Um Modelo para Geração de Prosódia de Palavras em Conversores Texto-Fala para a Língua Portuguesa Falada no Brasil.

Manuel Leonel da Costa Neto

Tese de Doutorado submetida à Coordenação dos Cursos de Pós-Gradução em Engenharia Elétrica do Centro de Ciências e Tecnologia da Universidade Federal de Campina Grande - Campus I como parte dos requisitos necessários para obtenção do grau de Doutor em Ciências no domínio da Engenharia Elétrica.

Área de Concentração: Processamento da Informação

Benedito Guimarães Aguiar Neto, Dr. Ing.
Orientador
Maria Auxiliadora Bezerra, Dra.
Orientadora

Campina Grande, Paraíba, Brasil

C837m Costa Neto, Manuel Leonel da
Um modelo para geracao de prosodia de palavras em conversores texto-fala para a lingua portuguesa falada no Brasil / Manuel Leonel da Costa Neto. - Campina Grande, 2004.
170 f.

Tese (Doutorado em Engenharia Eletrica) - Universidade Federal de Campina Grande, Centro de Ciencias e Tecnologia.

1. Conversores Texto-Fala 2. Sistema da Fala 3. Prosodia 4. Tese I. Aguiar Neto, Benedito Guimaraes II. Bezerra, Maria Auxiliadora III. Universidade Federal de Campina Grande - Campina Grande (PB) IV. Título

CDU 621.391:801.6(043)

UM MODELO PARA A GERAÇÃO DE PROSÓDIA DE PALAVRAS EM CONVERSORES TEXTO-FALA PARA A LÍNGUA PORTUGUESA FALADA NO BRASIL

MANUEL LEONEL DA COSTA NETO

Tese Aprovada em 30.04.2004

PROF. BENEDITO GUIMARÃES AGUIAR NETO,Dr.-Ing., UFCG
Orientador

PROFa. MARIA AUXILIADORA BEZERRA, Dr., UFCG
Orientadora

PROF. SIDNEY CERQUEIRA BISPO DOS SANTOS, Dr., FAC. MICHELANGELO -DF
Componente da Banca

PROF. RUI SEARA, Dr., UFSC
Componente da Banca

PROF. ADRIÃO DUARTE DÓRIA NETO, Dr., UFRN
Componente da Banca

PROF. HERMAN MARTINS GOMES, Dr., UFCG
Componente da Banca

PROF. FRANCISCO MARCOS DE ASSIS, Dr., UFCG
Componente da Banca

CAMPINA GRANDE – PB
Abril - 2004

Dedicatória

Dedico este trabalho a Deus em primeiro lugar, a minha esposa Maria José, aos meus filhos Emanuel e Camila, aos meus pais José Nilton (*in memorian*) e Carmelita e ao tio Pedro Paulo.

O temor do SENHOR é o princípio da ciência;
os loucos desprezam a sabedoria e a instrução.
Provérbios 1.7.

Agradecimentos

- A Deus, pelo dom da fé e da vida.
- Aos Professores Benedito e Auxiliadora pela incansável orientação e colaboração sem as quais não seria possível realizar este trabalho.
- A Universidade Federal do Maranhão e a CAPES que proporcionaram o suporte financeiro para viabilizar a realização deste trabalho.
- A minha família, pela paciência, apoio e incentivo sempre presentes.
- Aos professores do DEE-UFMA pelo apoio e incentivo neste trabalho.
- Aos meus amigos, Robson, Francisco Madeiro, Tomaz, Fabiano, Edmar, Albos e Jakobson, pelo apoio e incentivo neste trabalho.
- As minhas amigas, Fabrícia e Joseana que tanto me incentivaram no decorrer deste trabalho.
- As minhas sobrinhas Ana Danielle e Ana Gabrielle pelo apoio no desenvolvimento deste trabalho.
- Ao PET de Engenharia Elétrica da UFCG, sobretudo aos alunos Towar, Sérgio, Natasha, Murali, Mozart e Fídias, pelo apoio no desenvolvimento deste trabalho.
- A COPELE-UFCG, em especial à Ângela, à Pedrinho, e à Eleonora, pelo apoio constante.
- Ao Centro de Apoio Pedagógico à Deficientes Visuais do Estado do Maranhão, em especial a Diretora Josefa Lídia e aos alunos que contribuiram para o desenvolvimento deste trabalho.
- Aos membros da Banca Examinadora, Professores Rui Seara, Adrião Duarte Dória Neto, Francisco Marcos de Assis, Sidney Cerqueira Bispo dos Santos e Herman Martins Gomes, pela contribuição bastante significativa, que veio enriquecer o trabalho desenvolvido.
- A todos que, direta ou indiretamente, me incentivaram no decorrer deste trabalho.

Resumo

Este trabalho apresenta um modelo para geração automática da prosódia em um sistema texto-fala concatenativo para o Português Brasileiro. O modelo é baseado em regras e na tonicidade de palavras para determinar os contornos de entonação. Para tal, é realizada uma análise acústica em um *corpus* de palavras, contendo as mais diversas combinações de fonemas para as sílabas, e de frases foneticamente balanceadas, para identificação do comportamento da duração e sobretudo do *pitch*, ao longo de palavras oxítonas, paroxítonas e proparoxítonas com até cinco sílabas. Na primeira etapa do trabalho é apresentada a estrutura básica de um sistema texto-fala e destacada a importância de um modelo de prosódia para um sistema desse tipo. Na segunda etapa é apresentada a forma natural de produção da fala e conceitos importantes de lingüística. Na terceira etapa são apresentados os estágios de processamento lingüístico, processamento prosódico e processamento do sinal em um sistema texto-fala. Optou-se por um processamento lingüístico mais simples, contemplando os estágios de pré-processamento e transcrição fonética. Também optou-se pela síntese concatenativa, considerando-se as vantagens de simplicidade, flexibilidade e sobretudo porque o processamento do sinal de fala é feito na própria forma de onda, mantendo-se assim as características originais desse sinal. Na quarta etapa são apresentados o dicionário de unidades acústicas e o modelo prosódico com os resultados obtidos. São utilizadas sílabas e demissílabas como unidades acústicas do dicionário. A base de dados de unidades acústicas, relativamente elevada, incorpora informações prosódicas das mais relevantes, bem como das características articulatórias correspondentes aos fenômenos de coarticulação, para a obtenção de uma fala sintetizada de qualidade. A seleção das unidades é realizada através de uma estrutura de pesos atribuídos às sílabas tônicas, pretônicas e postônicas, considerando as curvas de entonação analisadas no *corpus*. O modelo foi avaliado através de testes informais de escuta em palavras e testes formais em um *corpus* de 20 frases foneticamente balanceadas, que constituem uma amostra representativa do universo de todos os fonemas e unidades do dicionário. Os testes com 20 frases foi realizado com 40 ouvintes e usando-se a escala MOS (*Mean Opinion Score*), obtendo-se um escore de 4,25, superior ao escore bom (4,0). A partir dos resultados obtidos, conclui-se que o modelo proposto pode ser aplicado a um sistema de síntese concatenativa com bons resultados, para palavras ou frases declarativas. Sugere-se que, em trabalhos futuros, o modelo possa ser ampliado para outros contornos de entonação como, por exemplo, contornos para frases interrogativas e exclamativas.

Abstract

This work presents a model for automatic generation of the prosody in a concatenative text-to-speech system for Brazilian Portuguese. The model is based on rules and on stressed syllables in the words, to determine the intonation contours. For such, an acoustic analysis is accomplished in a corpus of words, containing the most several combinations of phonemes for the syllables, and phonetically balanced sentences, for identification of the behavior of the duration and above all the pitch, along oxytons, paroxytons and proparoxytons words, with up to five syllables. In the first stage of the work, a basic structure of a text-to-speech system is presented and the importance of a prosody model for that kind of system is emphasized. The second stage presents the natural form of speech production and important concepts of linguistics. The third stage presents the linguistic processing, the prosodic processing and the signal processing levels in a text-to-speech system. A simpler linguistic processing was opted, contemplating the pre-processing and phonetic transcription levels. The concatenative synthesis was also opted, considered the advantages of simplicity, flexibility and, above all, because the processing of the speech signal is made in the own wave form, maintaining the original characteristics of the signal. The fourth stage presents the acoustic units dictionary and the prosodic model with the obtained results. Syllables and demissylables are used as acoustic units of the dictionary. The relatively large, database of acoustic units, incorporates the most important prosodic information, as well as the articulatory characteristics corresponding to the coarticulation phenomena, in order to obtain a good synthesized speech. The selection of the units is accomplished through a structure of weights attributed to the stressed, pre-stressed and post-stressed syllables, considering the curves of intonation analyzed in the corpus. The model was evaluated through informal listening tests in words and formal tests in a corpus of 20 phonetically balanced sentences, wich constitute a representative sample of the universe of all the phonemes and units of the dictionary. The tests with 20 sentences were accomplished with 40 listeners and using the MOS scale (Mean Opinion Score), obtaining a score of

4,25, above the good score (4,0). From the obtained results, it is concluded that the proposed model can be applied to a system of concatenative synthesis with good results, for words or declarative sentences. It is suggested that, in future works, the model can be enlarged for other kinds of intonation contours as, for instance, interrogative and exclamatory sentences contours.

Conteúdo

Lista de Tabelas

Lista de Figuras

Lista de Abreviaturas e Símbolos

A/D	:	Conversor Analógico Digital
AFLAPS	:	Alfabeto Fonético Desenvolvido no Laboratório de Automação e Processamento de Sinais
AMDF	:	*Average Magnitude Difference Funcion* (Função da Média de Diferenças de Amplitudes)
ASCII	:	*Americam Standard Code for Information Interchange* (Código Padrão Americano para Troca de Informações)
C	:	Consoante
DEE/UFCG	:	Departamento de Engenharia Elétrica da Universidade Federal de Campina Grande
DSPs	:	*Digital Signal Processors* (Processadores de Sinais Digitais)
FD - PSOLA	:	*Frequency Domain Pitch Syncronous Overlap Add* (Sobreposição e Soma em Sincronismo com o *Pitch* no Domínio da Freqüência)
FFT	:	*Fast Fourier Transform* (Transformada Rápida de Fourier)
HMMs	:	*Hidden Markov Models* (Modelos de Markov Escondidos)
IPA	:	*The International Phonetic Alphabet* (Alfabeto Fonético Internacional)
LAPS	:	Laboratório de Automação e Processamento de Sinais
LPC	:	*Linear Predictive Coding* (Codificação por Predição Linear)
MBE	:	*Multiband Excited* (Excitação Multibanda)
MBR - PSOLA	:	*Multiband Resynthesis Pitch Syncronous Overlap Add* (Ressíntese por Multibanda com Sobreposição e Soma em Sincronismo com o *Pitch*)
MBROLA	:	*Multiband Resynthesis Overlap and Add* (Ressíntese por Multibanda com Sobreposição e Soma)
MM	:	*Markov Model* (Modelo de Markov)

MOS	:	*Mean Opinion Score* (Escore de Opinião Média)
OLA	:	*Overlap Add* (Sobreposição e Soma)
PCM	:	*Pulse Code Modulation* (Modulação por Largura de Pulso)
PSOLA	:	*Pitch Syncronous Overlap Add* (Sobreposição e Soma em Sincronismo com o *Pitch*)
PUC-RJ	:	Pontifícia Universidade Católica do Rio de Janeiro
RELP	:	*Residual Excited Linear Predictive Coding* (Codificação por Predição Linear com Excitação Residual)
SAMPA	:	*Speech Assessment Methods Phonetic Alphabet* (Alfabeto Fonético Capaz de ser Lido pela Máquina)
SV	:	Semivogal
TD - PSOLA	:	*Time Domain Pitch Syncronous Overlap Add* (Sobreposição e Soma em Sincronismo com o *Pitch* no Domínio do Tempo)
UFCG	:	Universidade Federal de Campina Grande
UFRJ	:	Universidade Federal do Rio de Janeiro
UFRGS	:	Universidade Federal do Rio Grande do Sul
UFSC	:	Universidade Federal de Santa Catarina
UNICAMP	:	Universidade Estadual de Campinas
V	:	Vogal
VOCODER - LPC	:	*Voice Coder - LPC* (Codificador de Voz por Predição Linear)
F_0	:	Freqüência Fundamental
P_0	:	Período de *Pitch*.
D_{UR}	:	Duração Final do Segmento Fonético
D_{IN}	:	Duração Intrínseca do Segmento Fonético
D_{MIN}	:	Duração Mínima do Segmento Fonético
PRS	:	Percentagem de Redução do Segmento Fonético
μ_i	:	Média Associada a Distribuição Formada pelas Durações das Realizações do Fonema
σ_i	:	Desvio Padrão Associado a Distribuição Formada pelas Durações das Realizações do Fonema
$b_i(O(t))$	:	Probabilidade de Emitir o Símbolo $O(t)$ no Estado i, no Instante t
$\mathbf{T}_{01}$, $\mathbf{T}_{02}$, $\mathbf{T}_{03}$	:	Tempo de Ocorrência dos Comandos de Frases
$\mathbf{F}_{min}$	:	Freqüência Mínima

$\mathbf{A}_F$	:	Amplitude de um Impulso de Comando de Acento
$\mathbf{A}_{AC}$	:	Amplitude de um Impulso de Comando de Frase
$\mathbf{T}_{11}$, $\mathbf{T}_{12}$, $\mathbf{T}_{13}$	:	Instantes de Tempo de Inicialização de cada Pulso de Comando de Frase no Modelo de Fujisaki
$\mathbf{T}_{21}$, $\mathbf{T}_{22}$, $\mathbf{T}_{23}$	:	Instantes de Tempo de Finalização de cada Pulso de Comando de Frase no Modelo de Fujisaki
F_1, F_2, F_3, F_4, F_5	:	Valores Máximos de Freqüência Fundamental no Modelo de Estilização Acústica
V_1, V_2, V_3, V_4	:	Valores Mínimos de Freqüência Fundamental no Modelo de Estilização Acústica
IF	:	Valor Inicial de Freqüência Fundamental no Modelo de Estilização Acústica
FF	:	Valor Final de Freqüência Fundamental no Modelo de Estilização Acústica
$C^O(O_i, U_i)$	:	Custo Objetivo (Estimativa da Diferença entre a Unidade U_i do Dicionário e o Objetivo O_i o qual Supostamente a Representa)
$C^C(U_{i-1}, U_i)$	:	Custo Concatenação (Estimativa da Qualidade de Concatenação entre as Unidades Consecutivas U_{i-1} e U_i do Dicionário)
$C_j^O(O_i, U_i)$	:	p Sub-Custos Correspondentes as Diferenças dos Elementos dos Vetores da Unidade U_i do Dicionário e do objetivo O_i
w_j^O	:	Pesos dos p Sub-Custos do Objetivo O_i
$C_j^C(U_{1-i}, U_i)$	:	p Sub-Custos Correspondentes as Diferenças dos Elementos dos Vetores das Unidades U_{i-1} e U_i do Dicionário
w_j^C	:	Pesos dos q Sub-Custos da Unidade U_i do Dicionário
$V(z)$	:	Transformada z do Modelo do Trato Vocal
$U(z)$	:	Transformada z do Sinal de Excitação $u(n)$
$R(z)$	:	Transformada z do Modelo de Radiação $r(n)$
$R_n(z)$	:	Transformada z da Função de Transferência de um Ressonador na Síntese por Formantes
a_{1n}, a_{2n}, a_{3n}	:	Coeficientes da Função $R_n(z)$
f_n	:	Freqüência Central de Ressonância
B_n	:	Largura de Banda do Ressonador em Hertz
T	:	Período de Amostragem em Segundos
f_1, f_2, f_3	:	Três Primeiras Freqüências Formantes

$s(n)$	:	Sinal de Fala Digitalizado
$s_m(n)$	:	Sinais Elementares de Análise na Técnica PSOLA
$h_m(n)$	:	Janela de Hamming
α	:	Fator de Modificação da Duração das Unidades Acústicas
β	:	Fator de Modificação da Freqüência das Unidades Acústicas
$s_q(n)$	:	Sinais Elementares de Síntese
t_m	:	Marcas de Pitch de Análise
t_q	:	Marcas de Pitch de Síntese
δ_q	:	Atraso entre as Marcas de Pitch de Síntese e de Análise
$s_f(n)$	:	Sinal de Síntese da Fala com o Algoritmo PSOLA
$y(n)$	:	Amostra Atual do Sinal da Fala na Análise/Ressíntese LPC
P	:	Ordem do Preditor (Número de Amostras)
$e(n)$	:	Erro Residual do Sinal
a_k	:	k-ésimo Coeficiente do Preditor
$\tilde{y}(n)$	:	Estimativa da Amostra $y(n)$
N	:	Comprimento da Janela
$s_h(n)$	:	Componente Harmônico do Sinal de Fala
$r(n)$	:	Componente de Ruído do Sinal de Fala
k	:	Número de Harmônicos do Sinal de Fala
w_o	:	Freqüência Fundamental em rad/seg
kHz	:	kiloHertz
Pi	:	Palavra **i**
Pre1, Pre2, Pre3, Pre4	:	Sílabas Pretônicas nas Posições 1, 2, 3 e 4
Pos1, Pos2	:	Sílabas Postônicas nas Posições 1 e 2
D_0Nat	:	Duração do Segmento Fonético Correspondente à Sílaba da Palavra Natural
D_0Sint	:	Duração do Segmento Fonético Correspondente à Unidade do Dicionário
$FoNat$	:	Freqüência Fundamental do Segmento Fonético Correspondente à Sílaba da Palavra Natural
$FoSint$	:	Freqüência Fundamental do Segmento Fonético Correspondente à Unidade do Dicionário
K_{D_0}	:	Relação entre a Duração da Unidade do Dicionário e a Duração da Sílaba da Palavra Natural

K_{F_0} : Relação entre a Freqüência Fundamental da Unidade do Dicionário e a Freqüência Fundamental da Sílaba da Palavra Natural

E_i : Resultado do Escore para a Frase i Utilizando a Escala MOS

R_i : Freqüência Relativa de cada Escore para a Frase i Utilizando a Escala MOS

Capítulo 1

Introdução

A fala constitui-se na ação ou faculdade de falar, ou seja, de expressar-se ou exprimir-se por meio de palavras [1]. Também pode ser vista como um ato de criação individual da vontade e da inteligência.

A linguagem falada é o meio de comunicação mais fundamental para o ser humano, precedendo a linguagem escrita em termos tanto de surgimento histórico como de aprendizado da própria vida. A escrita nada mais é do que o registro simbólico da fala [2]. Apesar da escrita ser uma forma eficiente e necessária de se comunicar e preservar a informação, a fala se destaca em muitas aplicações sobretudo na comunicação vocal homem-máquina em que, por exemplo, uma mensagem de alarme ou advertência falada pode ser mais adequada do que uma mensagem escrita.

Nas últimas décadas o homem tem procurado reproduzir a fala artificialmente com o uso de sistemas computacionais. Esse procedimento, denominado síntese da fala, tem apresentado um grande desenvolvimento devido ao uso de novas tecnologias como, por exemplo, os Processadores de Sinais Digitais (DSPs), e computadores com maior capacidade de processamento e armazenamento de informações. Também o número de aplicações tem aumentado consideravelmente, tornando-se uma área atrativa para o desenvolvimento de sistemas de produção da fala, quer seja para fins de pesquisa ou comerciais.

Na comunicação vocal homem-máquina, vários sistemas de produção da fala têm sido desenvolvidos, porém os sistemas de síntese da fala apresentam várias vantagens devido à maior flexibilidade e capacidade de compressão da informação. De forma geral, esses sistemas podem ser divididos em dois grupos [3, 4, 5]:

- **Sistemas de Reprodução Vocal**. São sistemas que se baseiam no processo de concatenação de palavras isoladas ou partes de sentenças. Nesse caso, tem-se um melhor controle da qualidade e inteligibilidade da fala, mas em contrapartida só devem ser usados em aplicações que envolvam um vocabulário restrito (algumas centenas de palavras), tendo em vista que é impraticável armazenar em uma base de dados todas as palavras possíveis de serem originadas em uma determinada língua. Como exemplos, podem ser citados: anúncios de chegadas e partidas de vôos em aeroportos, saldo bancário por telefone, dentre outros.

- **Sistemas de Conversão Texto-Fala**. São sistemas que produzem a fala automaticamente a partir de um texto com base em algoritmos de processamento lingüístico e de sinais. Nesse caso, tem-se maior flexibilidade, pois o vocabulário pode ser irrestrito, de modo que não há necessidade de armazenamento de todas as possíveis palavras de uma língua. Na realidade são usadas unidades acústicas, através das quais as palavras são construídas. Dessa forma, tais sistemas se tornam mais complexos do que os anteriores, pois incluem um tratamento lingüístico para a obtenção de uma fala inteligível e natural.

Em geral, existem três fatores que influenciam o uso de sistemas de síntese da fala em determinada aplicação [6]:

- a qualidade da fala sintetizada (medida em termos de inteligibilidade, naturalidade e reconhecibilidade [1]);

- a abrangência do vocabulário, relacionada com a capacidade de produzir mensagens com diferentes palavras, ênfases, entonações, velocidades, etc.;

- complexidade (medida em termos de capacidade de processamento computacional e de armazenamento dos dados).

A Figura 1.1 apresenta a interação entre qualidade, abrangência de vocabulário e complexidade para sistemas de síntese da fala.

Um sistema de síntese ideal apresenta alta qualidade (a fala resultante é altamente inteligível e natural); mensagens com qualquer padrão de entonação e velocidade de fala requerida e baixa complexidade, de modo que pode ser integrado em qualquer

[1] Reconhecibilidade refere-se a produção da fala sintética mais próxima possível da fala natural do locutor [7].

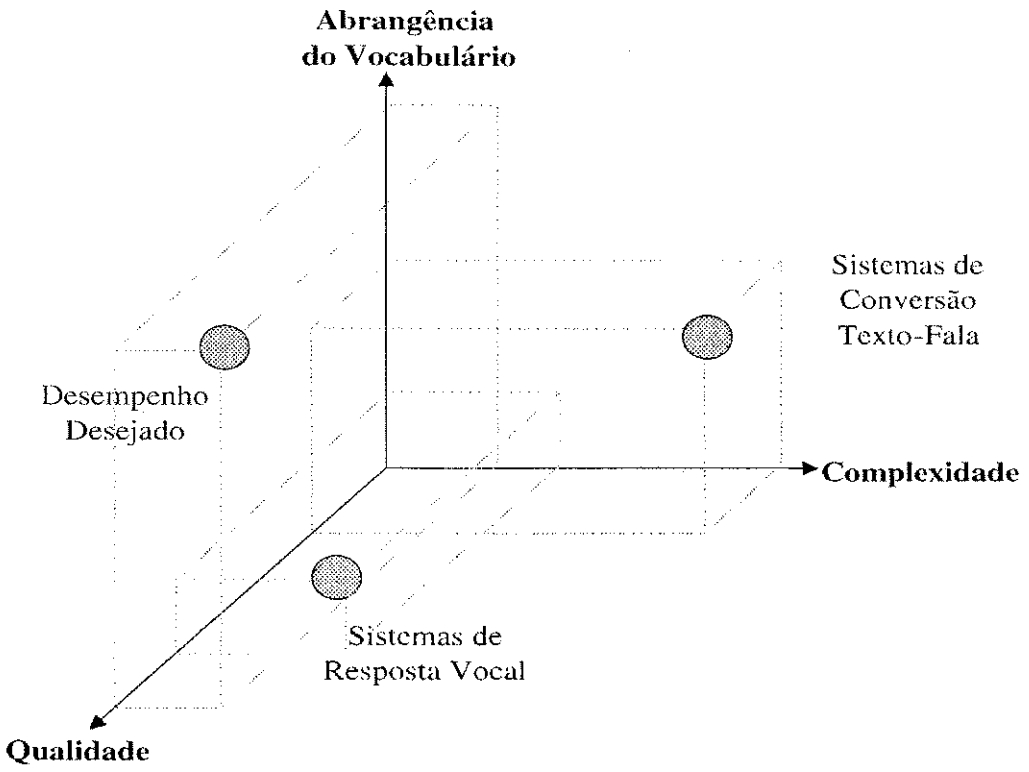

Figura 1.1: - Grupos de sistemas de síntese da fala em termos de qualidade, abrangência do vocabulário e complexidade [6].

ambiente de aplicação. Infelizmente, não existe atualmente um sistema que atenda simultaneamente aos três requisitos apresentados. Os sistemas de reprodução vocal têm uma alta qualidade e uma baixa complexidade, porém operam com um vocabulário limitado. Os sistemas de conversão texto-fala atuais têm uma complexidade e uma qualidade relativamente alta, e uma abrangência maior com relação ao vocabulário e, portanto, constituem-se na única alternativa para a conversão de um texto qualquer em fala.

A conversão texto-fala utilizando a concatenação de segmentos do sinal da fala tem apresentado melhores resultados do que as demais técnicas [8, 9], e por isso, tem sido mais usada atualmente. Evidentemente que a qualidade da fala produzida, nesse caso, também depende dos vários estágios utilizados no desenvolvimento e implementação do conversor. Dentre as características envolvidas na produção da fala tem-se dado um grande destaque atualmente à prosódia, a qual é relacionada ao ritmo e entonação, e, conseqüentemente, à qualidade da fala.

Considerando que a prosódia em um conversor texto-fala é o objeto de estudo deste trabalho, são apresentadas, neste capítulo, considerações básicas sobre a conversão texto-fala e suas aplicações típicas para a síntese da fala. Também é destacada a importância da prosódia dentro do contexto. Finalmente é feita uma apresentação sucinta do trabalho, incluindo os capítulos que contemplam as várias etapas para se chegar ao modelo prosódico proposto e a respectiva avaliação dos resultados alcançados, objetivando a geração da prosódia em conversores texto-fala para a Língua Portuguesa falada no Brasil a partir da tonicidade das palavras.

1.1 Sistemas de Conversão Texto-Fala

Os sistemas de conversão texto-fala não dispõem dos mesmos recursos do processo natural de leitura realizado pelo homem e, geralmente, são o resultado de uma imitação da capacidade de leitura humana, submetida a restrições tecnológicas e imaginativas que são características da sua época de criação. Apesar do estado atual do conhecimento humano, em função de técnicas e progressos em Processamento de Sinais, Inteligência Artificial e Lingüística, não se pode atribuir à máquina o mesmo efeito humano com relação à fala, pois o processo de leitura, que inclui além da decodificação do texto o seu significado, atua de forma mais profunda e envolve a inteligência humana. Além disso, o aperfeiçoamento do sistema implica na sua complexidade, a qual nem sempre é compatível com o critério econômico [10].

Em geral um sistema de conversão texto-fala pode ser representado pelo diagrama de blocos da Figura 1.2.

No modelo da Figura 1.2, o estágio de processamento do texto executa determinadas tarefas, tais como: escrever por extenso as siglas, números e abreviaturas, realizar uma análise gramatical e a transcrição fonética correspondente ao texto de entrada. Posteriormente, o estágio de processamento prosódico determina a duração e entonação correspondente aos fonemas, aplicados à sua entrada e produz na sua saída uma lista de fonemas com os parâmetros prosódicos correspondentes. Finalmente, o estágio de processamento do sinal transforma a informação simbólica recebida do processamento prosódico no sinal de fala. Nesse caso, devem ser definidas as etapas de cada estágio, como também devem ser desenvolvidos formalismos e algoritmos com base em conhecimentos lingüísticos e processamento de sinais para a obtenção da fala a partir de um texto. Esses estágios serão estudados com mais detalhes nos Capítulos 3, 4 e 5 deste trabalho.

