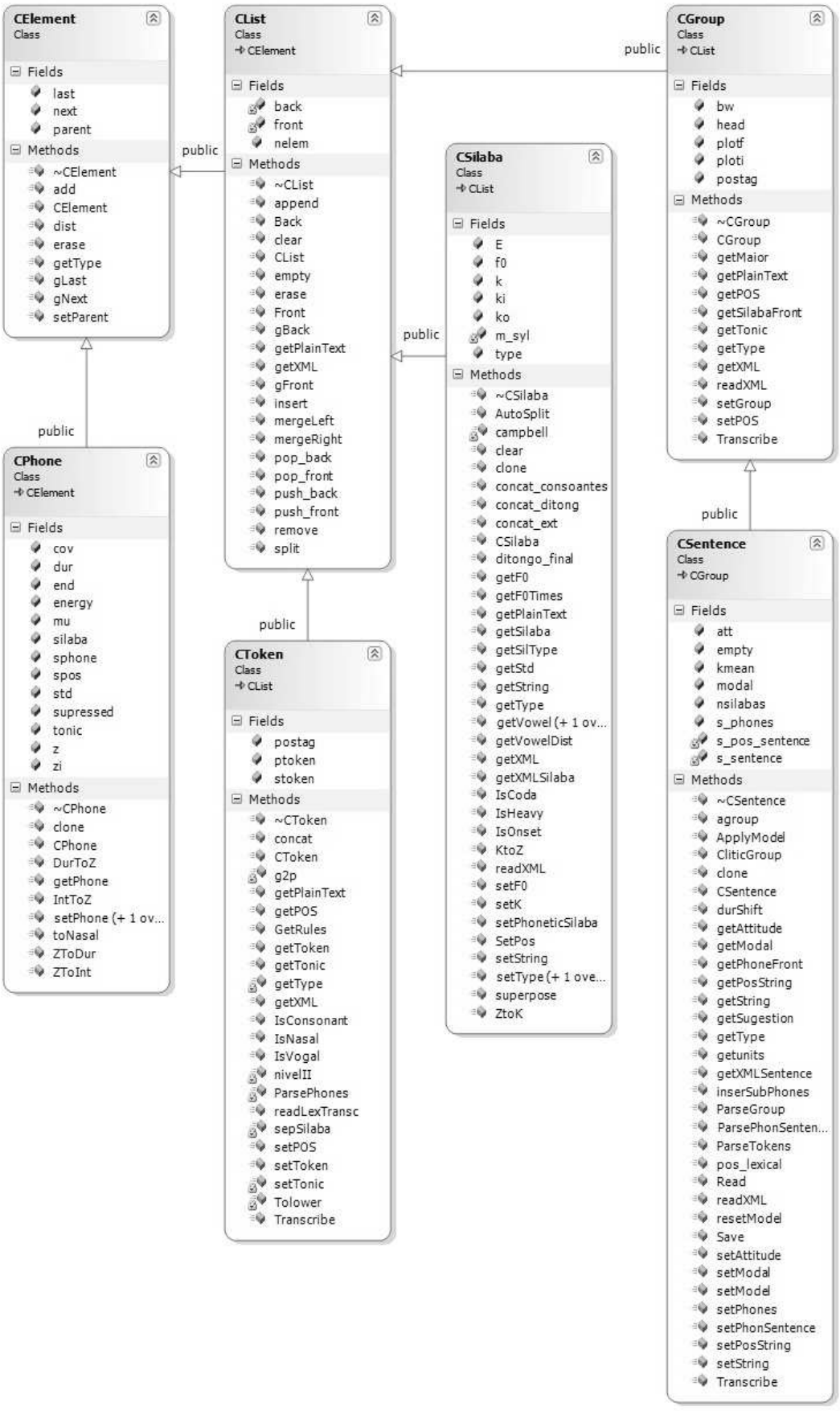


Figura D.2: Classes que implementam a estrutura de dados.