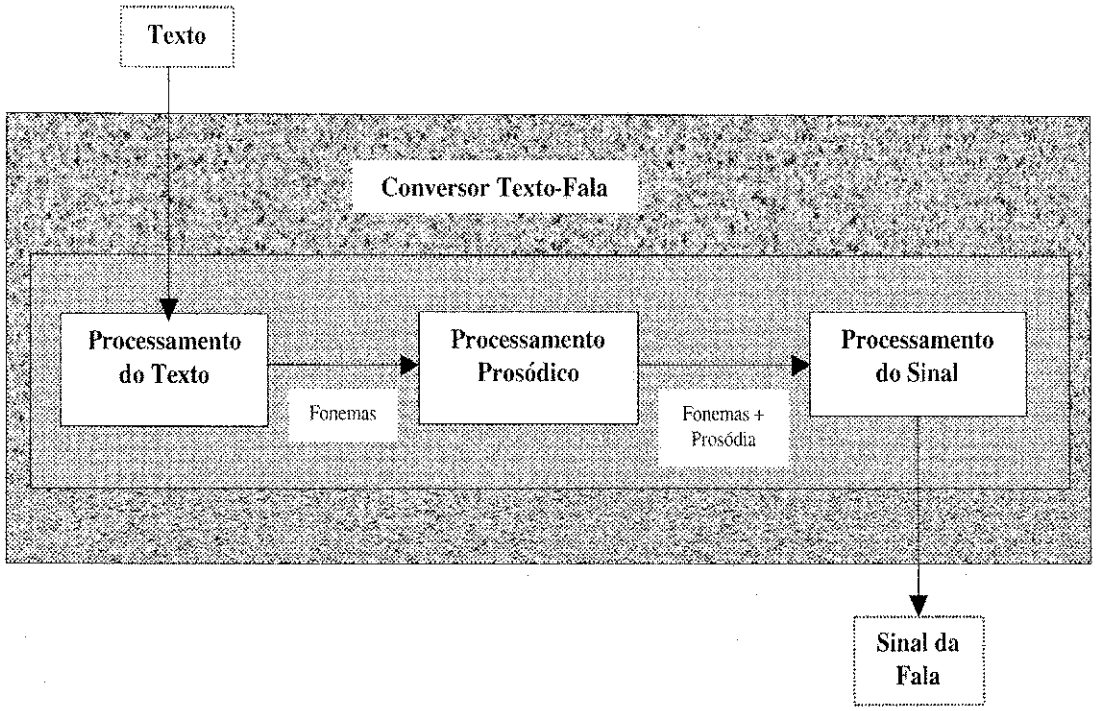

Figura 1.2: - Diagrama básico de um conversor texto-fala.

1.2 Aplicações Típicas de Conversores Texto-Fala

Os conversores texto-fala podem ser utilizados em várias aplicações, dentre as quais podem ser citadas:

- **Auxílio a deficientes visuais** - O auxílio a deficientes visuais é uma das mais importantes aplicações deste tipo de conversor. Com esta técnica é possível se obter uma máquina de leitura para cegos, que consiste na reprodução, sob a forma falada, de um texto qualquer impresso na forma escrita e escaneado ou digitado em um teclado [10]. Nesse sentido, já existem alguns conversores disponíveis para deficientes visuais, tais como: o *W Word*, desenvolvido pela Faculté Polytechnique de Mons (*http//tcts.fpms.ac.be*) [11], o DOSVOX, desenvolvido pela Universidade Federal do Rio de Janeiro (*http//www.nce.ufrj.br*) [12, 13], e o Virtual Vision, desenvolvido pela empresa brasileira MicroPower (*http//www.micropower.com.br*) [14], dentre outros.

- **Auxílio a deficientes vocais** - Deficiências vocais se originam devido a distúrbios de ordem motora ou mental. Um equipamento portátil contendo teclado, conversor texto-fala e estágio de saída sonora pode auxiliar a um portador com esse tipo de deficiência [11, 15].

- **Ensino de idiomas** - Um conversor texto-fala pode ser usado para o ensino de um novo idioma sobretudo acoplando-se ao mesmo um sistema de aprendizagem auxiliado por computador [10, 16].

- **Livros falantes** - Nesses sistemas as informações, na forma de texto, são armazenadas em um computador e podem ser recuperadas de forma auditiva quando solicitadas. Como exemplo, podem-se citar as enciclopédias e os dicionários falantes [17, 18].

- **Monitoramento vocal** - Nesses sistemas as instruções em forma de texto são armazenadas e podem ser usadas, de forma auditiva, em situações em que os olhos e as mãos precisam estar livres para realizarem outra tarefa. Essas instruções normalmente se referem à maneira de se usar ou de se dar manutenção em um determinado equipamento, e podem ser incorporados em sistemas de medição ou controle [10, 19].

- **Sistemas Remotos de Resposta Vocal** - Nesses sistemas é possível ter acesso a arquivos de texto através da resposta vocal, com o uso do telefone ou da internet. Esses arquivos podem variar de simples mensagens, como notícias sobre a chegada e partida de vôos, previsão meteorológica, programação de cinemas e teatros, cotações das bolsas de valores, etc. Questões a tais sistemas de informação podem ser feitas pelo usuário de forma vocal (com a ajuda de um sistema de reconhecimento da fala), ou através do teclado do telefone [10, 20].

- **Comunicação Vocal Homem-Máquina e Multimídia** - O desenvolvimento de conversores texto-fala de alta qualidade (como também sistemas de reconhecimento de fala mais robustos) é um passo importante para a comunicação mais completa entre o homem e os computadores. A multimídia é um primeiro passo neste sentido, porém, bastante promissor [21, 22].

- **Robótica** - O desenvolvimento de um sistema de reconhecimento e de síntese da fala, associado a um sistema de processamento de imagem, permite que um

robô seja capaz de reconhecer visualmente o seu usuário e interagir com ele verbalmente. Assim, é possível se obter robôs de diversos tipos, realizando tarefas domésticas ou no ambiente de trabalho, e interagindo com pessoas. Trabalhos nesse sentido estão sendo desenvolvidos pelo Grupo de Processamento da Fala e Robótica da UFRGS (Universidade Federal do Rio Grande do Sul) [23], pela Universidade de Aveiro em Portugal [24], e pelo Center for Spoken Language Research University of Colorado, USA [25], dentre outros.

- **Pesquisas fundamentais sobre a fala** - Um conversor texto-fala de alta qualidade pode se tornar uma ferramenta importante em um laboratório de lingüística para investigação da eficiência de modelos rítmicos e entonativos. Sistemas desse tipo baseados na descrição do trato vocal através das suas freqüências ressonantes, denominados sintetizadores por formantes, têm sido extensivamente usados por foneticistas para estudar a fala e seus fenômenos em termos de regras acústicas. Dessa forma, restrições articulatórias têm sido destacadas e formalmente descritas [10, 16].

1.3 A Prosódia em Conversores Texto-Fala

Uma oração (frase ou parte de uma frase que se estrutura em torno de um verbo ou de uma locução verbal [26]) na língua falada, conta com numerosos recursos para alcançar seu objetivo de unidade de comunicação. Em seu auxílio, além dos múltiplos recursos lingüísticos de entonação, há uma série de recursos extralingüísticos elocucionais, como: riso, suspiro, bocejo, etc., bem como os não elocucionais como a mímica. Pode-se então perceber que a fala não é um processo simples. Se a entonação e as pausas não fossem consideradas, o desenvolvimento de um sistema de conversão texto-fala não teria grande complexidade. Mas se esses fatores não fossem evidenciados, a fala produzida em tal sistema não apresentaria naturalidade, inteligibilidade e, poderia ser irreconhecível.

Desta forma, uma das maiores dificuldades que surgem no desenvolvimento e implementação de um conversor texto-fala consiste em determinar a prosódia, ou seja, a correta entonação (determinada pelas alterações da freqüência de vibração das cordas vocais durante a produção da fala), acentuação e duração da fala a partir de um texto escrito [27]. A determinação desses parâmetros não é uma tarefa trivial, pelo fato de não existir uma seqüência única de parâmetros prosódicos que possa ser associada a uma

determinada sentença. Obviamente existem características que constituem padrões da língua, como a presença de fronteiras prosódicas, curvas de entonação típicas para sentenças declarativas, imperativas, interrogativas, etc. [28]. Mesmo assim, a prosódia não possui uma estrutura fixa; cabe ao estágio de processamento prosódico, encontrar valores para os parâmetros prosódicos da sentença que a tornem o mais próximo possível de um enunciado da fala natural [29].

Para aplicar a prosódia em um conversor texto-fala é necessário definir um modelo prosódico, que determina a evolução temporal dos parâmetros prosódicos, de forma que seja possível identificar na fala a acentuação, o ritmo e a entonação [8]. Uma estrutura normalmente adotada é a separação do modelo prosódico em modelo de duração e modelo de entonação ou de *pitch*. No modelo de duração é realizado um tratamento automático no qual as durações dos fones de um enunciado possam ser determinadas [4, 30]. No modelo de entonação podem ser usadas várias técnicas dentre as quais se destacam:

- o uso de um contorno padrão, aplicado em todas as frases sintetizadas, nas quais são efetuadas alterações, a partir de um conjunto de regras baseadas em informações sintáticas e do conhecimento de fronteiras entre palavras e sílabas, além da informação sobre a posição da sílaba tônica de cada palavra [31, 32];
- o uso de contornos de entonação extraídos de elocuções naturais, que são ajustados às particularidades do texto a sintetizar [33, 34].

Apesar das dificuldades apresentadas, a inclusão de informações prosódicas em um conversor texto-fala resulta em uma fala com maior inteligibilidade e naturalidade [35, 36]. A inteligibilidade está relacionada à fidelidade com que são reproduzidos os sons surdos, ou seja, aos segmentos de mais baixa energia do sinal. Por outro lado, a naturalidade está relacionada à prosódia da fala, cujos fatores mais importantes são a fidelidade de reprodução dos contornos de freqüência fundamental nos segmentos sonoros e a correta reprodução da entonação dos referidos segmentos.

Vários métodos têm sido propostos para a geração da prosódia de forma automática, principalmente os baseados em regras [37, 38], em técnicas de aprendizagem com redes neurais [39, 40, 41], em árvores de classificação e regressão [42, 43, 44], ou em modelos de Markov escondidos (HMMs) [45, 46].

Na literatura são encontrados vários estudos para modelagem prosódica referente à entonação para a Língua Inglesa [31, 47], para a Língua Francesa [27, 38], para a

Língua Chinesa [16, 48] e para a Língua Portuguesa falada no Brasil [2, 4], dentre outros.

1.4 Motivação

Em geral, não existe um consenso sobre um modelo único de conversão texto-fala. Assim, surgem perspectivas para que vários tipos de conversores sejam desenvolvidos com técnicas diferentes, e na prática, são encontrados conversores texto-fala para várias línguas [11, 41, 47, 49, 50]. Para a Língua Portuguesa falada no Brasil algumas instituições de ensino e de pesquisa desenvolveram conversores deste tipo, como a PUC-RJ (Pontifícia Universidade Católica do Rio de Janeiro) [51] e UNICAMP (Universidade Estadual de Campinas - SP) [29, 52, 53, 54], com resultados satisfatórios segundo os autores. Outras instituições estão em fase de desenvolvimento de conversores como a UFCG (Universidade Federal de Campina Grande) e a UFSC (Universidade Federal de Santa Catarina), nas quais têm sido desenvolvidos relevantes trabalhos sobre o assunto [55, 56, 57, 58, 59, 60].

Relativamente à prosódia para a Língua Portuguesa, tem-se conhecimento de trabalhos realizados tratando isoladamente a duração ou *pitch* [28, 52, 61, 62, 63, 64], sobre técnicas de processamento de sinais para alteração de parâmetros prosódicos [8, 59], uma modelagem desenvolvida por Silva em [4], baseada no modelo de duração de Klatt para a Língua Inglesa [65] e um modelo entonacional baseado em constituintes prosódicos, e uma modelagem desenvolvida por Gomes em [2], na qual são usados contornos de duração e entonação obtidos a partir de dados extraídos da fala de um locutor.

Portanto, mesmo com os relevantes trabalhos realizados na área de conversão texto-fala, sobretudo para as Línguas Inglesa, Francesa, Japonesa e Chinesa, a questão da prosódia, nesses sistemas, tem sido objeto de muita atenção atualmente, pela sua importância quanto à obtenção de uma fala sintetizada com mais naturalidade e inteligibilidade [38, 49, 66, 67]. Assim, sobretudo para a Língua Portuguesa falada no Brasil, apesar também dos relevantes trabalhos realizados, existe ainda um campo de estudos bastante aberto em busca de uma conversão texto-fala com mais qualidade, constituindo-se, portanto, a modelagem prosódica uma área de pesquisa bastante atrativa.

1.5 Objetivo do Trabalho

O trabalho, aqui apresentado, trata do desenvolvimento de um modelo para geração automática da prosódia, baseado na tonicidade de palavras, em um conversor texto-fala para a Língua Portuguesa falada no Brasil, utilizando sílabas e demissílabas como unidades acústicas do dicionário. A base de dados das unidades acústicas, relativamente elevada, incorpora informações prosódicas das mais relevantes, bem como das características articulatórias correspondentes aos fenômenos de coarticulação, de forma a que se obtenha uma simplificação na etapa de processamento do sinal. O modelo aqui proposto de tonicidade é levado a efeito por palavras, considerando-se a complexidade que seria em uma abordagem por frases, permitindo assim que aspectos fonético-fonológicos sejam melhor explorados no âmbito de palavras, e posteriormente possam servir de suporte a futuros estudos de modelagens mais abrangentes. É observado que o modelo apresenta bons resultados, para um *corpus* de palavras e frases declarativas.

Para alcançar o objetivo proposto, foram desenvolvidas as seguintes etapas:

- Concepção de um dicionário de unidades acústicas, baseado em sílabas e demissílabas.
- Elaboração de regras para a normalização do texto, para a transcrição letra-fonema e para a separação dos segmentos fonéticos a serem rotulados no estágio de modelagem prosódica;
- Desenvolvimento de um modelo prosódico baseado na tonicidade de palavras, a partir de considerações fonético-fonológicas e da análise de um *corpus* de palavras oxítonas, paroxítonas e proparoxítonas com até cinco sílabas, e de um *corpus* de frases foneticamente balanceadas[2].
- Elaboração de regras para geração dos segmentos fonético-prosódicos para a síntese da fala.
- Desenvolvimento de um ambiente de testes, para análise de cada etapa da conversão texto-fala, como também da avaliação do modelo prosódico.

[2]Uma lista de frases é foneticamente balanceada quando a freqüência de ocorrência dos fones se aproxima de modo significativo daquela com que ocorrem na língua falada [68].

Portanto, após a etapa de geração e rotulamento dos segmentos fonético-prosódicos, em função do contexto em que se inserem, são buscadas as unidades acústicas correspondentes no dicionário e realizada a síntese da fala.

1.6 Organização do Trabalho

O presente capítulo tem por objetivo apresentar ao leitor uma visão geral sobre a síntese da fala, através da conversão texto-fala, destacando, sobretudo, a importância da prosódia no processo, que é o objeto de estudo deste trabalho. Os demais capítulos são apresentados de forma resumida nos itens descritos a seguir:

- Capítulo 2: são descritos o aparelho fonador e os tipos de sons produzidos. Também são apresentados alguns conceitos de Lingüística, necessários ao entendimento da prosódia.
- Capítulo 3: são apresentados os estágios básicos do processamento de um texto em um conversor texto-fala.
- Capítulo 4: são apresentados os principais modelos de duração e de entonação que podem ser utilizados no processamento prosódico de um conversor texto-fala.
- Capítulo 5: são apresentados os métodos mais usuais para realizar a síntese sonora a partir de informações fonéticas e prosódicas extraídas do texto a ser convertido em fala.
- Capítulo 6: é apresentado o desenvolvimento de um dicionário de unidades acústicas, contendo informações prosódicas necessárias à conversão texto-fala para a Língua Portuguesa.
- Capítulo 7: é apresentado um modelo prosódico para um conversor texto-fala concatenativo para a Língua Portuguesa, baseado na tonicidade de palavras e em um dicionário de unidades acústicas. Também é apresentado um ambiente de testes para a validação do modelo e os resultados obtidos.
- Capítulo 8: são apresentadas conclusões e sugestões importantes sobre o desenvolvimento de trabalhos futuros.
- Apêndice A: são apresentadas as unidades acústicas usadas no modelo prosódico.

- Apêndice B: são apresentadas as regras para a geração dos segmentos fonéticos.
- Apêndice C: são apresentadas as palavras usadas no modelo prosódico.
- Apêndice D: são relacionadas as frases usadas nos testes MOS.

Capítulo 2

Produção da Fala e Lingüística

Os sinais de fala resultam de uma seqüência de sons coordenados por regras de linguagem. O estudo científico da linguagem é denominado *Lingüística* [69] e a ciência que apresenta os métodos para a descrição, classificação e transcrição dos sons de fala, é denominada *Fonética* [70].

A produção da fala por um locutor também pode ser vista como o resultado de transformações que ocorrem em diferentes níveis: semântico, lingüístico, articulatório e acústico [71]. As diferenças nessas transformações surgem como diferenças nas propriedades acústicas do sinal de fala, que podem ser usadas no desenvolvimento de um modelo prosódico para um conversor texto-fala.

Assim, neste capítulo, é apresentada a forma natural de produção da fala pelo ser humano e os tipos de sons produzidos pelo aparelho fonador. Também são apresentados alguns conceitos básicos da Lingüística, considerados importantes para o entendimento e desenvolvimento de um modelo prosódico para um conversor texto-fala para a Língua Portuguesa.

2.1 O Aparelho Fonador

As partes do corpo humano utilizadas na produção da fala têm como função primária outras atividades como, por exemplo, mastigar, engolir, respirar ou cheirar. Entretanto, a produção de qualquer som de qualquer língua é feita com o uso de uma parte específica do corpo humano denominada *aparelho fonador* [70].

O aparelho fonador humano é constituído basicamente pelo sistema respiratório, pelo sistema fonatório e pelo sistema articulatório, conforme mostrado na Figura 2.1.

O sistema respiratório consiste dos pulmões, dos músculos pulmonares (diafragma), dos tubos brônquios e da tráqueia. Esse sistema encontra-se na parte inferior à laringe (onde está a glote), e produz um fluxo de ar que, passando através da laringe e do trato vocal e/ou nasal, dá origem à fala.

O sistema fonatório é constituído pela laringe. Nela encontram-se músculos estriados que podem obstruir a passagem da corrente de ar e são denominados *cordas vocais*. O espaço decorrente da não obstrução dos músculos laríngeos é chamado glote. A função primária da laringe é atuar como uma válvula que obstrui a entrada de comida nos pulmões por meio do abaixamento da epiglote. A epiglote é a parte com mobilidade que se localiza entre a parte final da língua (ao fundo da garganta) e acima da laringe.

O sistema articulatório consiste da faringe, da língua, do nariz, dos dentes e dos lábios, ou seja, do trato vocal e do trato nasal (parte superior à glote) [70].

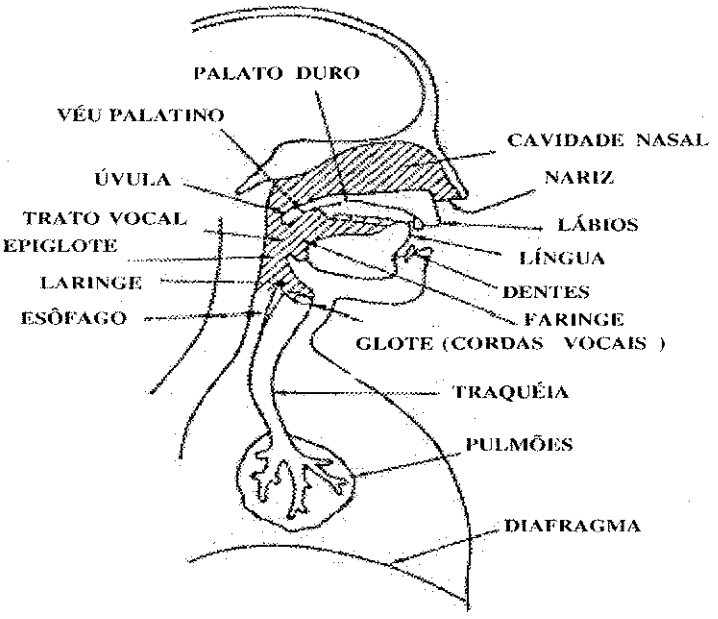

Figura 2.1: - Anatomia do aparelho fonador [8].

Para gerar os sons desejados, correspondentes a determinados fonemas, o ser humano exerce uma série de controles sobre o aparelho fonador, produzindo a configuração articulatória e a excitação apropriadas. O trato vocal, nome genérico dado ao conjunto de cavidades e estruturas que participam diretamente da produção sonora, começa na abertura entre as cordas vocais e termina nos lábios . O trato nasal começa na úvula e termina nas narinas. Quando a úvula é baixada, o trato nasal é acoplado acusticamente ao trato vocal para produzir os sons nasais da fala. Verifica-se que a forma do trato nasal não pode ser alterada voluntariamente pelo locutor. Após a filtragem, determinada pela conformação do aparelho fonador, o fluxo de ar injetado pelos pulmões é acoplado ao ambiente externo através dos orifícios dos lábios e/ou narinas [72].

2.2 Classificação dos Sons da Fala quanto ao Modo de Excitação

As características espectrais do sinal de fala são variantes no tempo, pois o sistema físico varia com o tempo. Como resultado, esse sinal pode ser dividido em segmentos que possuem propriedades acústicas semelhantes para curtos intervalos de tempo. A Figura 2.2, por exemplo, ilustra a forma de onda típica de um sinal de fala, correspondente à palavra *país*.

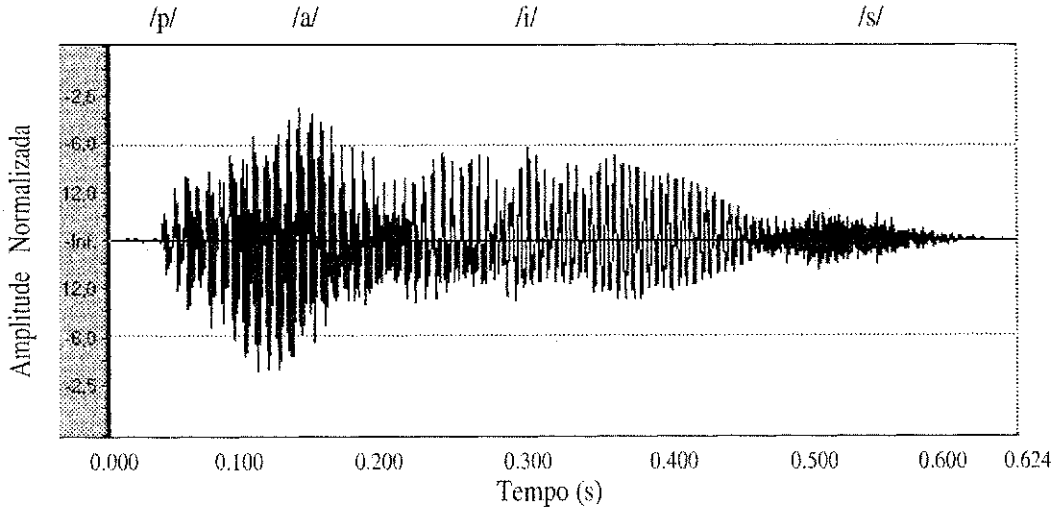

Figura 2.2: - Forma de onda no tempo da palavra *país*.

Na Figura 2.2, tem-se segmentos em que o sinal é quase periódico e segmentos em

que ele é aleatório. Assim, os sons da fala podem ser classificados em três classes distintas, conforme o modo de excitação. As classes são as seguintes [71]: sons sonoros, sons surdos e sons plosivos ou oclusivos.

2.2.1 Sons Sonoros

O fluxo de ar vindo dos pulmões é controlado pela abertura e fechamento das cordas vocais sob o controle do locutor. A abertura entre as cordas vocais é denominada de glote. Estando a glote totalmente fechada, o fluxo de ar proveniente dos pulmões é interrompido e a pressão sub-glotal aumenta até que as cordas vocais sejam separadas, liberando o ar pressionado, gerando um pulso de ar de curta duração. Com o escoamento do ar, a pressão glotal é reduzida, permitindo uma nova aproximação das cordas vocais. O processo se repete de uma forma quase periódica. Assim, são obtidas ondas de pressão, quase periódicas, excitando o trato vocal, que atuando como um ressonador modifica o sinal de excitação, produzindo freqüências de ressonância, denominadas de *formantes*, que caracterizarão os diferentes sons sonoros [72].

As vogais, cujo grau de nasalização é determinado pelo abaixamento da úvula, são exemplos típicos de sons sonoros. A Figura 2.3 mostra a forma de onda da vogal /a/, na palavra *país*. Algumas consoantes, como /l/ e /m/, também são produzidas com a excitação glotal.

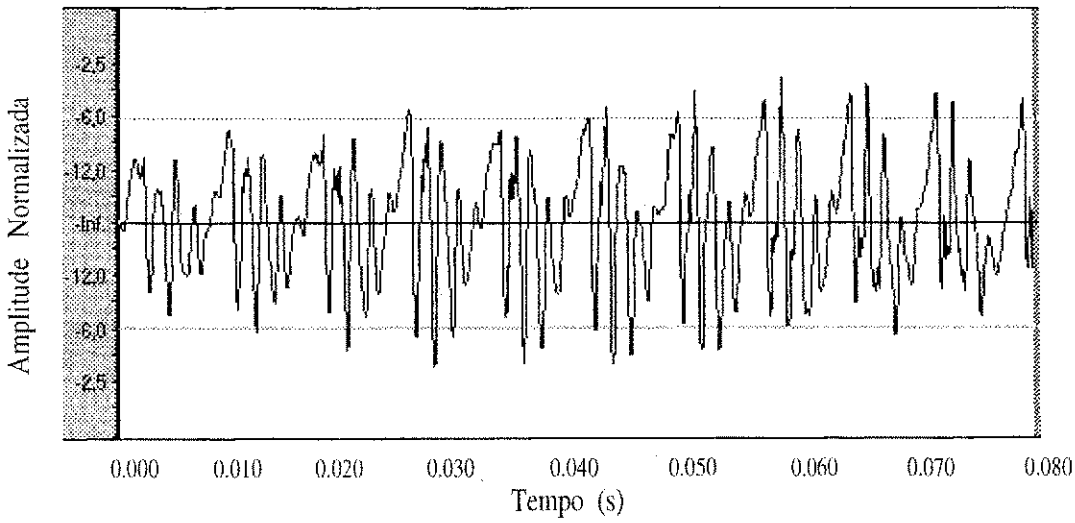

Figura 2.3: - Forma de onda correspondente ao fonema /a/ na palavra *país*.

A freqüência média de abertura e fechamento da glote é denominada *freqüência fundamental*, F_0, e o período correspondente, P_0, é denominado período de *pitch*.

A freqüência fundamental de sons sonoros possui um valor entre 70 e 200 Hz para homens, entre 150 e 400 Hz para mulheres e entre 200 e 600 Hz para crianças [10].

2.2.2 Sons Surdos

Os sons surdos são gerados pela produção de uma obstrução em algum ponto do trato vocal (normalmente próximo aos dentes e lábios). Assim, o ar adquire velocidade suficiente para produzir turbulência gerando um ruído de espectro largo (semelhante ao ruído branco) para excitar o trato vocal.

Na produção desses sons a glote permanece aberta, não havendo vibração das cordas vocais. Por exemplo, na produção do fonema /s/ na palavra *país* (Figura 2.4), os lábios e os dentes são ligeiramente pressionados, deixando assim uma passagem estreita para o ar, produzindo um fluxo de ar turbulento nas imediações da obstrução, o qual excita as cavidades do trato vocal. O som produzido dessa forma tem concentração relativa de energia nos mais altos componentes de freqüência do espectro do sinal de fala [71].

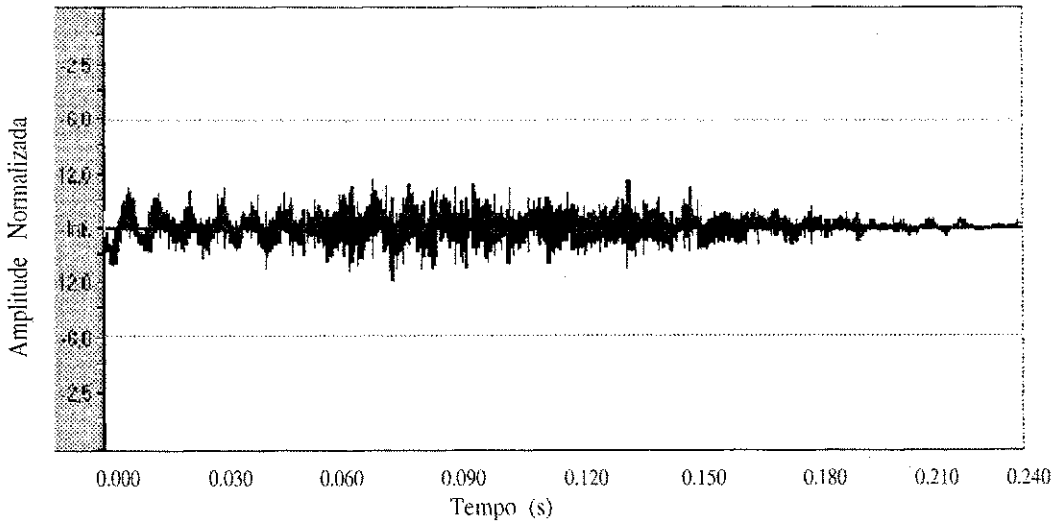

Figura 2.4: - Forma de onda correspondente ao fonema /s/ na palavra *país*.

2.2.3 Sons Plosivos ou Oclusivos

Na geração dos sons plosivos ou oclusivos, o ar é totalmente dirigido à boca, ocorrendo oclusão total. Com o aumento da pressão, a oclusão é rompida bruscamente, gerando um pulso que excita o aparelho fonador. Com a excitação ocorre um movimento rápido dos articuladores em direção à configuração do som seguinte. Sons plosivos são obtidos dos fonemas /p/, /t/, /k/, dentre outros [71]. A Figura 2.5 mostra a forma de onda correspondente ao fonema plosivo /p/, em *país*.

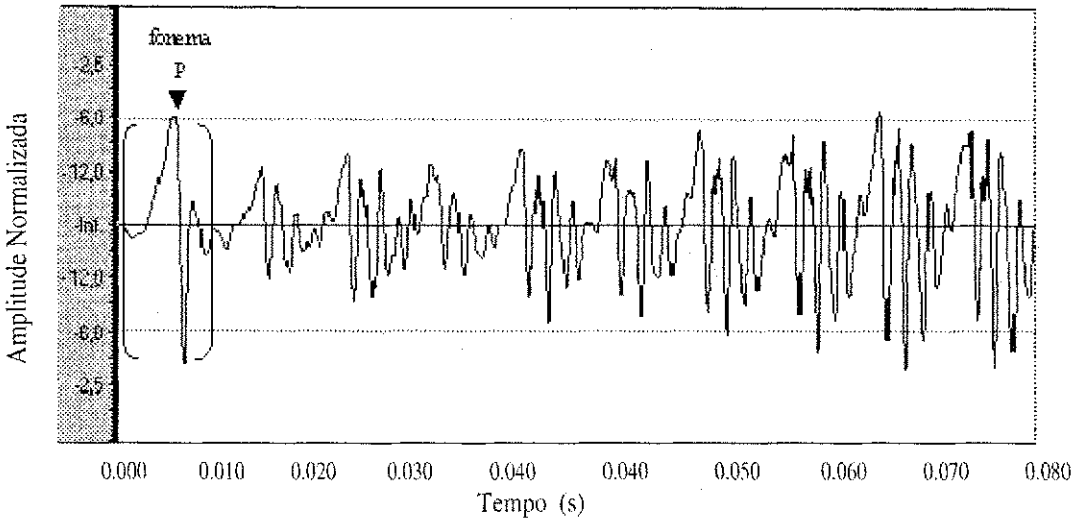

Figura 2.5: - Forma de onda correspondente aos fonemas /p/ e /a/ na palavra *país*, com destaque para o /p/.

2.2.4 Sons com Excitação Mista

Os sons fricativos sonoros, correspondentes aos fonemas /j/, /v/ e /z/, são produzidos combinando-se a vibração das cordas vocais com a obstrução no trato vocal. Nos períodos em que a pressão glotal atinge um máximo, o escoamento através da obstrução torna-se turbulento, gerando o caráter fricativo do som; quando a pressão glotal é reduzida abaixo de um dado valor, cessa o escoamento turbulento do ar e as ondas de pressão apresentam comportamento mais suave [71]. A Figura 2.6 mostra a forma de onda do som fricativo sonoro correspondente ao fonema /z/, obtido da palavra *anzol*.

Os sons plosivos sonoros, correspondentes aos fonemas /d/ e /b/, são produzidos de forma semelhante aos sons plosivos surdos, correspondentes aos fonemas /t/ e /p/,

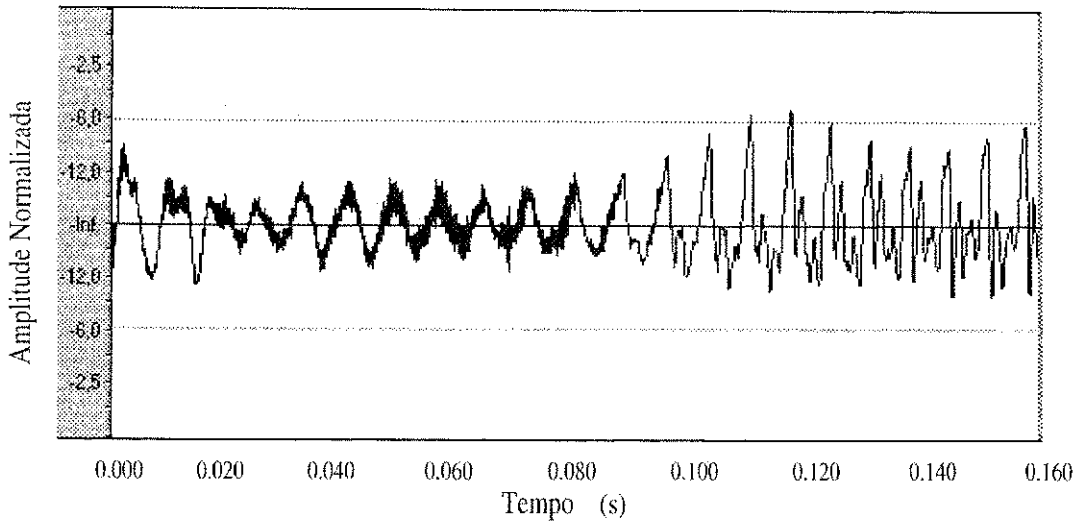

Figura 2.6: - Forma de onda correspondente ao fonema /z/ na palavra *anzol*.

porém há vibração das cordas vocais durante a fase de fechamento da cavidade oral.

A Figura 2.7 mostra a forma de onda do som plosivo sonoro correspondente ao fonema /b/, obtido na palavra *botão*.

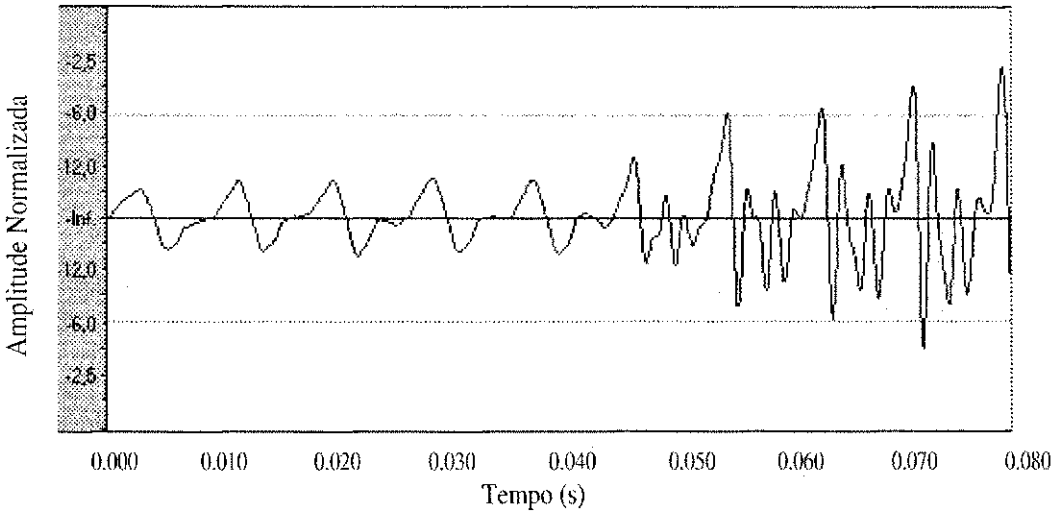

Figura 2.7: - Forma de onda correspondente ao fonema /b/ na palavra *botão*.

Assim, a partir da forma de onda do sinal da fala (gráfico da amplitude em função do tempo) é possível avaliar características importantes que permitem uma descrição do mesmo, tais como os parâmetros temporais. Dentre os parâmetros temporais destacam-se: Energia do Sinal, Taxa de Cruzamento por Zero, Coeficiente de Correlação Nor-

malizado e Número Total de Picos [73].

2.3 Fonética e Fonologia

Fonética e Fonologia são conceitos diretamente relacionados aos sons da fala e podem ser utilizados no desenvolvimento das várias etapas de um conversor texto-fala.

Enquanto a Fonética estuda os sons como entidades físico-articulatórias isoladas, a Fonologia estuda os sons sob o ponto de vista funcional como elementos que integram um determinado sistema lingüístico [74]. Assim, à Fonética cabe descrever os sons da linguagem e analisar suas particularidades articulatórias, acústicas e perceptivas. À Fonologia cabe estudar os elementos fônicos que distinguem em uma mesma língua, duas mensagens de sentido diferente (a diferença fônica no início das palavras do português *bala* e *mala*, a diferença de posição do acento, no português, entre *sábia*, *sabia* e *sabiá*, etc.), e aqueles que permitem reconhecer uma mensagem igual através de realizações individuais diferentes (voz diferente, pronúncia diferente, etc.) [69].

2.3.1 Fonemas

O fonema é a unidade mínima da Fonologia, ou seja, é a unidade mínima distintiva no sistema sonoro de uma língua [1].

Normalmente existe uma forte tendência de se confundir fonemas com letras, porém fonema é uma realidade acústica que nosso ouvido registra, enquanto letra é o símbolo empregado para representar na escrita o sistema sonoro de uma língua [75]. Assim, não existe identidade perfeita, muitas vezes, entre fonemas e letras, tornando impossível a obtenção de uma ortografia ideal. Como exemplo temos sete vogais orais tônicas, mas só existem cinco símbolos gráficos: *a, e, i, o, u*. Para distinguir um [e], tônico aberto, de um [e], tônico fechado, usamos o acento agudo (*fé*) ou circunflexo (*vê*). Há letras que se escrevem, mas por razões etimológicas não se pronunciam, e portanto não representam fonemas; é o caso do *h* em *homem*. Por outro lado, há fonemas que se ouvem e que não se acham registrados na escrita; assim, no final de *cantavam*, ouvimos um ditongo em (*am*), cuja semivogal não vem assinalada. Devido ao convencionalismo tradicional, a escrita nem sempre acompanha a evolução fonética.

Apesar de serem considerados unidades abstratas, os fonemas são definidos em termos de propriedades ou traços, que fazem a mediação entre a descrição lingüística

abstrata e a descrição fonética, a qual pode ser caracterizada acusticamente. A fim de distingüi-los dos sons realmente produzidos, os fonemas são representados entre barras (/ /), enquanto os sons são representados entre colchetes ([ ]). A palavra **tia**, por exemplo, é representada pelos fonemas */tia/* e pode ser pronunciada [*tʃia*], onde tʃ tem o som africado.

Os fonemas também são chamados de *segmentos* e, por extensão, as cadeias fonéticas, como, por exemplo, as sílabas, são consideradas no nível segmental da fala. Por outro lado, na produção da fala contínua são acrescentadas determinadas características nas cadeias fonéticas, como, por exemplo, a acentuação, o ritmo e a entonação, que determinam o nível suprasegmental, também chamado *nível prosódico*.

2.3.2 Fones e Alofones

O fone é a unidade da Fonética, ou seja, um fone é a realização acústica de um fonema [1]. Um fone não é uma classe de som, como um fonema, mas sim um sinal sonoro.

Um mesmo fonema pode produzir diferentes fones, isto é, um falante ao pronunciar um mesmo fonema diversas vezes produzirá sinais sonoros distintos a cada pronúncia. Contudo, essas distinções apresentarão um grau de semelhança que será suficiente para classificá-los como realizações acústicas de um mesmo fonema.

Na produção de sucessivos fones, os movimentos do trato vocal que se sucedem no tempo influenciam uns aos outros. Sendo assim as características acústicas e articulatórias dos fones produzidos são influenciados pelo contexto fonético em que eles ocorrem. Esse fenômeno é denominado de *coarticulação*.

Alofones também constituem a realização acústica dos fonemas e são os variantes de enunciação que cada fonema pode apresentar, ou seja, é a maneira como realmente são pronunciados. Como, por exemplo, na palavra *tia*, o fonema /t/ pode ser realizado como [tʃ] ou [t], dependendo do estado ou região do Brasil. Convém ressaltar que, embora existam pronúncias diferentes, não há permuta de fonema, o que não acarreta mudança de signo lingüístico.

Portanto, os fones e alofones têm características relacionadas aos fonemas e são importantes para a definição do tipo de fala sintetizada em um conversor texto-fala.

2.3.3 Classificação dos Fonemas

As palavras da Língua Portuguesa podem apresentar três tipos de fonemas [26, 75]: **vogais**, **consoantes** e **semivogais**.

Vogais

As vogais são fonemas produzidos por uma corrente de ar vibrante que passa livremente pela boca, proveniente dos pulmões. Elas podem ser classificadas em função do posicionamento dos músculos que delimitam a boca (língua, lábios e véu palatino), e também em função da intensidade, como relacionado a seguir.

1. Pela modificação no posicionamento do véu palatino:
   (a) **Orais (/a/, /é/, /ê/, /i/, /ó/, /ô/, /u/)** - A corrente de ar vibrante passa apenas pela cavidade bucal;
   (b) **Nasais (/ã/, /ẽ/, /ĩ/, /õ/, /ũ/)** - A corrente de ar vibrante passa ao mesmo tempo pelas cavidades bucal e nasal.
2. Pela altura da língua:
   (a) **Altas (/i/, /ĩ/ /u/, /ũ/)** - São produzidas quando a elevação da língua em direção ao palato duro é máxima;
   (b) **Médias (/e/, /ẽ/ /o/, /õ/)** - São produzidas com uma elevação média da língua em direção ao palato duro;
   (c) **Baixas (/a/, /ã/)** - São produzidas com uma elevação mínima da língua em direção ao palato duro.
3. Pela anterioridade/posteridade da língua:
   (a) **Anteriores (/ĩ/, /i/ /ẽ/, /ê/, /é/)** - São produzidas quando a ponta da língua se eleva em direção ao palato duro, determinando uma diminuição da abertura bucal e um aumento da abertura da faringe;
   (b) **Centrais (/ã/, /a/)** - São produzidas com a boca ligeiramente aberta e a língua na posição quase de repouso;
   (c) **Posteriores (/ũ/, /u/ /õ/, /ó/, /ô/)** - São produzidas quando o dorso da língua se eleva, recuando em direção ao véu palatino, o que provoca uma diminuição da abertura bucal e um arredondamento progressivo dos lábios.

4. Pela intensidade:

   (a) **Tônica** - vogal tônica é aquela em que recai o acento tônico da palavra. Exemplos: av**ó**, p**a**to, t**í**mido.

   (b) **Átona** - é a vogal não acentuada. Exemplos: **a**v**ó**, p**a**to, t**í**mido. as vogais átonas podem estar antes da tônica (pretônicas): **a**v**ó**, p**a**gar, ou depois (postônicas): t**í**mido.

Na Tabela 2.1, são relacionadas as vogais orais e nasais do Português Brasileiro. Na coluna da esquerda, tem-se a lista referente à altura da língua e na parte superior, tem-se a lista referente à anterioridade/posteridade da língua.

Tabela 2.1: Tabela fonética das vogais orais e nasais para o Português Brasileiro [70]

		Anteriores	Centrais	Posteriores
Altas		/i/ /ĩ/		/u/ /ũ/
Médias	Fechadas	/ê/ /ẽ/		/ô/ /õ/
	Abertas	/é/		/ó/
Baixas			/a/ /ã/	

Consoantes

As consoantes são fonemas em cuja produção a corrente de ar proveniente dos pulmões enfrenta obstáculos ao passar pela cavidade bucal. Esses obstáculos podem ser totais ou parciais, dependendo da posição da língua e dos lábios. Assim, elas são classificadas da seguinte maneira:

1. **Quanto ao modo ou maneira de articulação**:

   (a) **Oclusiva (/p/, /t/, /k/, /b/, /d/, /g/)** - Os articuladores produzem uma obstrução completa da passagem da corrente de ar através da boca. O véu palatino está levantado e o ar que vem dos pulmões encaminha-se para a cavidade oral. Nesse caso, tem-se as oclusivas sonoras (/b/, /d/, /g/) e as oclusivas surdas (/p/, /t/, /k/). Exemplos com estes tipos de fonemas podem ser encontrados nas palavras: **b**ola, **d**ado, **g**ato, **p**ato **t**ato, **c**arro;

(b) **Nasal (/m/, /n/ e /nh/)** - Os articuladores produzem uma obstrução completa da passagem da corrente do ar através da boca. O véu palatino encontra-se abaixado e o ar que vem dos pulmões dirige-se às cavidades nasal e oral. Exemplos com esses tipos de fonemas podem ser encontrados nas palavras: **m**ala, **n**uca, ba**nh**o.

(c) **Fricativa (/f/, /s/, /x/, /v/, /z/, /j/)** - Os articuladores se aproximam produzindo fricção quando ocorre a passagem central da corrente de ar. A aproximação dos articuladores entretanto não chega a causar obstrução completa e sim parcial que causa a fricção. Nesse caso, tem-se as fricativas sonoras (/v/, /z/, /j/) e as fricativas surdas (/f/, /s/, /x/). Exemplos com estes tipos de fonemas podem ser encontrados nas palavras: **f**eto, **s**apo, **x**ácara, **v**ala,; **z**apata, **j**aca;

(d) **Africada (/tʃ/, /dj/)** - Na fase inicial da produção de uma africada os articuladores produzem uma obstrução completa da passagem da corrente de ar através da boca e o véu palatino encontra-se levantado (como nas oclusivas). Na fase final dessa obstrução (soltura da oclusão) ocorre então uma fricção decorrente da passagem central da corrente de ar (como nas fricativas). As consoantes africadas podem ocorrer em algumas variedades do Português Brasileiro, como, por exemplos, em **t**ia (/tʃia/) e **d**ia (/djia/).

(e) **Tepe** (ou vibrante simples - /ɾ/) - O articulador ativo toca rapidamente o articulador passivo ocorrendo uma rápida obstrução da passagem de corrente de ar através da boca. O tepe ocorre em Português nos seguintes exemplos: ca**r**a, bra**v**a.

(f) **Vibrante** (múltipla - /rr/) - O articulador ativo toca algumas vezes o articulador passivo causando vibração. Em alguns dialetos de Português ocorre essa variante, em palavras como ma**rr**a e ca**rr**o.

(g) **Retroflexa - (/r/)** - O palato duro é o articulador passivo e a ponta da língua é o articulador ativo. A produção de uma retroflexa se dá com o levantamento e o encurvamento da ponta da língua em direção ao palato duro. Ocorre, por exemplo, em palavras como: ma**r** e ca**r**ta, no dialeto 'caipira' do Português Brasileiro.

(h) **Lateral (/l/, /ʎ/)** - O articulador ativo toca o articulador passivo e a corrente de ar é obstruída na linha central do trato vocal. O ar será então

expelido por ambos os lados desta obstrução tendo portanto a saída lateral. quando o obstáculo é parcial e o ar passa pelos lados da cavidade bucal. Ocorre, por exemplo, em palavras como: pal**lh**a e lata;

2. **Quanto ao lugar de articulação:**

   (a) **bilabial (/p/, /b/, /m/)** - Quando ocorre um contato dos lábios superior e inferior. Exemplos: **p**ata, **b**oa, **m**ala;

   (b) **labiodental (/f/, /v/)** - Quando o lábio inferior toca os dentes incisivos superiores. Exemplos: **f**aca, **v**ala;

   (c) **dental-alveolar (/t/, /d/, /s/, /z/, /n/, /l/)** - Quando o ápice ou a lâmina da língua toca a face interna dos dentes incisivos superiores. Exemplos: **d**ata, sapa, Za**t**a, **n**ada, la**t**a;

   (d) **alveopalatal (/tʃ/, /dj/)** - Quando a parte posterior da língua toca a parte medial do palato duro. Exemplos: **t**ia e **d**ia (no dialeto carioca);

   (e) **palatal (/x/, /j/, /$\tilde{\text{a}}$/, /$\tilde{\text{ɲ}}$/)** - Quando a parte média (dorso) da língua toca o palato duro, ou céu da boca. Exemplos: ba**nh**a, pa**lh**a;

   (f) **velar (/k/, /g/, /rr/)** - Quando a parte posterior da língua toca o palato mole ou véu palatino. Exemplos: **c**asa, **g**ata, **r**ata.

A relação entre os tipos de classificação para os fonemas consoantais, pode ser visualizada na Tabela 2.2. Na coluna da esquerda, tem-se a lista referente ao modo ou a maneira de articulação e na parte superior, tem-se a lista referente ao lugar de articulação.

Semivogais

As semivogais são fonemas produzidos de forma semelhante às vogais altas /i/ e /u/, mas diferem destas por não assumirem papel central em uma sílaba, ou seja, acompanham sempre uma vogal, com a qual formam a sílaba. Na escrita são representadas pelas letras **(i)** e **(u)**, como, por exemplo, em 'mais' e 'ma**u**'. Também, podem ser representadas pelas letras **e** e **o**, com o som de [i] e [u], como, por exemplo, nas palavras mã**e** e mág**oa**. Em alguns estados do Brasil como por exemplo, Paraíba. Pernambuco e Maranhão, tem-se os fonemas /l/ e /m/ no final de uma palavra com o som de [u] e formando um ditongo, como, por exemplo, nas palavras sol e fora**m**.

Tabela 2.2: Tabela fonética consonantal para o Português Brasileiro [70]

	Bilabial		Labiodental		Dental-alveolar		Alveopalatal		Palatal	Velar	
Oclusiva	/p/	/b/			/t/	/d/				/k/	/g/
Fricativa			/f/	/v/	/s/	/z/	/ ʃ /	/ʒ/			
Nasal	/m/				/n/				/ ɲ /		
Vibrante					/rr/						
Flape					/r/						
Lateral					/l/				/ ʎ /		

2.4 Prosódia

Prosódia pode ser definida como a parte da Fonética que trata da acentuação e entonação dos fonemas nas palavras e frases [75]. Portanto, a sua principal atribuição é destacar as sílabas predominantes (sílabas tônicas) na produção das palavras, o ritmo e a entonação na produção de frases.

Assim, a prosódia está relacionada a determinadas propriedades do sinal da fala, tais como mudanças audíveis na freqüência de vibração das cordas vocais (*pitch*), volume do som (energia do sinal) e duração das sílabas. O conjunto de características prosódicas também pode incluir outros fatores relacionados à duração da fala, tais como o ritmo (determinado pelo tempo de duração das sílabas acentuadas) e a taxa da fala (número de fonemas por segundo).

2.4.1 Níveis de Representação da Prosódia

Geralmente os modelos utilizados no processamento prosódico podem ser estudados ou representados em um dos três níveis relacionados abaixo [10]:

- **Nível Acústico.** Em nível acústico é possível obter uma representação numérica da prosódia na forma de uma seqüência de valores correspondentes à freqüência fundamental, duração e energia. Modelos de duração e entonação nesse nível são apresentados no Capítulo 4.

- **Nível Perceptivo.** Em nível perceptivo é produzida uma descrição quantitativa e compacta de eventos prosódicos percebidos no sinal da fala por determinado

ouvinte (em valores médios). Um estudo nesse nível decorre do fato de que as propriedades espectrais dos sons da fala e as características acústicas podem ser medidas, mas nem sempre percebidas. Nesse sentido foram desenvolvidos vários modelos como, por exemplo, o modelo de percepção de d'Alessandro e Mertens, apresentado em [76], que produz uma estilização automática dos contornos de freqüência fundamental de uma frase a partir dos contornos percebidos nas sílabas das palavras.

- **Nível Lingüístico.** Em nível lingüístico a prosódia de uma palavra ou frase é representada por uma sequência de unidades abstratas (signos ou símbolos), algumas das quais têm uma função comunicativa na fala, enquanto outras são adequadas à estrutura sintática. Um estudo nesse nível é mais complexo que os anteriores, pois a informação prosódica é relativa, em oposição aos valores absolutos de duração, freqüência fundamental e energia. Um exemplo de modelo prosódico para a língua inglesa em nível lingüístico foi desenvolvido por Black e Hunt em [77]. Nesse modelo, o tipo de evento prosódico e a sua localização são dados, sendo necessária apenas uma predição do contorno de F_0. São usados símbolos, como, por exemplo, B (*Break*) ou BB (*Break-Break*), para representar a pontuação de frases, e L (*Low*) e H (*High*) para representar o tipo de acento de sílabas dentro de cada frase. Além do tipo de acento de cada sílaba, são considerados fatores tais, como o tipo de acento das duas sílabas anteriores e das duas posteriores, o número de sílabas decorridas desde o início da frase, o número de sílabas que faltam para o final da frase, dentre outros. É utilizada a técnica de regressão linear para modelar os valores de F_0 no início, no meio e no final de cada sílaba.