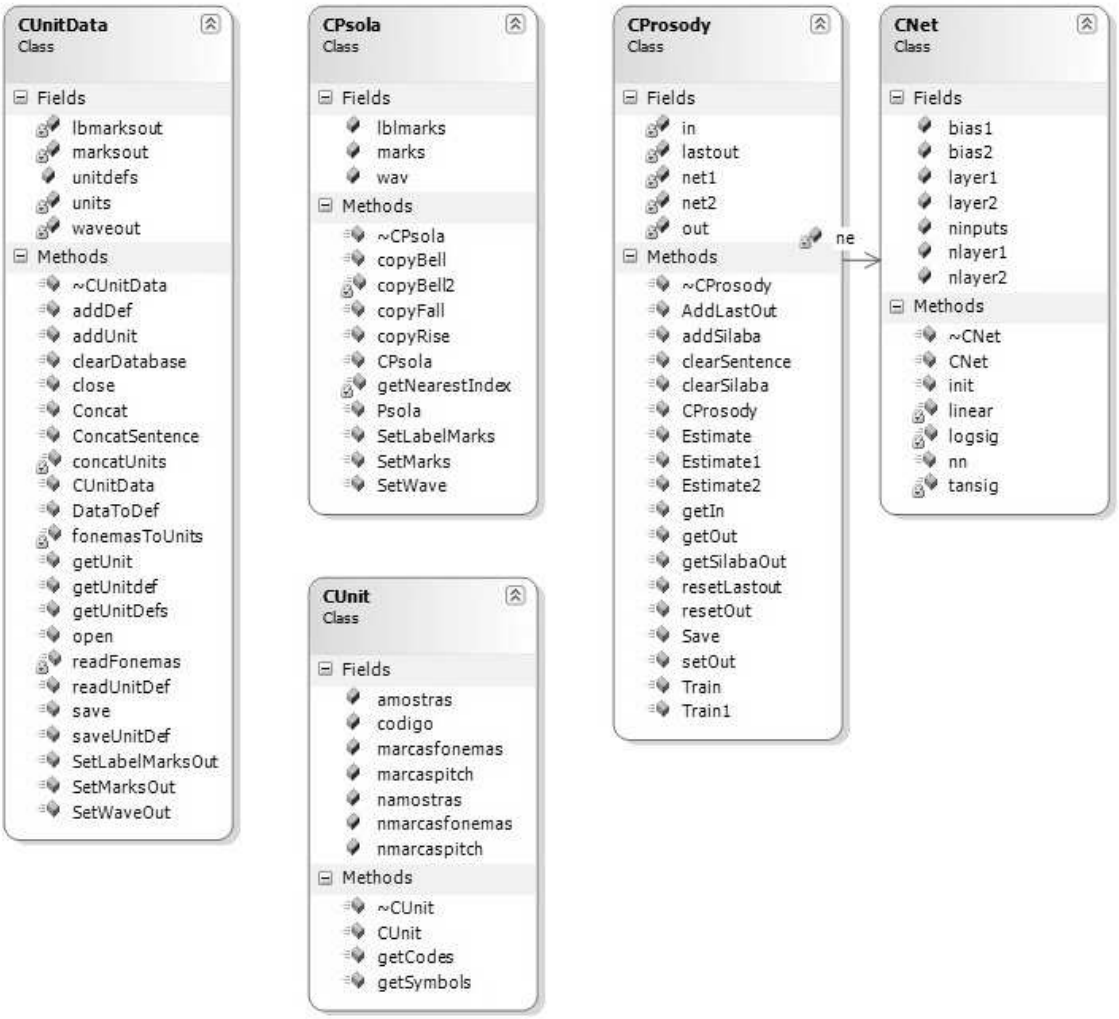


Figura D.3: Classes que implementam a etapa de síntese da fala do prótipo TTS.