Assim, em um conversor texto-fala, o modelo lingüístico é desenvolvido em função das características gráficas de determinada língua relacionadas à fala, enquanto que o modelo acústico é desenvolvido com base em processamento de sinais de fala e o modelo perceptivo é usado em nível de avaliação do conversor como um todo.

A Tabela 2.3 apresenta uma relação entre as características ou parâmetros básicos de cada um dos níveis de representação da prosódia. Esta associação pode ser usada como referência no desenvolvimento de um modelo prosódico como também na comparação desse com os demais.

Tabela 2.3: Relação entre as características dos modelos de transcrição da prosódia

Nível Acústico	Nível Perceptivo	Nível Lingüístico
Freqüência Fundamental (F_0)	*Pitch*	Tom, entonação, acentuação
Energia	Intensidade	Acentuação
Duração	Comprimento	Acentuação

2.4.2 Parâmetros Prosódicos e Acentuação

As alterações dos parâmetros prosódicos em nível acústico têm um destaque especial na produção da fala [10]. Assim, a freqüência fundamental, a duração e a energia, juntamente com os conceitos de acentuação são abordados nas subseções seguintes, considerando-se a importância desses parâmetros e o fato de que o modelo prosódico, desenvolvido neste trabalho, é realizado em nível acústico e baseado na tonicidade das palavras.

Freqüência Fundamental

A freqüência fundamental, F_0, é o parâmetro prosódico mais envolvido na entonação e, como foi descrito anteriormente, corresponde à periodicidade da forma de onda do sinal sonoro. Assim, não faz sentido falar em freqüência fundamental para segmentos de fala não-sonoros, pois nesse caso não ocorre vibração das cordas vocais, e a forma de onda tem características aperiódicas.

Geralmente o período de *pitch*, ou simplesmente o conceito de *pitch*, é associado a freqüência fundamental e na literatura sobre síntese de fala os dois termos constumam ser utilizados de forma equivalente (apesar do período ser o inverso da freqüência). Na realidade enquanto F_0 envolve medidas acústicas em Hz, o *pitch* é usado como conceito perceptual, e diz respeito à sensação de altura (grave - agudo), de modo que quanto maior for a freqüência fundamental, maior será o *pitch* ou, equivalentemente, mais agudo será o sinal [78]. A relação entre a freqüência fundamental e a sensação de altura do sinal é quase logarítmica e portanto não-linear [29].

Duração

O parâmetro duração refere-se ao intervalo de tempo entre o início e o final de um segmento fonético. Na prática, os segmentos fonéticos possuem durações médias da

ordem de dezenas a centenas de milisegundos. O valor médio e a dispersão da duração são características individuais de cada falante.

A determinação da duração não é uma tarefa fácil, pois a natureza contínua da fala não permite que limites bem claros entre segmentos fonéticos sejam estabelecidos.

Energia

A energia é o parâmetro prosódico relacionado à amplitude do sinal de fala, sendo as variações de amplitude produzidas pelas variações de pressão do ar vindo dos pulmões. O correlato perceptual da energia chama-se intensidade. A intensidade é medida através do julgamento feito pelo ouvinte no volume do som (forte ou fraco).

Aparentemente, a amplitude é um parâmetro prosódico tão importante como os demais, porém, tem-se observado na prática que a intensidade do sinal tem uma função de contraste menos significativa do que os outros parâmetros prosódicos, como a duração e a freqüência fundamental. O que se observa efetivamente é que as variações de durações e sobretudo de freqüência fundamental determinam, mais do que o aumento de energia, a localização do acento nas sentenças [29]. Portanto, em estudos desenvolvidos sobre modelagem prosódica, como os referenciados no capítulo anterior, são abordados apenas os aspectos de duração e/ou a freqüência fundamental.

Acentuação

De modo geral, a acentuação é o modo de realizar um som ou um grupo de sons com mais destaque do que outros [75].

Geralmente a acentuação se manifesta de duas maneiras: no vocábulo (acento vocabular) ou na enunciação da frase (acento frásico). Assim, toda palavra da nossa língua possui uma sílaba tônica (com exceção de monossílabos átonos), e muitas vezes, na escrita, essa sílaba tônica é indicada por meio de um acento gráfico. Já o acento frásico é utilizado para salientar certas palavras, de forma a facilitar a compreensão do enunciado por parte do ouvinte.

A primeira impressão de um ouvinte é que as sílabas acentuadas são mais fortes do que as demais, ou seja, que o fenômeno de acentuação está diretamente relacionado ao parâmetro prosódico de energia (intensidade). Entretanto, o que se observa efetivamente é que as diferenças dos valores de duração e *pitch* são mais importantes para a determinação do acento do que a intensidade em si [29].

Assim, na determinação da prosódia também podem ser consideradas a classificação gramatical e a tonicidade de cada palavra como também as pausas entre as palavras ou frases. A classificação gramatical e as pausas são essenciais em nível de sentença e a tonicidade é fundamental para as considerações de prosódia em nível de palavra, já que informa qual sílaba é mais proeminente, ou seja, qual apresenta maior destaque.

Portanto, a prosódia pode ser tratada em nível de fonema (microprosódia), de sílaba, de palavra ou de frase, e o desenvolvimento de um modelo prosódico em um conversor texto-fala torna-se necessário quando se deseja obter uma fala com maior naturalidade possível.

2.5 Discussão

Os sons da fala são produzidos pelo aparelho fonador, que é constituído pelo sistema respiratório, pelo sistema fonatório e pelo sistema articulatório. Para produzir a fala, o locutor exerce uma série de controles sobre o aparelho fonador produzindo a configuração articulatória e a excitação apropriadas, gerando os diversos sons da fala (sons sonoros, sons surdos e sons plosivos). Assim, a seqüência de sons, dirigida por regras de linguagem, é transmitida do locutor para o ouvinte, no processo de comunicação entre pessoas através da fala.

O entendimento de determinadas características dos sons da fala e de conceitos de lingüística, tratados neste capítulo, são fundamentais para a determinação de um modelo de produção da fala, e, portanto, são usados principalmente na concepção de um dicionário de unidades acústicas e de um modelo prosódico, para um conversor texto-fala para a Língua Portuguesa, desenvolvidos neste trabalho.

Capítulo 3

Processamento do Texto

O processamento do texto constitui o primeiro módulo de um conversor texto-fala e tem a função de transformar o texto escrito na sua forma ortográfica em unidades fonológicas, as quais serão posteriormente associadas à prosódia (*pitch* e duração), para informar ao módulo de processamento do sinal como o texto deve ser pronunciado. Esse módulo pode ser subdividido em duas etapas: na primeira etapa é feita uma análise do texto, incluindo uma transcrição de siglas, números e abreviaturas, por extenso, como também uma análise gramatical, e na segunda etapa é realizada uma transcrição fonética do texto resultante, após o processo de análise.

Neste capítulo são descritos os estágios básicos que compõem o analisador de texto como também o módulo de transcrição fonética implementado em um sistema de conversão texto-fala para o Português Brasileiro, utilizado para testes do modelo prosódico desenvolvido neste trabalho.

3.1 Analisador de Texto

No analisador de texto é realizado inicialmente um pré-processamento, no qual siglas, números e abreviaturas são escritos por extenso. Posteriormente é realizada uma análise morfológica, uma análise semântica e uma análise sintático-prosódica, conforme mostrado na Figura 3.1.

Devido à importância destes subestágios na conversão texto-fala, é feita uma descrição mais detalhada nas seções seguintes.

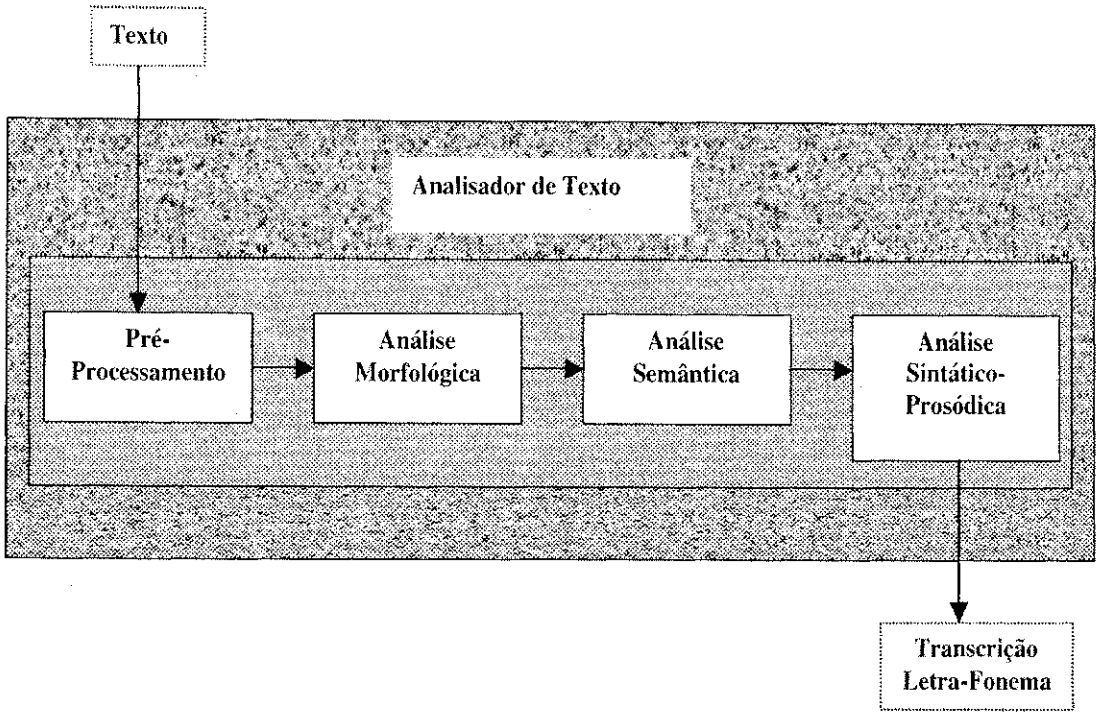

Figura 3.1: - Analisador de texto para um conversor texto-fala.

3.1.1 Pré-Processamento

No estágio de pré-processamento de texto deve ser identificado inicialmente o papel dos caracteres de pontuação, como, por exemplo, o caso do hífen, do ponto final, do ponto de interrogação, etc. Posteriormente deve ser realizada a expansão das abreviaturas, como também devem ser observadas as siglas, os caracteres especiais (como, por exemplo: '$', '%' e '&') e o formato correto dos números.

Abreviaturas

Abreviatura é a representação de uma palavra por meio de algumas de suas sílabas ou letras [1], como, por exemplo: Senhor $\Longrightarrow$ Sr.

A abreviação das palavras é realizada por uma simples questão de tempo e de espaço sobretudo na comunicação escrita, porém devem obedecer a regras já fixadas do nosso sistema ortográfico e não podem ser alteradas.

A maioria das abreviaturas são lidas de forma completa e freqüentemente incluem pausas. Nesse caso, deve-se ter um dicionário contendo as abreviaturas e as palavras correspondentes, para que seja efetuada a troca adequadamente, como, por exemplo:

(n)(º) $\Longrightarrow$ (número)

(e)(.)(g)(.) $\Longrightarrow$ (por)( )(exemplo)

Os nomes de unidades físicas, moeda de um país, ou caracteres especiais também podem ser incluídos no dicionário, apesar de alguns serem ambíguos (como *V* que pode ser *volt* ou *cinco* em algarismos romanos). No caso de ambigüidade, devem ser estabelecidas regras com base no contexto, o que pode tornar a conversão mais complexa e gastar maior tempo no processo de busca.

Siglas

Sigla é um tipo especial de abreviatura em que se reúnem as letras inciais (maiúsculas) de uma locução substantiva ou nome composto, como, nos exemplos: Produto Interno Bruto ⇒ PIB e cavalo-vapor ⇒ CV (ou HP).

A transcrição de uma sigla é realizada conforme a sua identificação. Em alguns casos, elas podem ser identificadas pela leitura individual das suas letras, ou seja, tem-se uma seqüência não articulada, como, por exemplo, em 'FGTS'. Em outros casos, elas podem ser identificadas pela leitura coordenada conforme os padrões silábicos da língua portuguesa, ou seja, tem-se uma seqüência articulada, como, por exemplo, em 'IBOPE', ou por pré-definição através de regras de pronúncias, como, por exemplo, em 'PASEP'. Assim, as siglas podem ser identificadas e marcadas no estágio de normalização de modo a evitar erros de deteção do final da sentença.

Números

Os números podem ser detectados inicialmente em cinco formatos:

- O primeiro englobando os casos que representam valores monetários, os quais podem ser identificados pelo R$ antes dos algarismos seqüenciados ou pela vírgula para identificar os centavos, como, por exemplo: R$ 1.200,00 (Um mil e duzentos reais).

- O segundo englobando os códigos postais, que podem ser identificados pela sigla CEP antes dos algarismos seqüenciados ou pelo hífen que separa os três últimos algarismos, como, por exemplo: CEP 58100-950.

- O terceiro englobando apenas as datas, as quais são identificadas por separação de barras ou por pontuação a cada dois algarismos, como, nos exemplos: 21/10/99

e 21.10.99.

- O quarto englobando apenas números de telefones, que podem ser identificados pelo número de algarismos ou pelo código local (do Estado) ou ainda pelo código da operadora, como, por exemplo: 0XX.83.3331000.

- O quinto englobando os números de forma geral e que não estão incluídos nos casos anteriores.

Após serem identificados por regras, os números são escritos por extenso.

Algoritmo para a Normalização

Um algoritmo para realizar a normalização do texto pode ser dividido em três etapas:

- Na primeira, é dado um tratamento às siglas.

- Na segunda, é determinado o formato correto dos algarismos;

- Na última, é dado um tratamento adequado às abreviaturas;

Neste caso, cada etapa deve apresentar um dicionário contendo informações essenciais para que o texto atinja um formato adequado.

Outras situações podem ser consideradas visando tornar o conversor texto-fala mais robusto: além dos dicionários, podem ser criadas regras de decisões com relação a determinadas siglas pronunciadas conforme a seqüência de letras - como é o caso de CPF -, e com outras pronunciadas conforme o modelo silábico - como é o caso INCRA. Também deve ser observado que as palavras abreviadas não vêm, necessariamente, finalizadas por um ponto (.) - como é o caso de "ha" (hectare). O ponto (.), por sua vez, também pode caracterizar o final de uma sentença e, assim, pode vir após uma palavra ou mesmo seguir uma sigla, como, por exemplo - "Este é o número do meu RG.".

Para possibilitar a ocorrência de todas as situações especificadas anteriormente, é necessário estabelecer algumas regras, como as descritas a seguir:

- Os vocábulos do texto podem ser classificados em quatro classes: formato adequado (escrito por extenso), sigla, algarismo e abreviação;

- Se o vocábulo for uma seqüência de letras maiúsculas, tem-se uma sigla;

- Se o vocábulo for uma seqüência de números, tem-se um algarismo;
- Se o último caracter for um ponto (.), deve-se verificar os caracteres anteriores:
  - Se os caracteres anteriores forem maiúsculos, tem-se uma sigla;
  - Se os caracteres anteriores fazem parte de um dicionário de abreviações;
  - Senão o vocábulo já se encontra no formato adequado.
- Se o vocábulo não se enquadrou em nenhuma consideração anterior, deve-se verificar se ele faz parte do dicionário de exceções de siglas (sem o sinal de pontuação);
- Se nenhuma das afirmações anteriormente citadas foi verificada, então o vocábulo já apresenta um formato adequado.

Com base nas considerações apresentadas, tem-se um fluxo de dados para a normalização do texto conforme mostrado na Figura 3.2.

Assim o texto é reescrito e encontra-se pronto para ser submetido diretamente ao estágio de transcrição fonética ou submetido a uma análise prévia nos estágios de análise morfológica, análise semântica e análise sintático-prosódica visando o aprimoramento da modelagem prosódica.

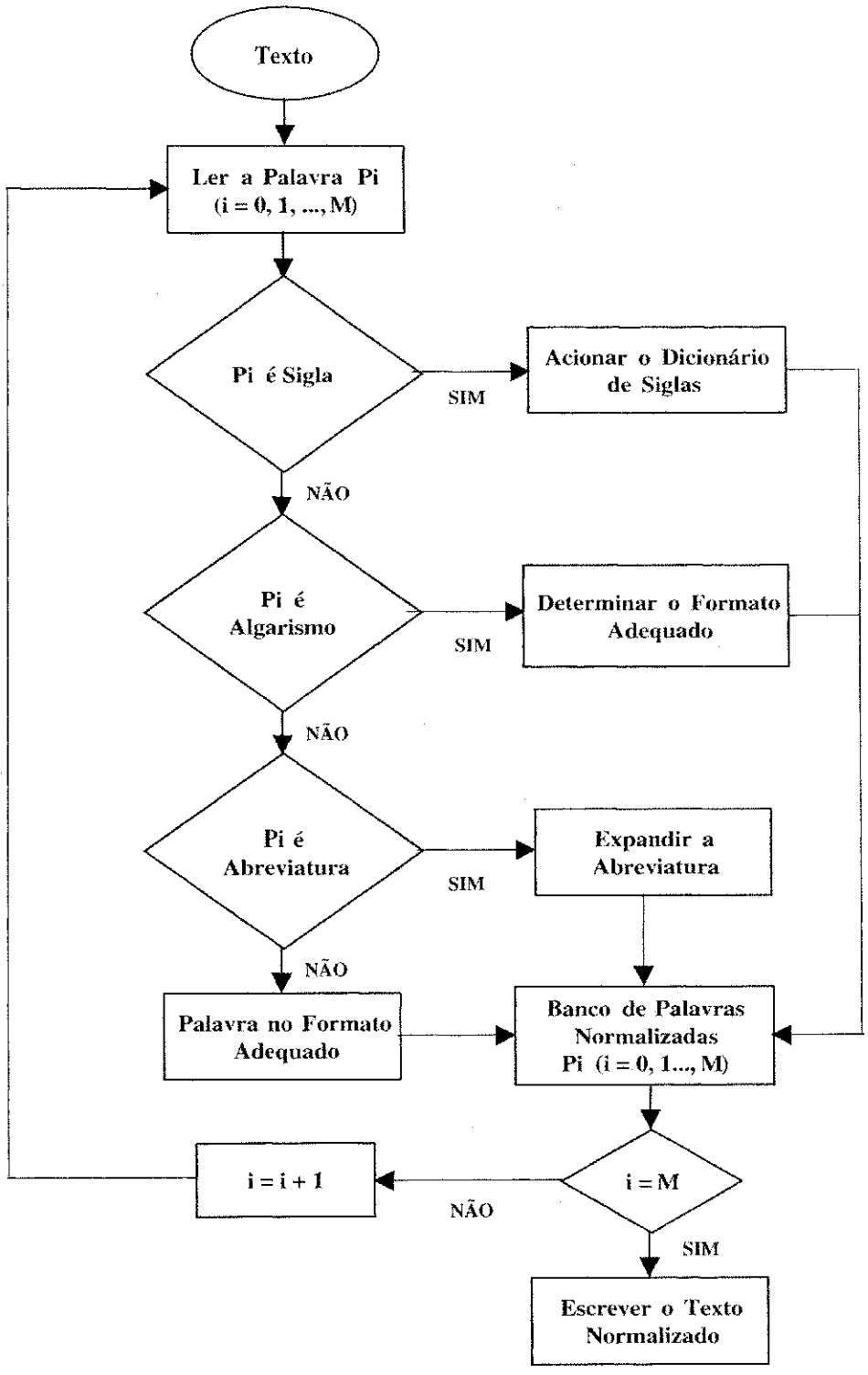

Figura 3.2: - Fluxograma correspondente ao algoritmo de normalização do texto.

3.1.2 Análise Morfológica

A morfologia pode ser definida como a parte da Lingüística que trata da descrição das regras que regem a estrutura interna das palavras em função de um conjunto mínimo de unidades com comportamento semântico (com significado) chamadas morfemas [10, 69].

Quanto à natureza de significação, os morfemas classificam-se em lexicais e gramaticais. Por exemplo na palavra 'ruas' tem-se: *rua* = morfema lexical e *s* = morfema gramatical.

Os morfemas gramaticais podem por sua vez ser divididos em [79]:

- Morfemas classificatórios, os quais são constituídos pelas vogais temáticas cuja função é enquadrar os vocábulos em classes de substantivos, adjetivos e verbos.

- Morfemas flexionais, os quais alteram os morfemas lexicais adaptando-os à expressão das categorias gramaticais que sua classe admite: nos substantivos e adjetivos com gênero e número e nos verbos com modo, tempo, número e pessoa. Assim, podem resultar no acréscimo ou supressão de um ou mais fonemas ao morfema lexical, como, nos exemplos: *"rapaz - rapazes"*, *"órfão - órfã"*, ou na alternância ou permuta de um fonema no interior do vocábulo, como, nos exemplos: *"povo -povos"*, *"formoso - formosa"*.

- Morfemas derivacionais, os quais criam novas palavras na língua a partir de determinado morfema lexical, como, por exemplo, a partir de *"livr-o"* pode-se obter *"livr-eiro"*, *"livr-aria"*, *"liv-rinho"*, etc.

- Morfemas relacionais, os quais ordenam os elementos da frase, possibilitando a concatenação dos morfemas lexicais entre si, como preposições, conjunções e pronomes relativos. Assim, a manipulação desses morfemas pertence à sintaxe.

Por sua vez, a análise morfológica consiste na descrição da estrutura da palavra, depreendendo suas formas mínimas ou morfemas, de acordo com uma significação e uma função elementares que lhes são atribuídas dentro da significação e da função total da palavra [79].

A análise morfológica pode ser feita usando-se um dicionário de morfemas lexicais, o qual reduz significativamente o número de dados armazenados para consultas, e realizando-se a classificação das palavras a partir dos morfemas lexicais e gramaticais através de um conjunto de regras. Assim, todas as palavras são sistematicamente

examinadas uma a uma, através de uma série de hipóteses de sufixos e dependendo da aprovação são definidas as classes. Nesse caso, são realizados os testes anexando-se alguns sufixos a cada morfema lexical da palavra que está sendo analisada, até a sua aprovação.

Uma análise morfológica apresenta um certo grau de importância em um conversor texto-fala, pois a partir dessa podem ser determinadas as classes das palavras e, assim, serem identificadas as suas pronúncias de forma correta.

3.1.3 Análise Semântica

A análise semântica (ou contextual) tem a função de eliminar a ambigüidade de palavras que podem constar em mais de uma classe como as homógrafas heterofônicas e que não conseguiram ser devidamente classificadas através da análise morfológica. Nessa análise, duas divisões podem ser evidenciadas: a primeira enquadra os vocábulos que têm seus sons diferenciados a partir de sua classe gramatical - por exemplo, a palavra acordo (verbo/substantivo). A segunda, por sua vez, considera as palavras que apresentam a mesma classe gramatical, porém possuem sons diferentes, como é o caso do vocábulo sede (substantivo/substantivo). Assim, deve-se proceder uma análise mais apurada no contexto para possibilitar a correta identificação do som das unidades.

Existem basicamente dois modelos de análise semântica: o modelo probabilístico, que é baseado na transição de probabilidades entre sucessivas classes de palavras, realizado na prática através de Modelos de Markov Escondidos (HMMs) ou através de Redes Neurais [10]. e o modelo determinístico, no qual um conjunto de regras sim/não são exploradas para identificar se aceitam ou rejeitam determinadas classes de palavras. O resultado de cada decisão é aplicado a um dicionário contendo as palavras homógrafas heterofônicas, de modo que na saída desse estágio se obtenha a classificação correta de cada palavra aplicada à sua entrada.

Assim, na análise semântica é necessário definir informações contextuais, para eliminar a ambigüidade entre palavras, com base no contexto em que se inserem. Se o conjunto for relativamente reduzido tem-se uma análise limitada, e pode ser usado o modelo determinístico, considerando-se a sua simplicidade em relação modelo probabilístico [10]. Caso contrário, pode ser usado o modelo probabilístico, como forma de eliminar o excesso de regras, e contemplar o maior número possível de desambigüidades de palavras.

3.1.4 Análise Sintático-Prosódica

Normalmente nem todas as seqüências de palavras contidas em um dicionário de determinada língua resultam em uma frase correta. Embora a lista de frases que se possa construir em determinada língua seja quase infinita, é necessário que exista determinada sintaxe para o seu entendimento.

A sintaxe pode ser definida como a parte da gramática que descreve as regras segundo as quais as palavras se combinam para formar frases [80]. Portanto, trata da relação lógica das palavras nas frases.

A análise sintática, por sua vez, implica na decomposição de uma frase em seus elementos constituintes a fim de verificar a relação lógica existente entre eles. Ela pode ser realizada através de uma estrutura em árvore, e a informação resultante tem uma grande vantagem sobre simples descrições das palavras, pois apresenta os destaques possíveis da(s) estrutura(s) interna(s) das frases [81]. Esses destaques são importantes na síntese da fala por duas razões:

- A precisão da transcrição fonética depende do conhecimento da classe das palavras e do conhecimento da relação de dependência entre as sucessivas palavras.
- A prosódia depende fortemente da sintaxe. Claro que ela tem também dependência com a semântica e a pragmática, mas se palavras estão sendo avaliadas sobre aspectos de dependência, os conversores texto-fala se concentram basicamente na sintaxe.

Observa-se neste processo que existem ocorrências nas quais uma palavra, dependendo da sua localização, faz parte de determinada classe gramatical. A palavra '*a*', por exemplo, pode funcionar como um artigo ou uma preposição, e a palavra '*entre*' pode ser uma preposição ou um verbo.

Observa-se também que, em uma língua, a colocação das palavras obedece a tendências variadas, quer de ordem estritamente gramatical, quer de ordem psicológica e estilística. Entretanto, o maior responsável pela ordem favorita, em uma língua, tende a ser a entonação oracional.

Assim, o estágio de análise sintático-prosódica no contexto de um sistema de conversão texto-fala realiza um processo que provê informações necessárias para a pronúncia das palavras de forma adequada, ou seja, a entonação da frase e a variação na proeminência que os falantes podem usar. Informações obtidas a partir desse estágio

são usadas para fazer decisões de limites de sentença e de tonicidade e, portanto, auxiliar à modelagem prosódica.

3.2 Transcrição Fonética

É sabido que a pronúncia das palavras geralmente difere de sua grafia. Esse fenômeno se origina em parte devido ao descompasso natural entre a língua falada e a forma mais rígida conservadora da escrita, ou seja, a escrita permanece fixa por muitos anos enquanto a pronúncia vai mudando com o tempo. Como conseqüência não existe uma correspondência única entre as letras (caracteres) e os fonemas (sons elementares e distintivos que o homem produz quando exprime seus pensamentos e emoções) [75]. Como exemplo, pode ser citada a letra "*g*" que pode assumir os fonemas */j/* e */g/* como nas palavras: "*gelo*" e "*gato*", respectivamente.

Para a obtenção da fala mais natural possível na síntese concatenativa, devem ser levadas em conta considerações fonéticas como as apresentadas no parágrafo anterior. A inteligibilidade, por sua vez, está relacionada a um conjunto mínimo e consistente de alofones (variantes de enunciação que cada fonema pode apresentar) [82] e sua determinação em alguns casos, como, por exemplo, o /e/ e o /o/ aberto ou fechado, pode tornar-se extremamente complexo, pois requer uma análise contextual.

Além disso, deve ser considerado que, na transcrição fonética ou letra-fonema, as palavras resultantes precisam adquirir um formato adequado para serem buscadas em um dicionário de unidades acústicas.

Em geral, a transcrição fonética pode ser feita com base no Alfabeto Fonético Internacional (AFI), que utiliza letras tomadas dos alfabetos grego e latino, para representar os sons da fala de qualquer língua [1, 69]. Observa-se, porém, que a simbologia do citado alfabeto é complexa e que alguns dos símbolos não fazem parte do código ASCII (*American Standard Code Information Interchange*) para uma conversão de leitura do texto em leitura da máquina. Assim, objetivando tornar o processo de transcrição mais simples para a Língua Portuguesa, pode ser adotada uma relação entre letra e fonema (ou mais precisamente entre grafia e fonema), como mostrado na Tabela 3.1 [56]. Nesta Tabela, pode-se perceber que algumas letras têm uma única representação fonética como "*a*", "*b*" e "*d*" que correspondem aos fonemas /a/, /b/ e /d/, como, por exemplo, nas palavras "*abril*", "*banco*", e "*dado*". Outras letras têm mais de uma representação fonética, com um destaque para a letra "*x*" que pode assumir os fonemas

/x/, /s/, /z/ e */ks/* como nas palavras: *"lixo"*, *"máximo"*, *"exame"* e *"fixo"*.

Tabela 3.1: Relação entre grafia e fonema para a Língua Portuguesa

GRAFIA	FONEMA	GRAFIA	FONEMA	GRAFIA	FONEMA
a	/a/	gü	/g/	qu	/k/ - /ku/
ã	/am/	h	-	r	/r/ - /rr/
b	/b/	i	/i/	rr	/r/
c	/k/ - /s/	j	/j/	s	/s/- /z/
ç	/s/	l	/l /- /u/	ss	/s/
ch	/x/	lh	/l/	t	/t/
d	/d/	m	/m/	u	/u/
e	/e/ - /i/	n	/n/ - /m/	v	/v/
f	/f/	nh	/n/	x	/x/- /s/ -/ z/ - /ks/
g	/j/ - /g/	o	/o/ - /u/	z	/z/- /s/ - /is/
gu	/g/	p	/p/	-	-

3.2.1 Letras que Representam mais de um Fonema

Quando uma letra (caractere) tem várias representações fonéticas, torna-se necessário realizar uma análise contextual, para determinar de forma correta o fonema considerado. Assim, com base na Tabela 3.1, são destacados os seguintes casos:

Letra c

A letra *c* representa dois fonemas, ou seja, /k/ e /s/. A determinação da forma correta pode ser representada a partir das seguintes considerações:

- quando *c* anteceder as vogais *a*, *o*, *u* ou quando anteceder uma consoante, representa o fonema /k/, como nas palavras: "casa", "copo", "culinária" e "tecla"; respectivamente;
- quando *c* anteceder as vogais *e* e *i*, representa o fonema /s/ - nessa regra têm-se as palavras "ceia" e "cinto", por exemplo.

Letra g

A letra *g*, por sua vez, representa dois fonemas distintos: /g/ e /j/. Portanto, a sua transcrição fonética deve obedecer às seguintes regras:

- se a letra *g* vier seguida das letras *e* ou *i*, corresponde ao fonema /j/ - como nas palavras "geladeira" e "girafa";

- se a letra *g* vier seguida da letra *u* e das letras *e* ou *i*, as letras *g* e *u* se unem, correspondendo ao fonema /g/ - é o que ocorre nas palavras "negue" e "lânguido", respectivamente;

- para as demais ocorrências, a letra *g* representa o fonema /g/ - como exemplo têm-se as palavras "gato" e "gosto".

Letra l

A letra *l* corresponde a dois fonemas: /l/ e /u/. Para determinar as ocorrências possíveis tem-se:

- se a letra *l* vier seguida de consoante diferente de *h*, corresponde ao fonema /u/, como, por exemplo, na palavra "falta".

- se a letra *l* vier seguida de uma vogal, então pronuncia-se como o fonema /l/, como é o caso da palavra "lata";

- se a letra *l* coincidir com o final da palavra, ela representa o fonema /u/, como, por exemplo, na palavra "fatal".

Letra q

- Normalmente a letra *q* representa o fonema /k/.

- Quando a letra *q* vier seguida da letra *u* e das letras *e* ou *i*, as letras *q* e *u* se unem, formando dígrafo, e representam o fonema /k/, como exemplo, tem-se as palavras: "queijo" e "quinta".

Letra s

Com relação à letra *s*, dois fonemas podem ser representados por ela, ou seja:

- se a letra *s* vier entre vogais, tem-se o fonema /z/, como, por exemplo, na palavra "casa";

- caso contrário, a letra *s* assume o fonema /s/, como, por exemplo, nas palavras "semáforo" e "solo".

Letra x

A letra x corresponde a quatro fonemas distintos: /x/, /z/, /s/ e /ks/, de modo que devem ser criadas regras conforme as várias ocorrências existentes. Assim, foi realizado um estudo considerando os vocábulos presentes no dicionário Aurélio [1]. Para um conjunto de cerca de quatrocentas palavras, contendo a letra x, foram observados os seguintes casos [56]:

- palavras em que ocorrem a seqüência (f ou s) + (vogal) + (x) + (vogal), troca-se o x por /ks/, como nas palavras: "sexo", "saxão", "saxofone", "sexagésimo", "sexênio", "sexual", "fixa", "crucifixo", "asfixia";

- palavras terminadas em x, troca-se o x por /ks/, como por exemplo: "fax", "tórax";

- palavras em que ocorrem a seqüência: (vogal) + (xc), troca-se o (xc) por /s/, como, por exemplo: "excesso" e "exceção";

- palavras iniciadas pela seqüência: (e) + (x) + (vogal), troca-se o (x) por /z/, como, por exemplo, em: "exato", "exagero", "exercício", "existir", "exuberância" e "exótico";

- palavras iniciadas por (a ou o) + (x) + (vogal), troca-se o (x) por /ks/, como, por exemplo, em: "axioma", "axial" e "oxigênio";

- palavras em que ocorre a seguinte seqüência: (vogal) + (x) + (consoante diferente de c), troca-se o (x) por /s/, como, por exemplo, em "extrair", "texto", "extenso", "explicar", "explosão".

- palavras em que se encontra a seqüência (axi) antecedida por (t) ou (l), ou sucedido por (l) ou (m), troca-se o (x) por /ks/, como, por exemplo, em "táxi", "galáxia", "maxilar" e "máxima";

- palavras em que se encontra a seqüência (exa) antecedida por (h) ou (fl) ou (pl), troca-se o (x) por /ks/, como, por exemplo, em "hexaedro", "reflexão" e "complexa";

- palavras em que se encontra a seqüência (exi) antecedida pelas letras (l), (fl) ou (pl), troca-se o (x) por /ks/, como, por exemplo, em: "léxico", "flexível", "complexidade";

- palavras em que se encontra a seqüência (*exo*) antecedida pelas letras (*n*) ou (*fl*), troca-se o (*x*) por /ks/, como, por exemplo, em: "anexo", "reflexo";

- palavras em que se encontra a seqüência (*oxe*) fora do início da palavra, troca-se o (*x*) por /ks/, como, por exemplo, em "boxe";

- palavras em que se encontra a seqüência (*oxi*) antecedida pelas letras (*pr*), troca-se o (*x*) por /s/, como, por exemplo, em: "próximo" e "aproximação";

- palavras em que se encontra a seqüência (*oxi*) não antecedida pelas letras (*pr*), troca-se o (*x*) por /ks/, como, por exemplo, em "tóxico";

- palavras em que se encontra a seqüência (*uxo*) antecedida pelas letras (*fl*) ou no início da palavra, troca-se o (*x*) por /ks/, como, por exemplo, em "uxoricida" e "refluxo".

3.2.2 Algoritmo para a Transcrição Fonética

Um algoritmo para realizar a transcrição fonética pode ser composto pelos casos apresentados na subseção anterior, com letras correspondendo a mais de um fonema, e por um conjunto de regras, também baseado em uma análise contextual, conforme descrito a seguir:

1. Se for encontrado um (*c*) seguido pela consoante (*h*), troca-se o (*ch*) por /x/. Exemplo: "cacho", "chapada".

2. Se for encontrado um (ç), troca-se esse caractere por /s/. Exemplo: "canção", "caçula", "paçoca".

3. Se for encontrado um (*h*) iniciando uma palavra, ele é excluído, pois não tem fonema correspondente. Exemplo: "humano", "hora", "Hélio".

4. Todo (*e*) sucedido por (*l*) no final da palavra, deverá ser trocado por /é/, como, por exemplo, em "móvel", "pincel", "pastel".

5. Se for encontrado um (*n*), deve-se verificar o caractere seguinte: - Se não for vogal nem (*h*), troca-se o (*n*) por /m/. Exemplo: "cantar", "contabilidade", "quente".

6. Se for encontrado um (qu), deve-se verificar o caractere seguinte:
   - Se for (e) ou (i), troca-se o (qu) por /k/. Exemplo: "queijo", "quilo", "quimera", "quente".
   - Se for (a), troca-se o (qu) por /ku/. Exemplo: "qual", "quanto", "quaisquer".
   - Se for encontrado $(q\ddot{u})$, troca-se por /ku/. Exemplo: "tranqüilo","seqüência".

7. Se for encontrado um (r) no início de uma palavra, troca-se por /rr/. Exemplo: "rei", "roda", "rua", "rural".

8. Se for encontrado um (s) entre duas vogais, troca-se o (s) por /z/. Exemplo: "asa", "casado", "asilo", "osmose". Se for encontrado (ss), troca-se por /s/. Exemplo: "assado", "assassino", "associação".

9. Se for encontrado um (y), troca-se por /i/.

10. Todo (a) será aberto /a/ se a letra seguinte não for (m) ou (n). Exemplo: "batata", "faca", "macaco". Se aparecer um $(\acute{a})$, troca-se por /a/, como, por exemplo, em "pá" e "tábua". Se aparecer um $(\tilde{a})$, troca-se por /am/; um $(\tilde{a}e)$ por /ame/; um $(\tilde{a}o)$ por /amo/; como, por exemplo, em "mão", "anã", "mãe", "lã".

11. Se for encontrado $(\acute{i})$, troca-se por /i/. Se for encontrado $(\acute{u})$, troca-se por /u/, como, por exemplo, em "viúva", "íngreme", "ímã".

12. Se for encontrado um (e) no final da palavra precedido por consoante, esse deverá ser trocado por /i/, como, por exemplo, em "cante", "contente", "frente".

13. Se for encontrado $(\tilde{o}e)$, troca-se por /ome/, como, por exemplo, em "põe" e "compõe".

14. Se houver um (o) no final da palavra precedido por consoante, esse deverá ser trocado por /u/, como, por exemplo, em "gato", "macaco", "canto".

15. Se for encontrado (qu) seguido por (e) ou (i), troca-se por /g/, como, por exemplo, em "guerra", "guitarra". Se for encontrado $(q\ddot{u})$ seguido por (e) ou (i), troca-se por /gu/, como, por exemplo, em "lingüiça", "agüentar".

16. Se a palavra terminar em (es), troca-se o (es) por /is/, como, por exemplo, em "fones", "meses", "testes".

17. Se a palavra terminar em (*os*), troca-se (*os*) por /us/, como, por exemplo, em "campos", "meninos".

18. Se for encontrado um (*z*) no final de uma palavra, deve-se verificar o caractere anterior: - Se for (*i*) ou (*u*), troca-se o (*z*) por /s/, como, por exemplo, em "feliz".

19. Se for encontrado acento circunflexo em uma palavra (exemplo: cômodo), esse deve ser eliminado no processo de transcrição.

3.3 Discussão

O processamento do texto em um conversor texto-fala é realizado por uma série de estágios, tais como: pré-processamento, no qual, siglas, números, abreviaturas e caracteres especiais são escritos por extenso; classificação gramatical das palavras, incluindo análise morfológica, análise semântica e análise sintático-prosódica e, finalmente, transcrição fonética do texto. Observa-se que cada estágio apresenta uma série de dificuldades, como, por exemplo: no pré-processamento o ponto pode causar ambigüidade, indicando um final de frase ou uma abreviatura a ser expandida; na análise morfológica é necessário realizar uma descrição da estrutura das palavras com base nas suas formas mínimas ou morfemas e, na análise semântica, deve ser avaliado o contexto anterior e contexto posterior, a cada palavra analisada, para a eliminação da ambigüidade e, conseqüentemente, uma correta classificação gramatical. Além disso, é necessária a geração de bases de dados e regras para a execução de cada estágio. Assim, neste trabalho, optou-se por criar um conversor texto-fala mais simples, para efetuar testes no modelo de prosódia desenvolvido, excluindo a classificação gramatical e contemplando os estágios de pré-processamento e a transcrição fonética, a qual é implementada com base nas regras descritas na seção anterior.

Capítulo 4

Técnicas de Modelagem Prosódica

A incorporação da prosódia em um conversor texto-fala é de extrema importância para a obtenção de uma fala sintetizada com inteligibilidade e naturalidade. Essa incorporação é realizada definindo-se inicialmente um modelo prosódico, que determina a evolução temporal dos parâmetros prosódicos, de modo que seja possível identificar, na fala, a acentuação, o ritmo e a entonação [8]. A definição de um modelo é feita com base em uma análise acústica de um *corpus* constituído por textos lidos por um locutor e em conhecimentos fonético-fonológicos referentes à língua cuja modalidade falada se pretende sintetizar [2].

Na prática, o processamento prosódico em um conversor texto-fala, utilizando-se síntese concatenativa, é realizado a partir de dados obtidos na saída do estágio de transcrição fonética. Para tal, são gerados e rotulados os segmentos fonéticos correspondentes às unidades acústicas de um dicionário, previamente estabelecido. Em função do tipo de segmento fonético e da técnica de modelagem prosódica é efetuada uma busca de unidades no dicionário e realizada a síntese do sinal. Os parâmetros prosódicos das unidades podem sofrer ou não alterações no estágio de síntese, dependendo das técnicas de modelagem prosódica e de síntese utilizadas [9, 11, 30, 32].

Assim, neste capítulo, são representadas as principais técnicas de modelagem prosódica, com base na literatura consultada. A apresentação tem como objetivos enfocar o *estado-da-arte* relativo ao assunto e estabelecer um comparativo entre as técnicas para a definição de um novo modelo prosódico para um conversor texto-fala, utilizando síntese concatenativa, para o Português Brasileiro. As técnicas de síntese serão abordadas no capítulo seguinte.

4.1 Modelos de Duração

Os modelos de duração segmental aplicados à síntese da fala são procedimentos pelos quais as durações das unidades acústicas (fones, difones, etc.) possam ser determinadas e ajustadas adequadamente [29]. A eficiência de cada modelo aumenta com a proximidade entre as durações determinadas no modelo, para a síntese de um enunciado, e as respectivas durações das unidades acústicas do mesmo enunciado dito por um locutor.

Existem várias técnicas utilizadas na elaboração de um modelo de duração, dentre as quais se destacam: técnicas baseadas em regras [4, 65, 83, 84], técnicas baseadas em redes neurais [11, 40, 65, 85], técnicas baseadas em árvores de classificação e regressão [27, 42, 43, 44, 86] e técnicas baseadas em HMMs [45, 46]. Uma síntese destas técnicas é apresentada nas subseções seguintes.

4.1.1 Modelo de Duração de Klatt

O modelo de predição da duração de Klatt (1987) para o Inglês serviu e serve de referência para modelos desenvolvidos por outros pesquisadores [4, 83, 84]. Todos eles consideram o segmento como paradigma para a obtenção da duração segmental, a qual é obtida após a aplicação sucessiva de um certo número de regras [65]. Um tipo de regra pode ser, por exemplo: *"reduzir a duração de um segmento pertencente a uma sílaba postônica por um fator de 0,9"*, ou *"aumentar a duração de uma vogal em uma sílaba tônica por um fator de 1,2"*.

Os princípios fundamentais deste modelo são:

1. A cada segmento associa-se uma duração, denominada *duração intrínseca* (inerente à natureza do segmento), que corresponde ao valor médio da distribuição de valores que a duração daquele segmento pode assumir;

2. Cada regra tenta prever uma variação percentual com o objetivo de efetuar um aumento ou diminuição na duração do segmento;

3. Os segmentos não podem ser reduzidos a valores menores do que certa duração mínima.

Assim, a duração segmental pode ser expressa por um modelo aditivo-multiplicativo dado pela Equação (4.1).

$$D_{UR} = \frac{(D_{IN} - D_{MIN}) \; . \; PRS}{100} + D_{MIN} \tag{4.1}$$

onde:

D_{UR} é a duração final do segmento após o ajuste;

D_{IN} é a duração intrínseca do segmento;

D_{MIN} é a duração mínima que o segmento pode assumir, calculada em função de D_{IN}. Para cada segmento não acentuado tem-se uma duração mínima dada pela Equação (4.2):

$$D_{MIN} = 0,45 \; . \; D_{IN} \tag{4.2}$$

PRS corresponde à percentagem de redução do segmento, determinada de maneira cíclica e cumulativa pela aplicação das regras (uma regra introduz em geral um fator multiplicativo que é replicado ao valor atual de PRS, fornecido por uma regra anterior). Os fatores que condicionam o valor final de PRS são o contexto fonético imediato e o ambiente sintático-prosódico do segmento.

Este modelo foi usado inicialmente no sintetizador de formantes de Klatt [87], e tem sido adaptado em vários modelos desenvolvidos por outros pesquisadores. Como exemplos podem ser citados o modelo de duração para o Português Brasileiro desenvolvido por Silva em [4], o modelo de duração para o Chinês desenvolvido por Shih em [83], e, mais recentemente, o modelo de duração para o Alemão desenvolvido por Hoffmann em [84], dentre outros. Evidentemente que a qualidade de cada modelo depende do número de regras utilizadas, que por sua vez depende do idioma considerado.

4.1.2 Modelo de Duração de Campbell

O modelo de Campbell (1992) para predição da duração segmental, para o Inglês britânico, é realizado separando-se o controle do tempo em nível de sílaba (procurando assim descrever a estruturação rítmica da fala) do cálculo da duração do segmento (em nível inferior), cálculo este que é efetuado a partir do paradigma temporal fornecido pela sílaba. Assim esse modelo opera em duas etapas [65]:

- Na primeira etapa a duração silábica é obtida por aprendizado automático pelo uso de uma rede neural do tipo *perceptron* multicamadas [88];

Nessa técnica, a rede neural é treinada para aprender a associar uma descrição fonológica da sílaba e de seu contexto frasal (no domínio simbólico) à duração real dessa mesma sílaba (no domínio físico). Isso é realizado por meio de uma regra de aprendizagem que, pela modificação dos pesos das conexões, busca aproximar a saída desejada (dos pares entrada/saída apresentados) à saída atual da rede (obtida aplicando-se os processos calculatórios definidos para os neurônios). Assim, a rede realiza uma passagem complexa entre o código simbólico e uma realização.

- Na segunda etapa a duração silábica é distribuída entre os segmentos que formam a sílaba pelo uso de um modelo estatístico chamado modelo de repartição [65].

O modelo de repartição estabelece que todos os fonemas de uma determinada sílaba possuem um único fator de alongamento z (denominado *z-score*) o qual impõe que a duração dessa sílaba seja dada pela Equação (4.3):

$$\text{Duração (sílaba)} = \sum_{i=1}^{n} \exp(\mu_i + z \ . \ \sigma_i) \tag{4.3}$$

onde a duração de cada segmento i é obtida pelas parcelas $\exp(\mu_i + z \ . \ \sigma_i)$.

O par estatístico (μ_i, σ_i) representa a média e o desvio-padrão associados à distribuição formada pelas durações das realizações do fonema /i/. Esta distribuição é obtida pela análise de um *corpus* pré-estabelecido de frases lidas.

Observa-se também que na equação (4.3) é utilizado um único valor de z para todos os segmentos que compõem a sílaba, ou seja, todos os segmentos que compõem a sílaba estão sujeitos ao mesmo alongamento (ou compressão).

Na literatura consultada foram encontrados modelos desse tipo, para a Língua Francesa [85], para o Português Brasileiro [65], para a Língua Chinesa [11] e, mais recentemente, para a Língua Alemã [40], dentre outros. Nesses modelos tem-se procurado aperfeiçoar a técnica tradicional existente como também ampliar a capacidade de processamento das informações prosódicas relativas à duração dos segmentos fonéticos, para a produção de uma fala sintetizada mais natural possível.

4.1.3 Modelo de Duração de Hunt e Black

O modelo de duração de Hunt e Black (1997) utiliza árvores de classificação e regressão, para a predição das durações dos fonemas, a partir de regras (questões) relativas ao contexo em que se inserem [47, 86, 89]. Uma árvore é constituída basicamente por

nós, ramos e folhas [90]. Em cada nó, há uma questão referente a uma determinada característica de interesse e, dependendo da resposta sim ou não, o fluxo da informação é dirigido a um nó seguinte através de um ramo. O processo continua até que seja atingida uma folha, contendo o valor de predição em questão. Assim, através de uma árvore é possível predizer valores específicos dentro de determinada classe (árvore de classificação), ou valores contínuos através de uma média e desvio padrão (árvore de regressão) [86].

Um exemplo de árvore de classificação, que inclui a predição dos valores de duração dos segmentos fonéticos /a/ e /xa/, resultantes da palavra *acha*, é apresentado na Figura 4.1. Cada nó, representado por um pequeno círculo, contém uma questão relativa a determinada característica fonética. Após a decisão sim ou não o ramo correspondente (representado por uma seta) conduz o fluxo de informação ao próximo nó, e o processo se repete até a classificação atingir uma folha (retângulos: rotular a sílaba /a/ e rotular a sílaba /xa/) ou uma subclassificação seguinte. Em cada folha os segmentos fonéticos são rotulados com o nome dos fonemas correspondentes e com um valor de duração arbitrado em função do contexto em que se inserem. Caso os segmentos fonéticos não sejam /a/ e /xa/, o fluxo da árvore segue outros caminhos de classificação, indicados pelos retângulos que não tratam do rotulamento.

Assim, as seguintes questões são formuladas nos nós da árvore da Figura 4.1:

1. O número de caracteres do segmento fonético é 1? Se for vá para o nó 2, se não, vá para o nó 4.

2. O segmento fonético é palavra? Se for classifique qual o tipo de palavra e a folha correspondente, se não, vá para o nó 3.

3. O segmento fonético é a vogal /a/? Se for rotule o segmento /a/ com o nome do fonema e o valor de duração correspondente, se não, classifique qual o outro tipo de vogal e a folha correspondente.

4. O número de caracteres do segmento fonético é 2? Se for vá para o nó 5, se não, teste se o segmento tem três ou mais caracteres e o classifique juntamente com a folha correspondente.

5. O segmento fonético é palavra? Se for classifique a palavra e a folha correspondente, se não, vá para o nó 6.

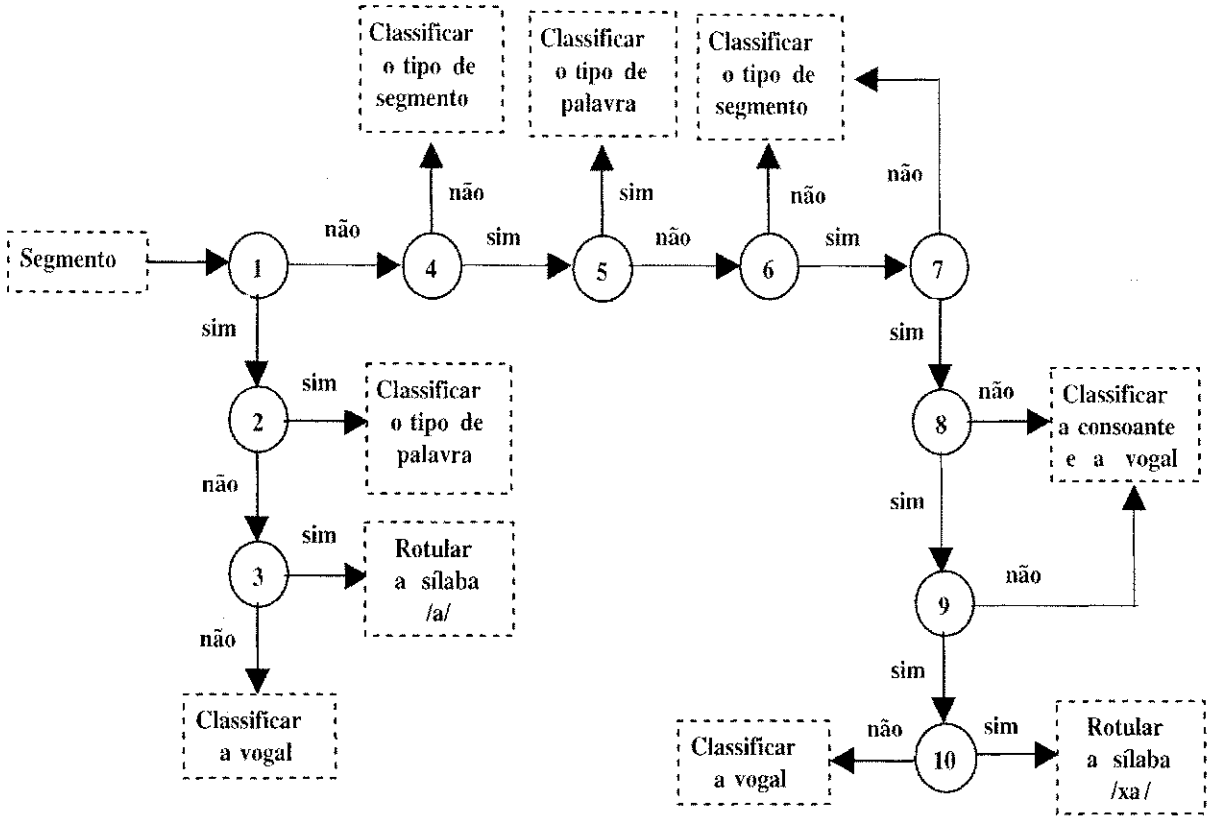

Figura 4.1: - Árvore de classificação para a predição da duração dos segmentos /a/ e /xa/ na palavra *acha* (adaptado de Dusterhoff [47]).

6. O segmento fonético corresponde a uma sílaba átona? Se for vá para o nó 7, se não, classifique o outro tipo de segmento e a folha correspondente.

7. O segmento fonético é CV (consoante-vogal)? Se for vá para o nó 8, se não, classifique o outro tipo de segmento e a folha correspondente.

8. A consoante é fricativa surda? Se for vá para o nó 9, se não, classifique qual o outro tipo de consoante, a vogal que a sucede e a folha correspondente.

9. A consoante do segmento CV é /x/? Se for vá para o nó 10, se não, classifique qual o outro tipo de consoante, o tipo de vogal que a sucede e folha correspondente.