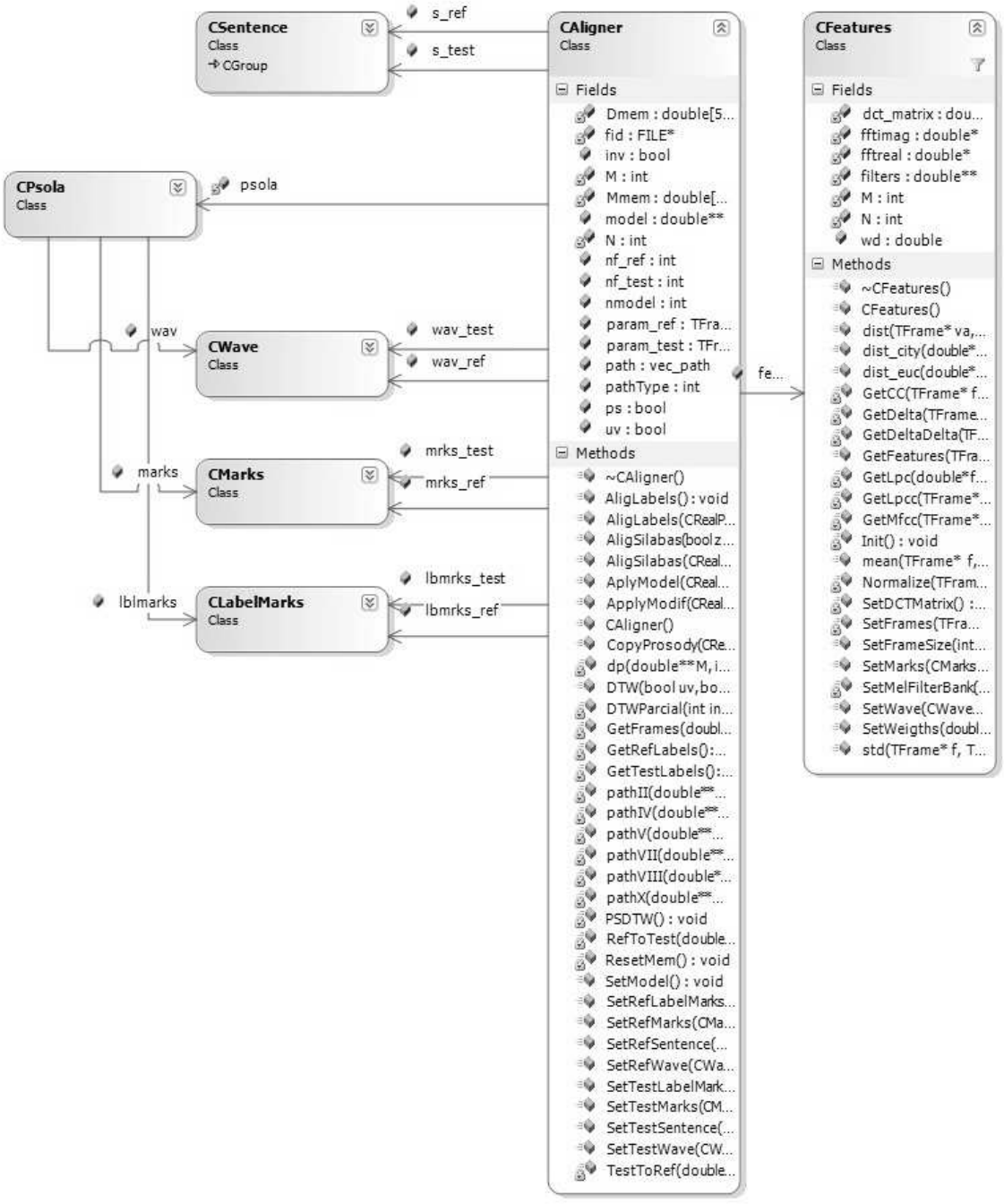


Figura D.4: Classes que implementam os métodos de transplante.