10. A vogal do segmento CV é /a/? Se for rotule o segmento /xa/ com o nome dos fonemas e o valor de duração correspondente, se não, classifique qual o outro tipo de vogal e a folha correspondente ao segmento /x/ + vogal.

Na literatura consultada foram encontrados modelos de duração usando essa técnica para a Língua Inglesa [86], para a Língua Francesa [27], e mais recentemente para a Língua Chinesa [42, 43, 44], dentre outros. Nesses modelos tem-se procurado incorporar o maior número possível de características com relação à duração dos fonemas, no sentido que a fala sintetizada seja a mais natural possível.

4.1.4 Modelos de Duração Utilizando HMMs

Modelos de Markov (MM) são representações utilizadas para modelar um sinal através de uma seqüência de observações [91]. Em uma Cadeia de Markov supõe-se uma fonte gerando tais saídas observáveis, denominada Fonte de Markov. Os símbolos gerados a partir dessa fonte são dependentes apenas de observações anteriores, as quais são geradas da mesma forma e assim sucessivamente. O número de seqüências anteriores consideradas para gerar uma saída é conhecido como ordem da Cadeia de Markov. Na maioria das aplicações conhecidas cadeias de primeira e segunda ordem são suficientes, mesmo porque a complexidade computacional cresce exponencialmente a partir dessas ordens [92].

Assim, um canal de Markov básico tem um número finito de estados, e um conjunto de funções aleatórias, com cada função aleatória associada a cada um dos estados. Para instantes de tempo discretos, assume-se que o processo está em algum estado e uma seqüência de observações é gerada por uma função aleatória correspondendo ao estado corrente [71]. O canal seleciona o estado de acordo com uma matriz de probabilidade de transição associada. O observador vê somente a saída da função aleatória associada a cada estado e não pode observar diretamente os estados do canal de Markov básico; resultando, então, no termo Modelo de Markov Escondido (HMM) [92].

No domínio da fala, HMMs tem sido de grande interesse devido ao seu baixo custo computacional e por basear-se em modelos estocásticos do sinal da fala, sendo capazes de modelar vários eventos, tais como fonemas, sílabas, etc. [71].

Em geral existem dois tipos de estrutura: HMMs sem restrições entre os estados da cadeia (ergódicos) e HMMs seqüenciais denominados de *left-right* (esquerda-direita) [71]. Os modelos 'esquerda-direita' apresentam melhor desempenho do que os ergódicos no caso de reconhecimento de locutor [71], e tem sido usados em trabalhos de reconhecimento de fala como os desenvolvidos por por Alcaim e Santos em [93], e Ynoguti e Violaro em [94], dentre outros. Um exemplo de HMM utilizando a estrutura 'esquerda-

direita', que pode ser usado para a predição de duração de um segmento é apresentado na Figura 4.2. Neste modelo, a_{ij} é a probabilidade de efetuar uma transição do estado i para o estado j e $b_i(O(t))$ é a probabilidade de emitir o símbolo $O(t)$ no estado i, no instante t. Um estudo detalhado sobre os parâmetros que caracterizam o HMM 'esquerda-direita' pode ser encontrado em [71].

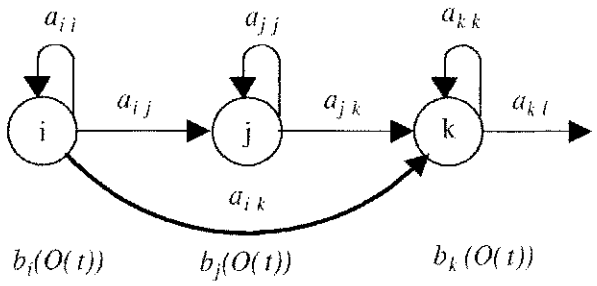

Figura 4.2: - Modelo de Markov que pode ser usado para a determinação da duração de um segmento [94]).

Aplicações mais recentes dessa técnica, em modelos de duração, são encontradas em um trabalho realizado para a Língua Espanhola em [45] e para a Língua Inglesa em [46], dentre outros. Nesses trabalhos tem-se procurado aperfeiçoar a técnica tradicional no sentido de tornar o sistema desenvolvido mais robusto e produzindo uma fala sintetizada mais natural possível.

4.2 Modelos Acústicos de Entonação

Modelar a entonação com que uma pessoa produz a fala é um problema complexo, principalmente devido à necessidade de se determinar uma curva de *pitch* com base apenas em um texto escrito. Na língua falada, a entonação pode variar para um mesmo texto dependendo do falante. Como não é possível extrair de imediato a entonação de qualquer texto escrito, normalmente atribui-se às sentenças sintetizadas um contorno entonacional neutro (por exemplo, uma linha de declinação simples). Assim, o contorno entonacional pode ser obtido a partir de algoritmos de extração do *pitch* [71, 95, 96, 97].

Vários tipos de modelos de entonação foram desenvolvidos em nível lingüístico [98, 99, 100], perceptual [47, 89] e acústico [41, 43, 49]. Nesta seção é apresentada uma síntese dos modelos de entonação em nível acústico, considerando que o modelo

prosódico, proposto neste trabalho, é realizado nesse nível.

4.2.1 Modelo de Entonação de Fujisaki

O modelo de Fujisaki é baseado no fato de que as curvas de entonação, embora contínuas no tempo e em freqüência, se originam de eventos discretos produzidos pelo leitor, os quais surgem como contínuos devido a mecanismos fisiológicos relacionados ao controle da freqüência fundamental [101].

Neste modelo são definidos dois tipos de eventos discretos, denominados comandos de frase (modelados como impulsos, atuando na entonação da frase) e comandos de acento (modelados como pulsos retangulares, destacando a sílaba tônica da palavra), conforme mostrado no exemplo da Figura 4.3. Observa-se nesta figura que cada comando é aplicado, de forma independente, à entrada de um mecanismo de controle (filtro linear de 2ª ordem criticamente amortecido) [102]. As saídas dos dois filtros são somadas para produzir a curva com os valores de F_0. Os efeitos produzidos na saída do sistema pelos comandos de frase são representados pela curva tracejada e a soma dos efeitos dos comandos de frase e acento são representados pela curva continua. A função de transferência dos dois filtros e os dois tipos de comandos são determinados a partir de fundamentos fisiológicos [102].

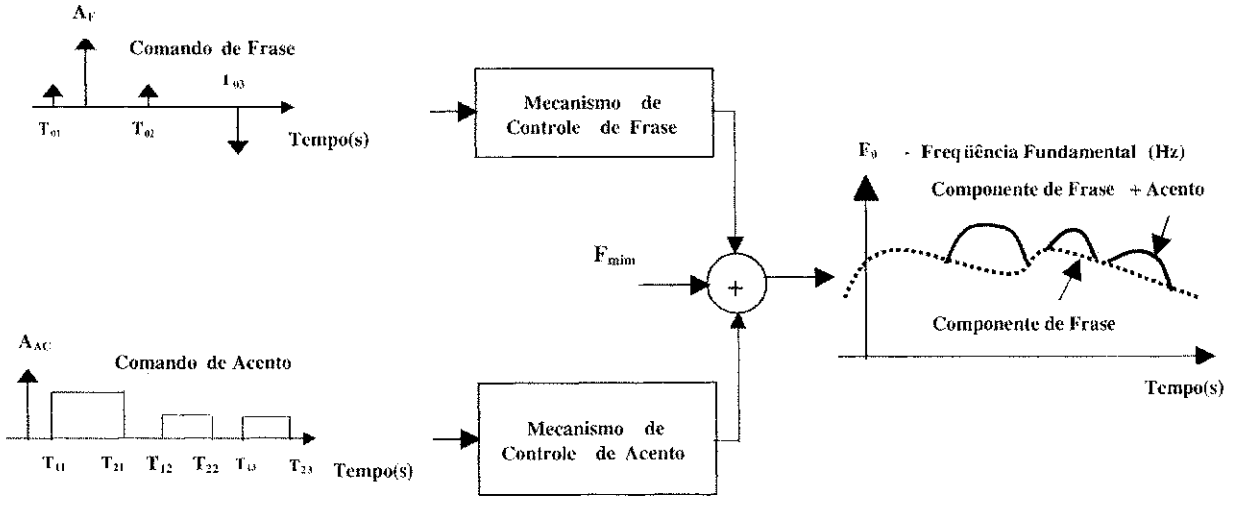

Figura 4.3: - Esquema do modelo de Fujisaki incluindo um exemplo de um contorno de freqüência fundamental obtido da superposição de três comandos de frase e três comandos de acento [101].

Considerando que os parâmetros dos filtros são fixos para um dado locutor (desde que eles sejam relacionados a restrições fisiológicas), os parâmetros mais importantes para o modelo são [10, 102]:

- Um valor mínimo de freqüência ($\mathbf{F}_{min}$) no qual todos os valores dos comandos de acentos e frases são sobrepostos;
- O número de comandos de frases e dois parâmetros de cada comando: amplitude de cada impulso ($\mathbf{A}_F$) e cronometragem do tempo de ocorrência ($\mathbf{T}_{01}$, $\mathbf{T}_{02}$, $\mathbf{T}_{03}$, etc.);
- O número de comandos de acento e três parâmetros de cada comando: a amplitude ($\mathbf{A}_{AC}$ de cada pulso retangular), a cronometragem do tempo de inicialização ($\mathbf{T}_{11}$, $\mathbf{T}_{12}$, $\mathbf{T}_{13}$, etc.) e a cronometragem do tempo de finalização ($\mathbf{T}_{21}$, $\mathbf{T}_{22}$, $\mathbf{T}_{23}$, etc).

Estes parâmetros (incluindo os parâmetros dos filtros) são obtidos estimando-se inicialmente o número de comandos de frase e de acento a partir de uma análise visual do contorno de F_0, e posteriormente minimizando-se o erro quadrático médio entre esse contorno e o do modelo através de uma busca no espaço paramétrico. Existem vários métodos de análise do *pitch* e determinação destes parâmetros, dentre os quais se destaca um método desenvolvido por Mixdorff's para o alemão em [103] e para o chinês em [104], que deduz automaticamente o número de comandos de frase e de acento.

Assim, o modelo de Fujisaki tem sido aplicado em várias línguas (incluindo Alemão, Inglês e Basco) nas quais foram obtidas aproximações satisfatórias segundo os autores [101, 102, 105]. Para tal, foi utilizado um número relativamente elevado de comandos de acento e de frase, com um acréscimo de restrições sobre os modelos desenvolvidos.

4.2.2 Modelo de Entonação com Estilização Acústica

No modelo de Estilização Acústica são destacadas as invariantes acústicas que as curvas de freqüência fundamental podem conter, através do cálculo das linhas de declinação e/ou aproximando-se as curvas por uma seqüência de pontos rotulados [10, 106].

Nas línguas entonacionais[1], como a Língua Portuguesa, diferentes tipos de enunci-

[1] Línguas entonacionais são aquelas em que a sílaba tônica é fundamental para o entendimento do significado de cada palavra. Elas podem ser classificadas em dois tipos: línguas de acento livre, como, por exemplo a Língua Inglesa e línguas de acento fixo, como, por exemplo, a Língua Portuguesa [10].

ados carreiam padrões melódicos predeterminados pelo aparelho fonador. Nesse caso, as 'frases declarativas' se destinguem das 'frases interrogativas', porque as primeiras apresentam um padrão entonacional descendente e as segundas, um padrão ascendente [107]. Assim, nas frases declarativas as curvas de F_0 estão em torno de valores médios e decrescem com o tempo, de modo que a freqüência fundamental tende a ser maior do que o seu valor médio na primeira parte da curva e menor na segunda parte. Essa tendência é chamada *declinação*. As linhas de declinação podem ser determinadas matematicamente como a melhor aproximação linear que se ajusta aos extremos locais (picos e vales) da curva de F_0 (uma linha reta passando pelos picos, - linha superior -, e outra linha reta passando pelos vales - linha inferior), conforme mostrado no exemplo da Figura 4.4. Através desta figura é possível visualizar um valor médio de F_0 representado por uma linha tracejada, os valores de F_0 correspondentes aos picos F_1, F_2, F_3, F_4, e F_5 e os valores de F_0 correspondentes aos vales V_1, V_2, V_3 e V_4, ao longo da curva, que tem início em IF e termino em FF.

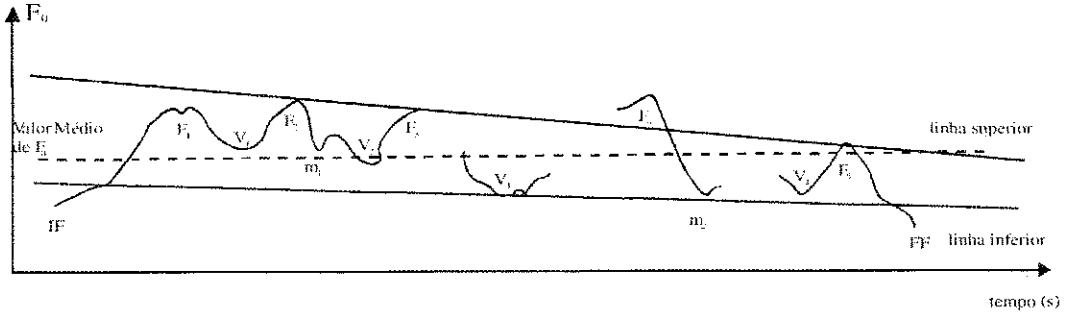

Figura 4.4: - Linhas de declinação obtidas de uma análise acústica [10].

Na prática, a tarefa de determinar as linhas de declinação apresenta uma certa dificuldade, pois os locutores tendem a eliminar seus registros vocais após a produção de um baixo valor de F_0 (tipicamente após as pausas separando os maiores constituintes das frases), os quais induzem o cálculo da linha inferior [10]. A linha superior nem sempre é determinada de forma precisa, pois F_0 máximo depende da acentuação da palavra e da frase.

Por outro lado, as curvas de F_0 podem ser expressas como seqüências de pontos rotulados, de modo que as transições entre eles podem ser completadas por um processo de interpolação, como por exemplo, o modelo desenvolvido por Taylor em [108]. Nesse caso, os parâmetros da função de interpolação em questão são impostos ou ajustados

através de regressão linear [86], aumentando a complexidade do sistema.

O modelo de estilização acústica também foi usado por Delmote [109], para desenvolvimento de um sistema de aprendizagem de língua estrangeira e, mais recentemente, em um trabalho desenvolvido por Shih [110], para confirmação de palavras na Língua Inglesa.

4.2.3 Modelo de Entonação de Silva

O modelo de entonação de Silva (1995) para o Português Brasileiro é baseado em regras e em níveis de hierarquia, para a geração das curvas de entonação de um enunciado, conforme ilustrado na Figura 4.5. Cada nível hierárquico deve obedecer às determinações do nível superior e, por sua vez, gerar determinações para o nível inferior, pois a curva de F_0 pode ser vista como uma sobreposição de efeitos [4, 87]. Assim, para cada enunciado que se queira sintetizar deve ser derivada uma estrutura desse tipo, na qual tem-se: 1º nível: Frase; 2º nível: Constituinte Prosódico; 3º nível: Palavra; 4º nível: Sílaba; 5º nível: Fone.

Os constituintes prosódicos nesse modelo são definidos como um grupo de palavras adjacentes na sentença, possuindo cada grupo a propriedade de influenciar a evolução dos parâmetros prosódicos ao longo das palavras que o constituem, como, por exemplo, na frase: 'as crianças de rua são o principal problema brasileiro' pode ser dividida nos constituintes: 'as crianças de rua' e 'são o principal problema brasileiro'.

Além disso devem ser observados os padrões de entonação em nível de frase. No caso da língua portuguesa, em que os padrões de entonação de frases declarativas são descendentes [107], devem ser estabelecidas linhas de declinação (superior e inferior), de forma semelhante ao modelo de estilização acústica [106], para cada constituinte prosódico, e posteriormente para palavras, sílabas e fones, entre as quais ocorrem as variações de F_0. Assim, são determinados os valores de F_0 inicial e final de cada constituinte prosódico dentro de uma senteça, de cada palavra dentro de cada constituinte prosódico, de cada sílaba dentro da palavra e de cada fone dentro da sílaba, para os devidos ajustes de F_0 no sinal da fala, no estágio de síntese.

No modelo desenvolvido por Silva em [4], foram consideradas apenas as curvas de F_0 de constituintes prosódicos, palavras e sílabas, baseadas em um *corpus* de 164 frases declarativas. Para cada fone foi realizada uma interpolação linear entre os valores de F_0 inicial e final, de modo que a curva de F_0 para cada enunciado deve ser contínua

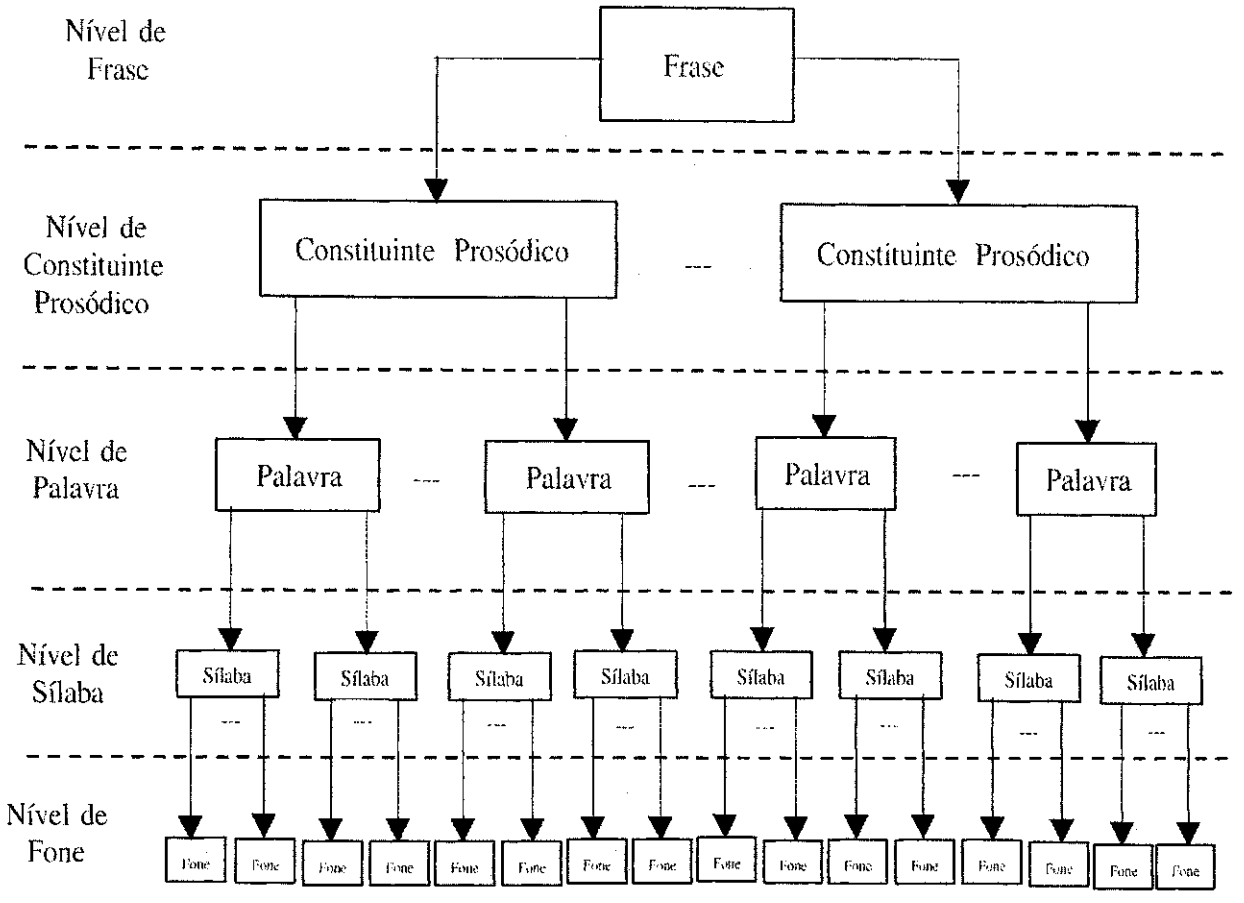

Figura 4.5: - Árvore correspondente ao modelo de entonação usando níveis de hierarquia [4].

e composta por uma sucessão de segmentos de reta, onde cada segmento corresponde a uma sílaba [4, 111]. Uma aplicação mais recente desse modelo é encontrada em um trabalho desenvolvido por Sheng em [33], para a Língua Chinesa.

4.2.4 Modelo de Entonação de Campbell

O modelo de entonação de Campbell utiliza uma seleção de unidades para determinar os contornos de entonação. Nesse modelo, o sinal de fala sintetizado é obtido pela concatenação de formas de onda de unidades acústicas rotuladas e selecionadas em um dicionário produzido por um só locutor. O dicionário contém unidades com diferentes características prosódicas e espectrais e, assim, é possível escolher as de interesse, de modo que o sistema seja capaz de produzir a fala soando de forma mais natural possível

[112, 113, 114, 115].

Nessa técnica é desenvolvido um dicionário de modelos de entonação com base em um *corpus* de fala. Cada modelo corresponde ao grupo de entonação.

No processo de geração da curva de F_0, os modelos são selecionados no dicionário, de modo a se obter um *custo total* mínimo possível na seleção. O *custo total* é composto do *custo objetivo*, $C^O(O_i, U_i)$, e do *custo concatenação*, $C^C(U_{i-1}, U_i)$, conforme mostrado na Figura 4.6.

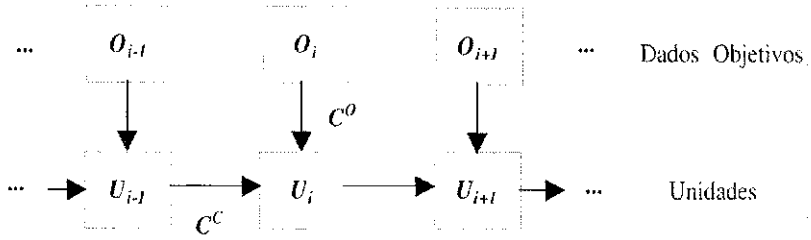

Figura 4.6: - Representação dos custos para a seleção de unidades [115].

O *custo objetivo* é uma estimativa das diferenças das características prosódicas (duração, freqüência fundamental, etc.) da unidade U_i do dicionário, e do objetivo O_i, o qual supostamente a representa. Esse custo é calculado através da Equação 4.4 [115, 116].

$$C^O(O_i, U_i) = \sum_{j=1}^{p} w_j^O C_j^O(O_i, U_i) \tag{4.4}$$

onde: $C_j^O(O_i, U_i)(j = 1, ..., p)$ são os p sub-custos correspondentes as diferenças dos elementos dos vetores da unidade U_i e do objetivo O_i, e w_j^O são os pesos dos p sub-custos.

O *custo concatenação* é uma estimativa da qualidade de uma concatenação entre unidades consecutivas (U_{i-1} e U_i), e pode ser dado pela Equação 4.5 [115, 116].

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

onde: $C_j^C(U_{1-i}, U_i)(j = 1, ..., q)$ são os q sub-custos correspondentes as diferenças dos elementos dos vetores da unidade U_{i-1} e da unidade U_i, e w_j^C são os pesos dos q sub-custos.

Portanto, no processo de geração da prosódia de uma sentença a curva de entonação é determinada inicialmente através da busca, no dicionário, de modelos que mais se

aproximam dos modelos requeridos para cada grupo de entonação. Após ser encontrada a seqüência ótima, a curva completa de F_0 é construída com base nos modelos obtidos.

Após esta etapa, tem-se uma descrição acústica da prosódia de uma sentença, ou seja, uma seqüência de fonemas rotulados com duração e F_0 a qual deve ser aplicada à entrada de um sintetizador para a geração do sinal da fala correspondente.

O modelo de Campbell foi otimizado por Hunt e Black em [115], e posteriormente utilizado por Malfrère em [27], no desenvolvimento de um modelo de entonação para a Língua Francesa. Também foi utilizado em aplicação mais recentemente por Erdem [40], no desenvolvimento de um modelo de entonação para a Língua Alemã.

4.3 Discussão sobre os Modelos Prosódicos

A obtenção da prosódia em um conversor texto-fala implica na determinação de parâmetros prosódicos de unidades acústicas e na definição e implementação de um modelo de duração e/ou de entonação, com base nos segmentos fonéticos considerados no conversor.

Observa-se a existência de vários modelos de duração e/ou entonação, aplicados a conversores que utilizam técnicas diferentes, de modo que se torna difícil um estudo comparativo entre eles, principalmente em termos práticos.

Os resultados obtidos com modelo de duração de Klatt dependem do conjunto de regras que o constitui. As regras são dependentes da língua com a qual se está trabalhando e podem atuar em nível de fonema, sílaba, palavra ou frase. Quanto mais completo for o conjunto de regras melhor serão os resultados [29].

Por outro lado, os modelos de Campbell, Malfrère e baseados em HMMs, para a predição de duração, produzem bons resultados segundo os autores, mas têm a desvantagem de serem modelos mais complexos do que o de Klatt e de, eventualmente, produzirem erros bastante grosseiros principalmente quando forem encontrados contextos fonético-prosódicos mais raros, não contemplados no *corpus* da base de dados [29, 39, 44, 46]. Esse fato decorre da dificuldade de elaborar um *corpus* que constitua um espaço amostral completo dos fenômenos prosódicos que ocorrem em uma língua.

O modelo de entonação de Fujisaki tem apresentado uma precisão considerável na determinação da curva de F_0 para algumas línguas, como, por exemplo, para a Inglesa, para a Alemã e para a Chinesa [101, 102, 104]. A determinação dos parâmetros desse modelo é uma tarefa árdua, considerando que os componentes relativos aos comandos

de frase e de acento são sobrepostos e não podem ser calculados de forma direta. Nesse sentido, foram desenvolvidos métodos para a extração automática desses parâmetros, com resultados promissores segundo os autores [103, 117].

O modelo de estilização acústica, para a determinação dos contornos da curva de F_0, tem sido usado em sistemas de síntese da fala como, por exemplo, o desenvolvido Taylor *et al.* em [108]. Nesse sistema, os contornos são obtidos de forma automática de um *corpus* de fala natural e têm apresentado resultados mais significativos do que os sistemas baseados em regras, segundo o autor, no que se refere à qualidade de fala sintetizada.

O modelo de entonação de Silva [4], baseado em níveis de hierarquia, foi desenvolvido para o Português Brasileiro e mais recentemente por Sheng [33], para a Língua Inglesa. O modelo para o Português Brasileiro foi inserido em um sistema de conversão texto-fala utilizando a técnica de síntese PSOLA [118] e, apesar de ser um trabalho inicial, apresentou resultados satisfatórios, para um *corpus* de 164 frases, segundo o autor. Com relação ao modelo de Sheng para a Língua Chinesa os resultados também foram satisfatórios para o *corpus* de frases considerado no trabalho. Porém, verifica-se a necessidade de ampliar o *corpus* com frases contendo vários tipos de entonação, como também a incorporação de uma análise sintática no processamento lingüístico, para que o sistema possa produzir o maior número possível de frases sintetizadas com vários tipos de entonação, segundo o autor.

O modelo de entonação de Campbell apresenta um certo grau de complexidade, sobretudo na parte de classificação e seleção das unidades acústicas do dicionário, porém produz uma fala sintetizada de qualidade, segundo o autor, e pode ser adaptado para outras línguas, conforme as características de cada uma [27].

4.4 Modelo Prosódico Proposto

Em função das técnicas de modelagem prosódica apresentadas neste capítulo e do grau de complexidade que surge na implementação de cada uma delas, é proposto neste trabalho um modelo para a geração de prosódia em conversores texto-fala para o Português Brasileiro. O modelo é baseado em regras aplicadas a técnica de seleção de unidades para determinar os contornos de entonação. Uma característica importante é a redução do número de regras para a seleção de unidades, comparado a sistemas que utilizam unidades menores, como, por exemplo, difones [27], como também a simplicidade na

determinação dos parâmetros do modelo comparando-se com os modelos de Fujisaki [101], de Estilização Acústica [10] e de Silva [4]. Além da técnica de seleção de unidades, o modelo é baseado na tonicidade de palavras, utilizando sílabas e demissílabas como unidades acústicas do dicionário, que incorporam informações prosódicas das mais relevantes, bem como características articulatórias correspondentes aos fenômenos de coarticulação, para a obtenção de uma fala sintetizada com naturalidade. Para tal, é realizada uma análise acústica em um *corpus* de palavras, contendo as mais diversas combinações de fonemas para as sílabas, e de frases foneticamente balanceadas, para identificação do comportamento da duração e sobretudo do *pitch*, ao longo de palavras oxítonas, paroxítonas e proparoxítonas com até cinco sílabas.

Além disso, a aplicação deste modelo em conversores texto-fala, utilizando-se síntese concatenativa, resulta em uma simplificação nas etapas de processamento lingüístico e de sinal (necessita de concatenação simples), como mostrado na Figura 4.7. Neste caso, o tempo de processamento das informações para a obtenção da síntese da fala é relativamente menor do que os sistemas de conversão texto-fala que incorporam análise morfológica, sintática e semântica no processamento do texto, como também técnicas de síntese mais complexas, conforme será apresentado no capítulo seguinte.

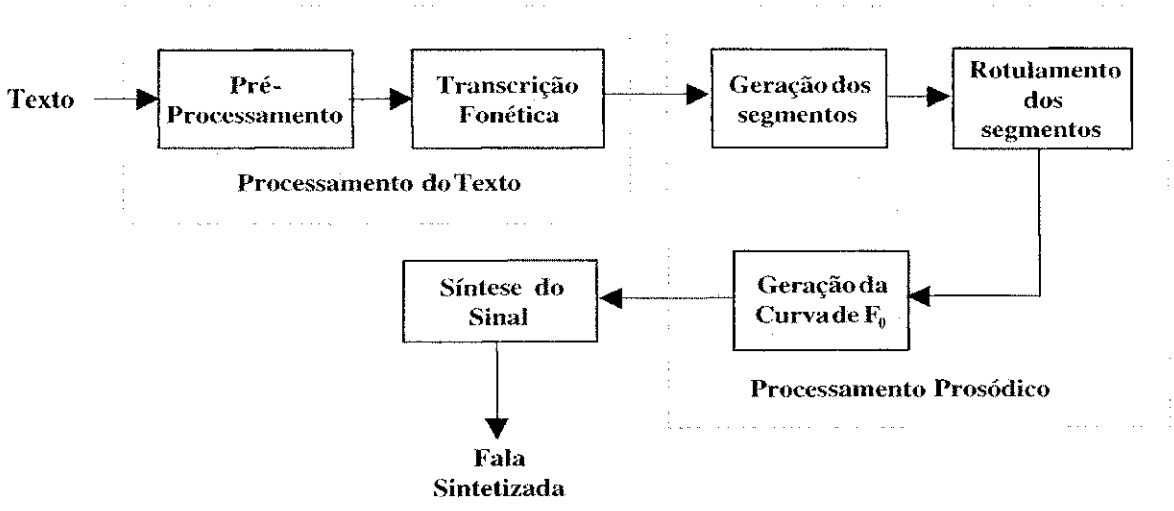

Figura 4.7: - Sistema proposto para a geração automática da prosódia baseado na tonicidade das palavras.

Portanto, após o processamento do texto, geração e rotulamento dos segmentos fonético-prosódicos (com o nome e valor de *pitch* correspondente), é realizada uma seleção de unidades acústicas, no dicionário, para a síntese do sinal. Esta seleção é

realizada através de uma estrutura de pesos atribuídos às sílabas tônicas, pretônicas e postônicas, considerando a curva de entonação das palavras analisadas no *corpus*, definido neste trabalho. Os valores de duração são incorporados a cada unidade acústica correspondente, em função da tonicidade, porém não são utilizados na classificação. Apesar deste modelo ser baseado em tonicidade de palavras, tem-se conseguido bons resultados na prosódia quando as palavras são inseridas em frases. Detalhes de implementação e resultados obtidos serão apresentados no Capítulo 7.

Capítulo 5

Técnicas de Síntese do Sinal de Fala

A síntese do sinal de fala, na conversão texto-fala, é realizada a partir das informações fonético-prosódicas, resultantes dos estágios de processamento lingüístico e processamento prosódico, as quais controlam a evolução temporal dos parâmetros: freqüência fundamental, duração e energia, durante o processo de síntese.

Atualmente existem várias técnicas de síntese, que podem ser classificadas em três grupos [87]:

- Síntese articulatória, na qual é modelado o mecanismo de produção da fala de forma direta;
- Síntese por formantes, na qual é modelada a função de transferência do trato vocal com base na teoria acústica de produção da fala;
- Síntese concatenativa, na qual são utilizadas unidades acústicas da fala natural de comprimentos variados e pré-gravadas, para a produção da fala sintetizada de forma contínua.

Neste capítulo são apresentadas as principais técnicas de síntese que podem ser implementadas em um conversor texto-fala, incluindo os algoritmos mais utilizados na síntese por concatenação de unidades acústicas. Também é apresentada a análise/resíntese LPC (*Linear Predictive Coding*), que tem sido bastante usada na síntese da fala, e a síntese híbrida desenvolvida no intuito de produzir uma fala natural e inteligível. É feito um estudo comparativo entre as técnicas e, em função das suas vantagens e desvantagens, é escolhida a síntese concantenativa para implementação em um conversor texto-fala, no qual está inserido o modelo prosódico proposto neste trabalho.

5.1 Síntese Articulatória

A síntese articulatória é baseada em uma modelagem mais próxima do aparelho fonador humano, de modo a simular sobretudo a dinâmica dos diversos articuladores no processo de produção da fala [29]. Assim, as dimensões e posições dos diversos articuladores (língua, mandíbula, lábios, véu palatino, etc.) como também a abertura glotal, tensão nas cordas vocais e a pressão dos pulmões estão associadas aos parâmetros do modelo. A área de abertura dos lábios, a constrição formada pela lâmina da língua, a abertura para as cavidades nasais, a área glotal média e a taxa de expansão ou contração do volume na região do trato vocal correspondente à faringe são exemplos de parâmetros de controle articulatório [119].

Normalmente os parâmetros de um modelo articulatório são obtidos pela observação da fala natural através de raios-X. Entretanto, as imagens de raios-X encontram-se em duas dimensões enquanto o trato vocal real encontra-se em três dimensões, de modo que se torna difícil a otimização desse modelo devido à ausência de determinados dados relativos aos movimentos dos articuladores. Também os graus de liberdade dos articuladores não são caracterizados através dos dados contidos nos raios-X e os movimentos da língua são complicados, tornando-se quase impossível fazer uma modelagem de forma precisa [87].

Em geral, a produção dos sons da fala na síntese articulatória pode ser modelada em três etapas:

- Na primeira etapa é realizada uma modelagem da vibração das cordas vocais [120].
- Na segunda etapa é modelado o formato do trato vocal, através da *função área*. Essa função é definida como a área instantânea da seção reta do trato vocal, da glote aos lábios, determinada pelo posicionamento dos articuladores [121].
- Na terceira etapa é realizada uma modelagem do movimento dos lábios. Esta etapa é essencial em aplicações incluindo a síntese visual, pois amplia a capacidade de compreensão da fala [122].

Portanto, a implementação de um sintetizador articulatório é uma tarefa bastante complexa, tendo em vista que é difícil de simular-se todos os possíveis movimentos dos articuladores. Uma aplicação dessa técnica é encontrada em um trabalho recente

desenvolvido por Huang et al. em [123]. Segundo os autores, foi obtida uma fala inteligível.

5.2 Síntese por Formantes

A síntese por formantes é baseada nas freqüências e larguras de banda dos formantes como também no modelo fonte-filtro [124], de modo que o processo físico de produção da fala é descrito matematicamente por meio de três componentes básicos: fonte de excitação, característica de filtragem do trato vocal e característica de radiação para o meio externo, conforme mostrado no diagrama de blocos da Figura 5.1 [8]. Assim, um sintetizador utilizando esse método recebe como entrada um conjunto de parâmetros de controle atualizado periodicamente e produz como saída amostras do sinal da fala. Os parâmetros incluem amplitudes dos sinais, freqüência fundamental do sinal sonoro e freqüências e larguras de banda dos formantes, dentre outros. A obtenção dos parâmetros é realizada por meio de um conjunto de regras, denominadas *regras de síntese*, que atuam com base na transcrição fonética do texto aplicado na entrada do conversor texto-fala [2].

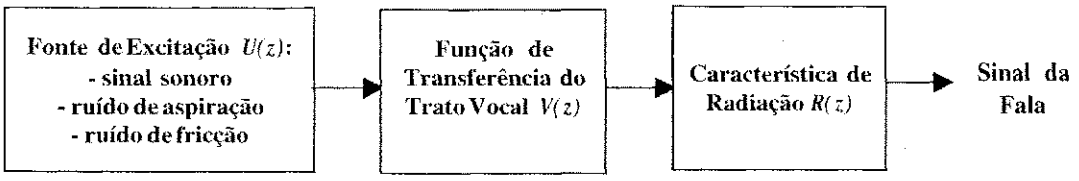

Figura 5.1: - Modelo simplificado de produção da fala na síntese por formantes [8].

No modelo da Figura 5.1 a fonte de excitação produz o sinal sonoro a partir de um gerador de impulsos separados por um intervalo de tempo igual ao período de *pitch* e produz os ruídos de aspiração e fricção (correspondentes à parte não sonora) a partir de um gerador de ruídos. Estes sinais podem ser misturados por meio de um modulador, de modo a produzir sons mistos como as consoantes fricativas sonoras [2].

No estágio seguinte, tem-se a função de transferência do trato vocal *V(z)* (principal componente da técnica) que é implementada utilizando-se uma associação de seções de segunda ordem. Cada seção é conhecida como ressonador, pois tem como objetivo modelar a freqüência e largura de banda de cada formante, e tem a seguinte função de transferência [8]:

$$R_n(z) = \frac{a_{1n}}{1 - a_{2n}z^{-1} - a_{3n}z^{-2}} \tag{5.1}$$

Os coeficientes a_{1n}, a_{2n} e a_{3n} estão relacionados com a freqüência central de ressonância, f_n, e a largura de banda do formante B_n, e são dados por:

$$a_{3n} = -e^{2\pi(B_n)T} \tag{5.2}$$

$$a_{2n} = 2e^{-2\pi(B_n)T}\cos(2\pi f_n T) \tag{5.3}$$

$$a_{1n} = 1 - a_{3n} - a_{2n} \tag{5.4}$$

onde:

f_n é a freqüência central de ressonância em Hz;

B_n é a largura de banda do ressonador em Hz;

T é o período de amostragem em segundos.

Na prática são necessários no mínimo três formantes para produzir uma fala inteligível e, no mínimo, cinco para produzir uma fala de alta qualidade [124]. Também os ressonadores podem ser associados em cascata ou em paralelo. Na associação em cascata, a saída de cada ressonador é aplicada à entrada do seguinte, conforme o exemplo com três formantes (f_1, f_2 e f_3), mostrado na Figura 5.2.

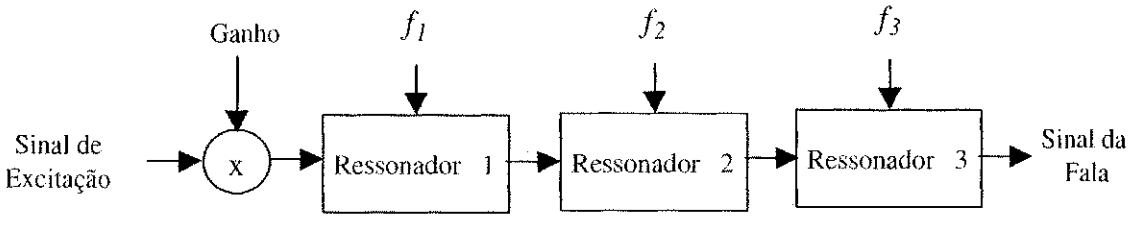

Figura 5.2: - Estrutura básica do sintetizador por formantes em cascata [87].

A vantagem de uma estrutura em cascata, como mostrado na Figura 5.2, é que as amplitudes relativas aos formantes das vogais não precisam de controles individuais e o controle da informação é feito apenas pelas freqüências formantes. Assim, ela tem apresentado uma qualidade melhor na produção de sons sonoros não nasalizados e como necessita de menos controles do que uma estrutura em paralelo torna-se mais simples

de implementar. A desvantagem é que a função de transferência dessa estrutura não pode ser modelada adequadamente para a produção dos sons fricativos e plosivos [87].

Por outro lado, em uma estrutura em paralelo, conforme o exemplo mostrado no diagrama de blocos da Figura 5.3, o sinal de excitação é aplicado simultaneamente nas entradas de todos os ressonadores e as saídas dos ressonadores são somadas. Nesse caso tem-se um controle individual do ganho e da largura de banda de cada formante e, portanto, é necessário mais controle da informação.

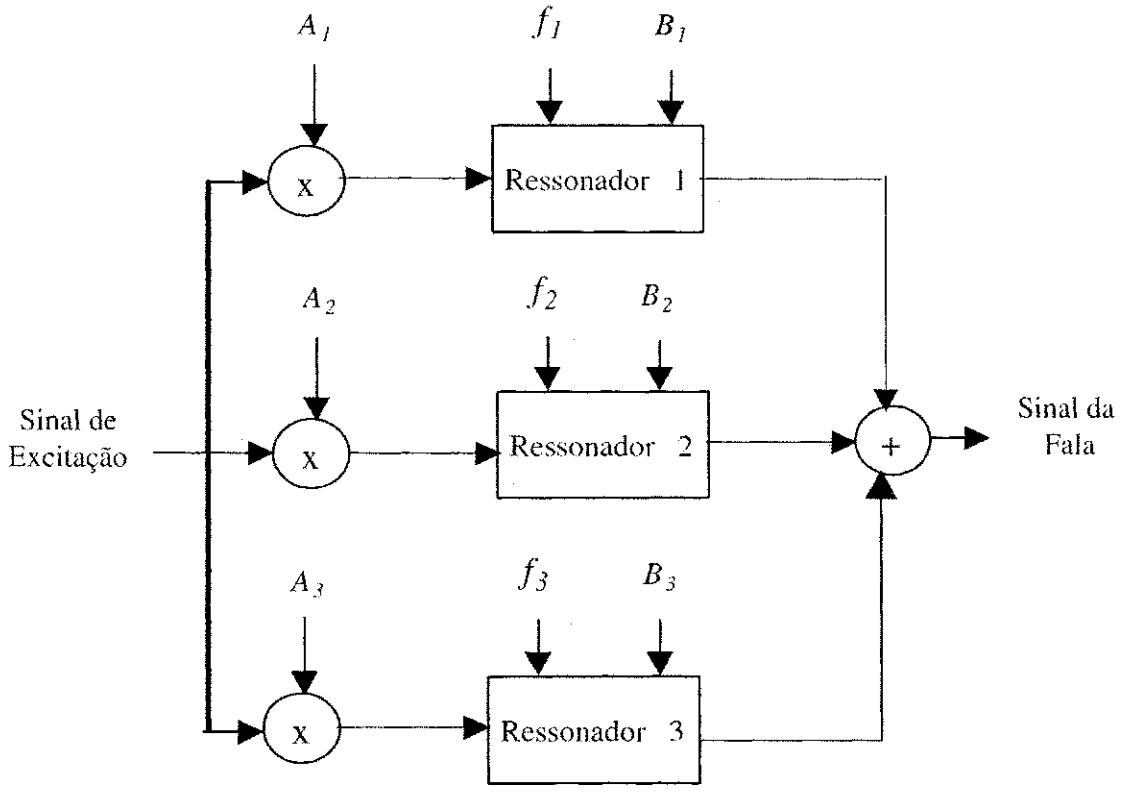

Figura 5.3: - Estrutura básica do sintetizador por formantes em paralelo [87].

Uma estrutura em paralelo produz os sons nasais, fricativos e plosivos, com melhor qualidade do que a estrutura em cascata. Por outro lado, a função de transferência dessa estrutura não é modelada adequadamente para a produção de vogais [87].

No sentido de melhorar a qualidade da fala sintetizada utilizando o método por formantes, foi desenvolvido um sintetizador mais complexo, denominado sintetizador de Klatt, que incorpora as estruturas em cascata e em paralelo [87]. Na estrutura em cascata são utilizados seis ressonadores e um antiressonador (para implementar zeros

na função de transferência dada pela Equação (5.1), sendo cinco ressonadores para atender os sons sonoros não nasalizados e um ressonador juntamente com um antiressonador para atender os sons nasalizados. Na estrutura em paralelo são utilizados seis ressonadores para os sons sonoros não nasalizados, um antiressonador para os sons nasalizados e uma conexão de *by-pass* para permitir a simulação de sons que não têm características de ressonância bem definida.

O sintetizador de Klatt tem uma fonte de excitação complexa e é controlado por 39 parâmetros que são atualizados a cada 5 ms. Apesar da complexidade, esse sintetizador foi incorporado em vários sistemas TTS (*Text-to-Speech*) tais como o Klattalk [87], o DECtalk [125] e em um conversor texto-fala para a língua portuguesa desenvolvido por Gomes em [2].

5.3 Síntese Concatenativa

Em síntese concatenativa, a fala sintética é produzida pela concatenação de segmentos correspondentes a unidades acústicas, que são previamente gravados e armazenados em uma base de dados (dicionário) [8, 126, 127, 128, 129]. As unidades podem ser: fones, difones, polifones, sílabas, demissílabas, palavras e frases [87]. Palavras e frases têm sido usadas em sistemas de reprodução vocal e difones, polifones, sílabas e demissílabas têm sido usadas em conversores texto-fala. A concatenação das unidades é realizada de forma ordenada, conforme a seqüência de comandos oriunda do processamento lingüístico e do processamento prosódico em um conversor texto-fala, ou através de comandos em um sistema de reprodução vocal.

Um sintetizador utilizando esse método de síntese pode ser representado pelo diagrama de blocos da Figura 5.4.

Assim, a fala sintetizada pode ser obtida a partir das seguintes etapas:

1ª Etapa:

Na primeira etapa uma seqüência de segmentos fonéticos com os respectivos parâmetros prosódicos, relativos a determinado texto, é aplicada na entrada do sintetizador.

2ª Etapa:

Na segunda etapa são selecionadas as unidades acústicas no dicionário, correspondentes à informação recebida na entrada do sintetizador.

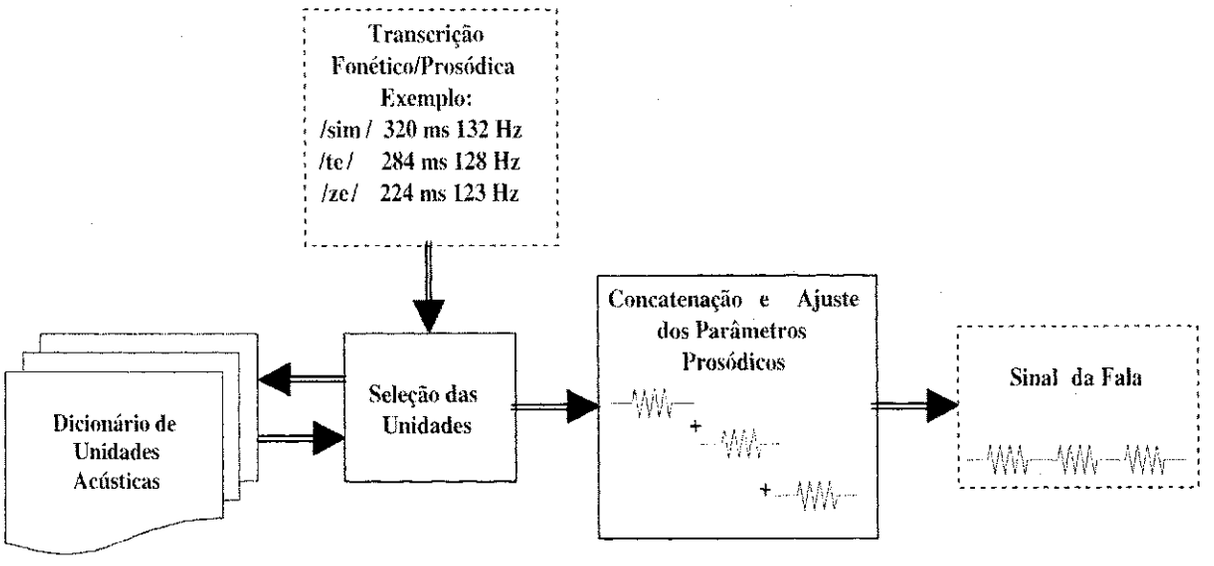

Figura 5.4: - Diagrama esquemático de um sintetizador concatenativo [8].

3ª Etapa:

Na terceira etapa as unidades acústicas podem ser concatenadas na forma em que estão ou após ajustes de duração e/ou *pitch*, conforme a descrição prosódica.

A principal vantagem deste método reside no fato de que o processamento do sinal de fala (modificações na freqüência fundamental, duração, etc.) é feito na própria forma de onda, como mostrado na seção seguinte, com a técnica PSOLA, mantendo-se assim as características originais do sinal, sobretudo o timbre [1]. Além disso, é mais simples de se produzir uma fala sintetizada de forma inteligível e natural, pois não há necessidade de definição de regras de transição entre os sons, como ocorre na síntese por formantes. Essas transições podem estar incorporadas nas unidades acústicas armazenadas em um dicionário.

Por outro lado, vários problemas podem ocorrer na síntese concatenativa quando comparada a outros métodos, tais como: descontinuidades no envelope espectral, descontinuidades de amplitude, de *pitch* e de fase entre os segmentos, e necessidade de uma quantidade de memória relativamente grande (dezenas de *megabytes*), quando segmentos maiores, tais como sílabas, são usados [87]. As descontinuidades espectrais

[1] Timbre é o efeito acústico resultante dos diversos graus de abertura da cavidade bucal, isto é, da distância entre a língua e o céu da boca, distância que é a máxima para o *a*, a mais aberta das vogais, e a mínima para o *i* e para o *u*, as mais fechadas [1].

ocorrem quando os formantes de segmentos adjacentes não têm os mesmos valores e estão relacionadas, principalmente, à coarticulação. Esse problema pode ser atenuado com um suavizamento nas bordas dos segmentos [10, 130]. Também a quantidade de memória requerida não é um problema relevante, pois atualmente tem-se memórias para computadores da ordem de dezenas de *gigabytes*, porém, a complexidade de acesso aos dados pode dificultar o processamento do texto e do sinal de fala em tempo real.

Visando, então, solucionar alguns problemas surgidos na concatenação, como também realizar ajustes nos parâmetros prosódicos (duração e *pitch*), quando necessário, foram desenvolvidos alguns algoritmos como os apresentados nas subseções seguintes.

5.3.1 Técnica PSOLA

Através da técnica PSOLA (*Pitch-Syncronous Overlap-Add*) é possível realizar a concatenação de unidades acústicas com a superposição e soma de blocos do sinal, deslocados no tempo, de maneira síncrona com o período de *pitch* [87]. Também é possível realizar alterações na duração e *pitch* do sinal da fala. As alterações podem ser realizadas no domínio da freqüência (Frequency Domain: FD-PSOLA), ou diretamente no domínio do tempo (Time Domain: TD-PSOLA). Na técnica FD-PSOLA, tem-se uma solução mais complexa, porém, uma flexibilidade maior de modificar as características espectrais do sinal, enquanto que, na técnica TD-PSOLA, tem-se soluções mais simples e eficientes para a implementação em tempo real, sendo, portanto, a mais utilizada em conversores texto-fala [29, 43, 131, 132, 133].

Etapas da Técnica TD-PSOLA

A síntese utilizando a técnica TD-PSOLA envolve basicamente três etapas [29]:

1ª Etapa:

Na primeira etapa, o sinal original digitalizado, $s(n)$, é transformado em uma seqüência temporal de sinais '*janelados*' de curta duração, $s_m(n)$, denominados *sinais elementares de análise*, os quais constituem os blocos básicos utilizados pelo processo de síntese e são determinados pela Equação (5.5).

$$s_m(n) = h_m(t_m - n).s(n) \tag{5.5}$$

Na prática, este resultado é obtido submetendo-se o sinal original a uma seqüência de '*janelamentos*' de Hamming [134, 135]. $h_m(n)$, no qual o posicionamento das janelas

de análise é síncrono com o período de *pitch* do sinal. O '*janelamento*' é feito de modo que haja sobreposição de 50%, como ilustrado na Figura 5.5.

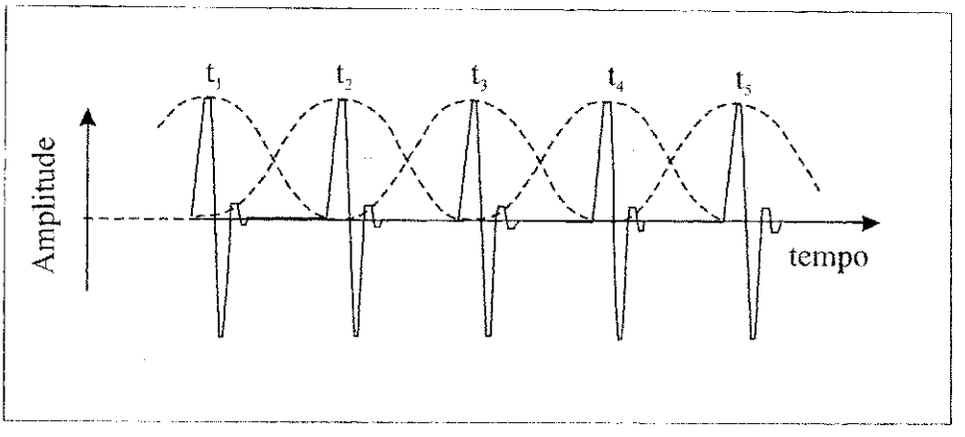

Figura 5.5: - Janelamento do sinal de análise.

Desta forma, as '*janelas*' são centradas em torno de sucessivos instantes t_m denominados marcas de *pitch* ou marcas de análise, os quais são colocados em uma taxa síncrona com o *pitch* nas porções sonoras do sinal e a uma taxa constante nas porções não sonoras. As marcas de *pitch* podem ser obtidas manualmente ou de maneira automática, por meio de técnicas de estimação de *pitch* [71, 95, 96, 97].

2ª Etapa:

Na segunda etapa é gerada uma seqüência de sinais elementares de síntese, $s_q(n)$, também de curta duração, a partir da seqüência de sinais elementares de análise, $s_m(n)$. Por sua vez, os sinais $s_q(n)$ são sincronizados com um novo conjunto de marcas de *pitch*, t_q, denominadas marcas de síntese. Nesse caso, os instantes em que ocorrem as marcas de síntese dependem dos fatores de modificações na escala de tempo (duração) e na escala de *pitch*, denominados α e β, respectivamente, e o intervalo de tempo $[t_q - (t_{q-1})]$ entre duas marcas de *pitch* sucessivas deve ser igual ao período de *pitch* de síntese.

Assim, é realizado um mapeamento entre as marcas de *pitch* de síntese e análise, $t_q \rightarrow t_m$, especificando quais sinais de análise $s_m(n)$ devem ser selecionados para produzir um dado sinal de síntese $s_q(n)$.

Os sinais de síntese $s_q(n)$ são obtidos, na prática, copiando-se uma versão dos sinais de análise $s_m(n)$ correspondentes, trasladados conforme a seqüência de atrasos dada pela Equação (5.6):

$$\delta_q = (t_q - t_m) \tag{5.6}$$

De modo mais específico os sinais de síntese podem ser determinados pela Equação (5.7).

$$s_q(n) = s_m(n - \delta_q) = s_m(n + t_m - t_q) \tag{5.7}$$

Portanto, cada sinal elementar de síntese provém de um único sinal elementar de análise, determinado por uma função de deformação temporal t_m a t_q, relacionando as marcas de análise às marcas de síntese. A natureza exata dessa função depende do efeito desejado: modificação da duração ou da freqüência fundamental.

Modificação da Duração

A modificação da duração das unidades acústicas é obtida através da relação entre as marcas de análise e de síntese, o que acarreta eliminação ou adição de períodos do sinal de fala, ou seja, para produzir diminuição na duração da fala se faz necessário eliminar períodos, enquanto que para se obter um aumento na duração, períodos devem ser acrescidos. Na Figura 5.6 tem-se um exemplo do aumento da duração da fala por um fator $\alpha = 5/4$.

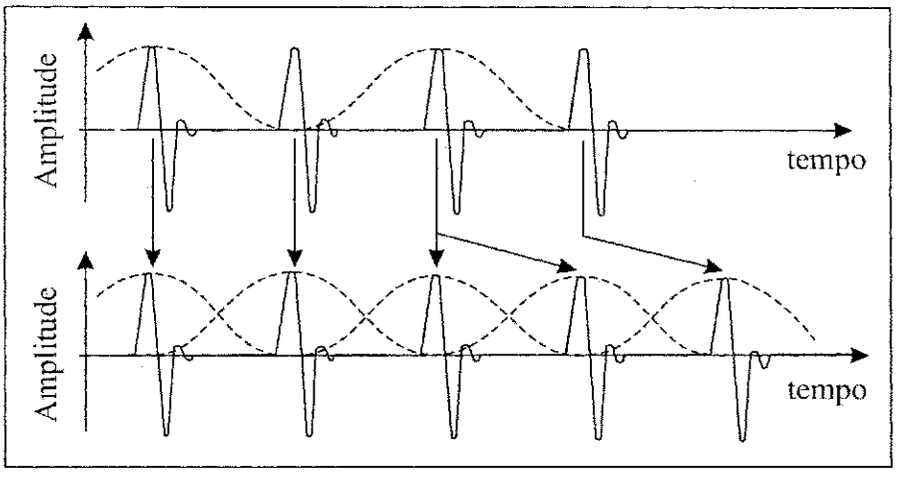

Figura 5.6: - Modificação da duração do sinal da fala por um fator de 5/4.

Modificação da Freqüência Fundamental

A modificação da freqüência fundamental do sinal original por um fator β é obtida alterando-se proporcionalmente, por esse fator, os intervalos de tempo (ou atrasos) entre dois sinais elementares sucessivos, conforme determinado na Equação (5.8).

$$t_{q+1} - t_q = (t_{m+1} - t_m)/\beta \tag{5.8}$$

Assim, se for desejado duplicar a freqüência fundamental, se faz necessário dividir o intervalo entre os blocos elementares por dois ($\beta = 2$), caso contrário, se for desejado reduzir a freqüência fundamental pela metade, deve-se multiplicar o intervalo entre os sinais elementares por dois ($\beta = 1/2$). Na Figura 5.7 tem-se um acréscimo na freqüência com um fator $\beta > 1$, no qual os períodos de síntese são menores do que os períodos de análise.

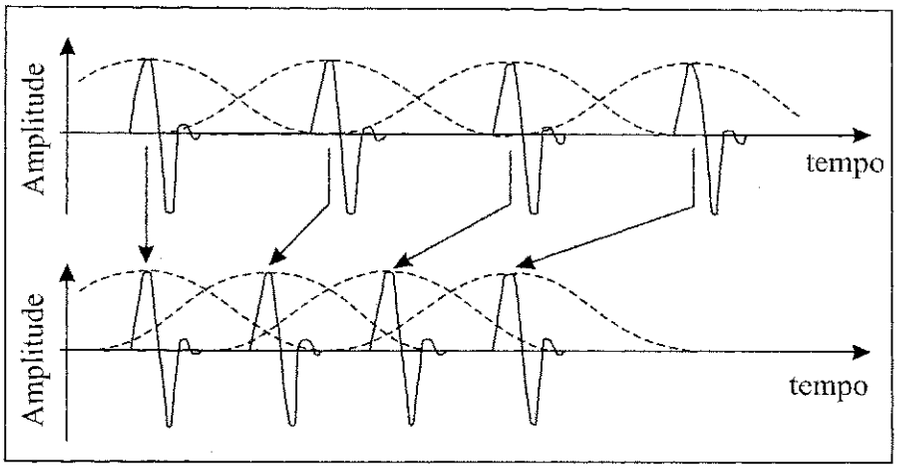

Figura 5.7: - Modificação da freqüência fundamental por um fator maior do que 1.

Modificações na freqüência fundamental são ligeiramente mais complicadas do que modificações na duração devido à influência da primeira sobre a segunda, sendo necessário realizar uma compensação na duração.

3ª Etapa:

Na terceira e última etapa é obtido o sinal de síntese da fala, $s_f(n)$, por simples superposição e adição dos sinais elementares de síntese conforme a Equação (5.9).

$$s_f(n) = \sum_q s_q(n) \tag{5.9}$$

Na Equação (5.9) o sinal sintetizado surge como uma simples combinação linear de versões janeladas e transladadas do sinal original. Todas as operações envolvidas são lineares com exceção da operação de '*janelamento*'.

Limitações

A técnica TD-PSOLA, na sua forma original, apresenta algumas limitações, principalmente para grandes variações prosódicas, conforme relacionado a seguir [136]:

- As modificações de duração só podem ser implementadas em determinados níveis de valores fracionados ou inteiros pré-estabelecidos (...,1/2, 2/3, 3/4, ..., 4/3, 3/2, 2/1, ...);

- As modificações de freqüência introduzem uma alteração na duração, causada pela maior ou menor superposição dos segmentos '*janelados*', variação essa que deve ser compensada de forma adequada;

- Durante o aumento de duração efetuado em porções não sonoras do sinal de fala, a repetição de segmentos introduz uma periodicidade que é responsável por uma aparência '*metálica*' da fala sintetizada;

No sentido de resolver os problemas apresentados, foram desenvolvidas novas técnicas como, por exemplo, a técnica MBR-PSOLA (*Multi-Band Resyntesis Pitch Syncronous OverLap Add*) [10] e a técnica híbrida, que combina uma análise/ressíntese LPC para a parte não sonora do sinal de fala, com um modelo harmônico (somatório de senóides com freqüências múltiplas da freqüência fundamental) da parte sonora desse sinal [137, 138]. Uma síntese dessas técnicas é apresentada a seguir.

5.3.2 Técnica MBR-PSOLA

A técnica MBR-PSOLA é baseada na resíntese harmônica [10], em que são feitos ajustes em quadros sonoros de segmentos do dicionário de unidades acústicas, usado na técnica TD-PSOLA, de modo que as descontinuidades de fase, *pitch* e envelope espectral sejam eliminadas. Como resultado tem-se um esquema melhorado para a síntese com a técnica TD-PSOLA e, conseqüentemente, uma melhor qualidade da fala sintetizada.

A estrutura geral de um sintetizador MBR-PSOLA é apresentada na Figura 5.8. Como se observa nesta figura, além de um novo dicionário obtido através da ressíntese harmônica (Dicionário MBR-PSOLA) é adicionado uma Interpolação Linear Temporal ao estágio de concatenação com a técnica TD-PSOLA. Essa interpolação atua nos períodos próximos à região de junção das unidades acústicas, resultando em uma suavização espectral impossível de se obter com a técnica TD-PSOLA [29].

O custo computacional acrescentado nessa técnica é pequeno comparado à técnica TD-PSOLA, pois o processo de análise/ressíntese é efetuado uma única vez durante a criação do dicionário. Atualmente, a técnica MBR-PSOLA foi modificada, dando lugar

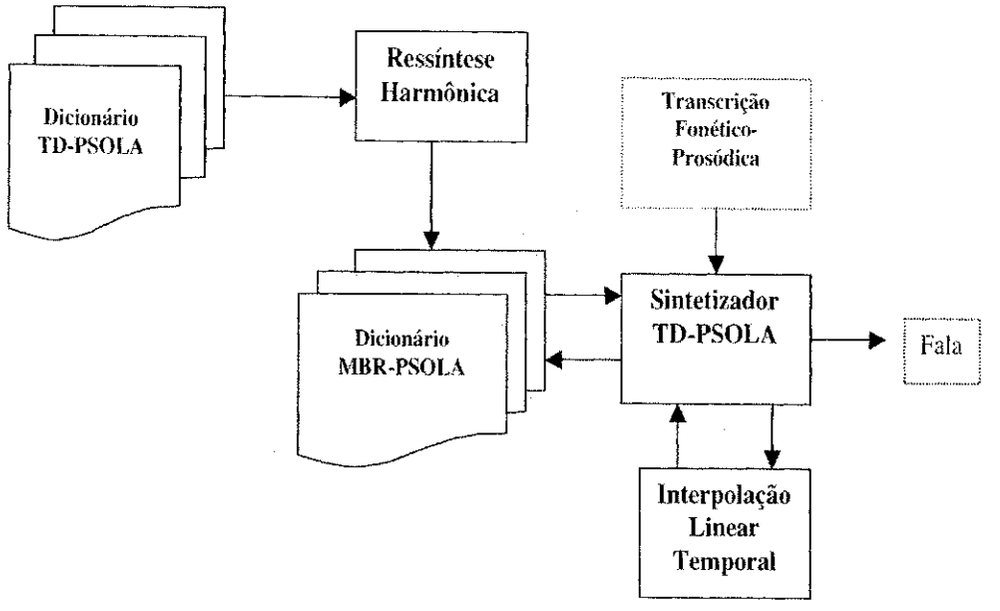

Figura 5.8: - Diagrama de blocos do sintetizador MBR-PSOLA.

a uma versão mais eficiente denominada MBROLA (*Multi-Band Resynthesis Overlap and Add*) [9, 11, 139].

5.4 Análise/Ressíntese LPC

A técnica de análise/ressíntese LPC (*Linear Predictive Coding*) foi desenvolvida inicialmente para sistemas de codificação da fala, mas pode também ser usada na síntese da fala [87, 140]. Essa técnica é baseada no modelo fonte-filtro, de forma semelhante à técnica de formantes [124]. Em codificação, a técnica LPC tem como objetivo sintetizar o sinal de saída o mais próximo do sinal de entrada, através da transmissão dos parâmetros da fonte e dos coeficientes do filtro que representam o trato vocal. Em síntese da fala, ela tem como objetivo permitir as alterações nas escalas de duração e *pitch* sem degradar o sinal [8]. Assim, ao invés de se transmitir ou armazenar as amostras do sinal de fala, se transmitem ou armazenam os parâmetros do modelo de produção da fala, tornando mais fácil o controle da prosódia da fala que está sendo sintetizada.

A predição linear, utilizada no filtro digital do modelo, é baseada no princípio de que a amostra atual do sinal da fala $y(n)$ pode ser aproximada ou predita por um número finito de P amostras anteriores de $y(n\text{-}1)$ a $y(n\text{-}k)$ através de uma combinação

linear, acrescentada de um pequeno erro $e(n)$ chamado *erro residual do sinal* [87]. Logo:

$$y(n) = e(n) + \sum_{k=1}^{P} a(k)\ y(n-k) \tag{5.10}$$

Então, o erro residual pode ser dado por:

$$e(n) = y(n) - \sum_{k=1}^{P} a(k)\ y(n-k) = y(n) - \tilde{y}(n) \tag{5.11}$$

onde $\tilde{y}(n)$ é um valor predito (estimado), P é a ordem do preditor, e $a(k)$, com $k = 1, ..., P$ são os coeficientes do preditor, calculados minimizando-se a diferença quadrática entre as amostras originais e as amostras linearmente preditas [124].

Existem vários métodos para a determinação dos coeficientes do preditor, dentre os quais se destacam: o método de autocorrelação, em que é realizado um *janelamento* do sinal da fala e o erro é minimizado no intervalo $0 \leq n \leq (N + P - 1)$, e o método da covariância, em que é feito um *janelamento* do erro, calculado no intervalo $0 \leq n \leq (N - 1)$, onde N é o comprimento da janela e P a ordem do preditor. O método da autocorrelação é mais eficiente em termos computacionais do que o método da covariância, pois o último requer um número maior de multiplicações para a resolução das equações matriciais [141].

O exemplo clássico de um sistema utilizando análise/ressíntese é o VOCODER - LPC (*voice coder* - LPC) . Esse sistema é formado por um módulo de análise e por um módulo de síntese, conforme mostra a Figura 5.9 [73, 142].

No módulo de análise, é feita a detecção sonoro/não-sonoro. Quando detetado um som sonoro, é determinado o período de *pitch*. Paralelamente é feita uma análise LPC para a determinação dos coeficientes do preditor do filtro de síntese e do Ganho. Esses parâmetros são determinados e transmitidos para o módulo de síntese.

No módulo de síntese o período de *pitch* é utilizado para habilitar o gerador de excitação a produzir uma seqüência de pulsos para a excitação sonora. De outra forma, o gerador produz um ruído para a excitação não-sonora. Além disso, os coeficientes do preditor e o Ganho são aplicados ao filtro de síntese para controle da fala sintetizada.

Observa-se também, nessa técnica, que a função de transferência do modelo básico é constituída apenas por pólos, de modo que todos os fonemas que contêm antiformantes como consoantes nasais e vogais nasalizadas não são modelados de forma adequada, reduzindo a qualidade do sinal de fala produzido [87]. Assim, foram desenvolvidos

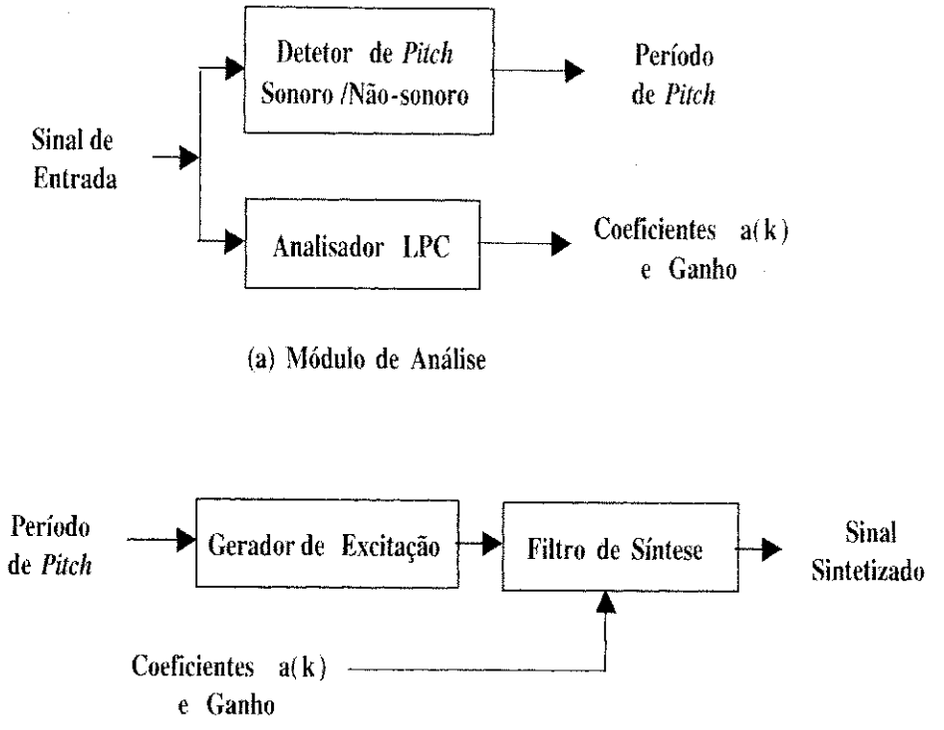

Figura 5.9: - Diagrama de blocos de um VOCODER - LPC.

vários algoritmos com o objetivo de melhorar a qualidade do sinal de fala, como, por exemplo, Codificação por Excitação Residual Preditiva Linear (*Residual Excited Linear Predictive Coding* - RELP), no qual um sinal, denominado *sinal de resíduo*, obtido através de um filtro introduzido no módulo de análise, é usado como um sinal de excitação para o filtro de síntese [124, 142]. A técnica RELP foi usada em um trabalho desenvolvido por Pacheco [8], para alteração dos parâmetros prosódicos em um conversor texto-fala para a Língua Portuguesa falada no Brasil, com bons resultados, segundo o autor. Um VOCODER de alta qualidade também foi usado recentemente por Toda *et al.* [140], para testar um algoritmo de seleção de unidades acústicas, na síntese da fala para a língua japonesa. Segundo os autores, foram obtidos resultados satisfatórios, sobretudo com relação à naturalidade da fala sintetizada.

5.5 Síntese Híbrida

Na síntese híbrida, são desenvolvidos modelos no intuito de eliminar os problemas inerentes às técnicas PSOLA e LPC. Nesses modelos são dados tratamentos diferentes aos componentes sonoro e não-sonoro do sinal da fala, de modo que o componente sonoro é representado por um modelo harmônico (somatório de senóides com freqüências múltiplas da freqüência fundamental) [137, 138], enquanto que o componente não-sonoro é modelado como uma excitação aleatória aplicada a um filtro LPC [73].

A síntese híbrida é, portanto, composta por duas etapas: Uma etapa de *análise*, que é efetuada uma única vez com a criação de um dicionário de unidades acústicas, na qual são determinados os componentes harmônicos e de ruído do sinal original; e uma etapa de *síntese*, na qual cada um dos componentes do sinal é submetido às alterações prosódicas necessárias, de modo que o sinal sintetizado corresponde ao somatório dos dois componentes após efetuar-se o processo de ajustes [29].

Vários algoritmos têm sido propostos para a síntese híbrida, que diferem basicamente quanto à forma de manusear os parâmetros harmônicos e estocásticos [10]. Como exemplo, pode ser citado o algoritmo desenvolvido Violaro e Boeffard em [143], no qual o sinal da fala $s(n)$ é inicialmente submetido a uma análise em sincronismo com o *pitch* e decomposto em um componente harmônico $s_h(n)$, com freqüência variável até determinado valor máximo, e um componente de ruído $r(n)$, conforme descrito pelas Equações (5.12) e (5.13).

$$s(n) = s_h(n) + r(n) \tag{5.12}$$

$$s_h(n) = \sum_{k=0}^{K} A_k cos(k w_o n + \theta_k) \tag{5.13}$$

onde K é o número de harmônicos, w_o é a freqüência fundamental, A_k e θ_k são amplitude e fase do k-ésimo harmônico.

Para efeito de simplificação, é considerado no modelo que os componentes ocupem faixas de freqüências distintas, de modo que o componente harmônico vai de 0 a Kw_o radianos, e o componente de ruído, de Kw_o a π radianos. O somatório da equação (5.13) é iniciado em $k = 0$, considerando-se a possibilidade da presença de um nível DC no segmento de fala de curta duração, e a freqüência máxima de Kw_o é tornada variável para evitar qualquer periodicidade no componente de ruído e qualquer ruído no compo-

nente harmônico. Nos segmentos não-sonoros, o componente harmônico é considerado zero. Entretanto, nos segmentos sonoros devem ser considerados ambos os componentes, para modelar sons mistos como os fricativos sonoros, os plosivos sonoros, segmentos transacionais nos limites sonoro/não-sonoro, como também qualquer outro tipo de ruído presente no sinal da fala. Quando são realizadas modificações de freqüência, um novo conjunto de parâmetros harmônicos é avaliado pela reamostragem do envelope espectral. Para a síntese dos componentes harmônicos com modificações de duração e/ou freqüência, é introduzida uma correção de fase nos parâmetros harmônicos [143].

Observa-se, também, que a síntese híbrida requer um custo computacional relativamente elevado, porém ela apresenta várias vantagens sobre a técnica PSOLA original, relacionadas a seguir [143, 144] :

1. evita sons '*metálicos*' produzidos com TD-PSOLA;
2. permite modificações de *pitch* e duração maiores;
3. permite modificações contínuas de *pitch* e duração;
4. permite modificações de *pitch* sem compensação de duração;
5. permite o controle de *pitch* e duração sobre cada marca de *pitch*, de forma mais flexível do que no TD-PSOLA, no qual normalmente apenas um controle de duração e dois ou três de *pitch* estão disponíveis por fone;
6. permite uma suavização entre as unidades de forma simples e bastante flexível.

A principal desvantagem do modelo é a extrema sensibilidade a erros de classificação sonoro/não-sonoro. Esse problema pode ser atenuado com ajustes feitos no algoritmo que determina as marcas de *pitch* [143]. Um trabalho recente utilizando essa técnica foi desenvolvido por Jilka [145], para a língua inglesa. Nesse trabalho foi testada a qualidade da síntese da fala para vários métodos de variação dos parâmetros prosódicos.

5.6 Discussão

Na literatura consultada foram encontradas várias técnicas de síntese do sinal de fala, que podem ser incorporadas em um conversor texto-fala. A qualidade da fala produzida na saída do estágio de síntese depende, não só das informações recebidas do

processamento do texto e processamento prosódico, como também da técnica de síntese utilizada.

Na síntese articulatória o trato vocal é modelado de forma direta e envolve conhecimentos de fonética acústica, como também de fonética articulatória [146]. Apesar de ser um método atrativo, pelas razões apresentadas, é quase impossível modelar perfeitamente todos os movimentos dos articuladores [29, 87]. Além disso, a coleta de dados e a implementação de regras são complexas, o que pode resultar em uma carga computacional relativamente grande quando comparada com as de outras técnicas de síntese [119]. Aplicações dessa técnica são encontradas em trabalhos desenvolvidos por Perkell, Prado e Guiard-Marigny em [119, 121, 122] e, mais recentemente, por Huang em [123]. Devido aos problemas apresentados, não se tem conseguido o mesmo nível de sucesso nessa técnica quando comparada às demais, porém, pode ser considerada como uma opção promissora para o futuro.

A síntese por formantes é baseada no modelo fonte-filtro, no qual é reproduzida a trajetória dos formantes de cada fone, na produção da fala [124]. Assim, é possível se obter uma variedade "infinita" de sons da fala, a partir do ajuste de parâmetros do modelo da fonte, tornando o método mais flexível do que outros, como, por exemplo, o concatenativo [87]. Por outro lado, a obtenção dos parâmetros de controle do modelo apresenta certo grau de dificuldade, principalmente na transição entre segmentos fonéticos diferentes, a fim de levar em conta os fenômenos articulatórios, resultando em um conjunto de regras relativamente grande [8, 29]. Apesar das dificuldades apresentadas, esse método resulta em uma qualidade da fala relativamente boa quando operado corretamente [10] e tem sido utilizado em vários sistemas comerciais tais como: Klattalk [87], o DECtalk [125] e em um conversor texto-fala para a língua portuguesa desenvolvido por Gomes em [2], dentre outros.

A síntese concatenativa é baseada na junção de unidades acústicas (segmentos de fala), extraídas de gravações prévias realizadas por um só locutor. Normalmente, as unidades acústicas utilizadas nesse tipo de síntese, como, difones, sílabas, demissílabas, etc., contêm transições entre os fonemas as quais precisam ser implementadas em sistemas de síntese paramétricos, como, por exemplo, os sistemas de síntese por formantes. Também o processamento do sinal da fala (modificações de *pitch* e duração) é feito diretamente na forma de onda, mantendo-se as características originais do sinal, que pode resultar em uma fala mais inteligível e natural do que em outros métodos [87]. Essa técnica é limitada a um locutor e um tipo de voz e pode requerer mais memória

no computador para armazenar os dados do que os outros métodos. Também podem surgir distorções devido a descontinuidades nos pontos de concatenação. Alguns dos problemas apresentados podem ser resolvidos com a técnica MBROLA [9, 11, 139], ou até mesmo com a técnica híbrida [143, 144, 145]. Em função da simplicidade, do baixo custo, do aumento da capacidade de memória nos computadores atuais, de técnicas de compressão de dados e da qualidade de fala produzida, a síntese concatenativa tem sido usada em vários sistemas de conversão texto fala atualmente [127, 128, 129, 132, 133].

A análise/ressíntese LPC é uma técnica de codificação da fala baseada no modelo fonte-filtro, bastante usada em reconhecimento de fala. O uso dessa técnica na síntese da fala tem a vantagem de permitir o controle, de forma independente, do *pitch*, da duração, do envelope espectral e do ganho dos segmentos de fala, como também permite realizar uma representação precisa e compacta do sinal de fala [8]. O modelo básico tem a desvantagem de não manter a forma de onda do sinal de fala natural durante o processo de síntese, como também as vogais nasalisadas não são modeladas de forma adequada, reduzindo a qualidade do sinal de fala produzido [87]. Assim, foram desenvolvidos vários algoritmos com o objetivo de melhorar a qualidade do sinal de fala, usando-se essa técnica, com resultados satisfatórios [8, 140].

Na síntese híbrida são dados tratamentos diferentes aos componentes não-sonoro e sonoro do sinal da fala, para eliminar os problemas ocorridos com a técnica PSOLA e LPC, na tentativa de produzir uma fala de qualidade superior aos demais métodos [138]. As principais desvantagens do modelo são a grande sensibilidade aos erros de classificação sonoro/não-sonoro e ao elevado custo computacional [143]. Um sistema de síntese da fala utilizando essa técnica foi desenvolvido por Violaro em [143], e mais recentemente por Jilka em [145], com resultados satisfatórios, segundo os autores.

Conclui-se que todas as técnicas de síntese apresentam vantagens e desvantagens e a escolha de uma delas depende do tipo de aplicação e da qualidade da fala requerida. Dentre as técnicas apresentadas, foi escolhida a síntese concatenativa para os testes do modelo prosódico proposto neste trabalho. A concatenação é realizada de forma simples, através da junção das unidades acústicas selecionadas no dicionário, de forma seqüencial, conforme os comandos provenientes do estágio de processamento prosódico. A escolha foi realizada considerando-se as vantagens de simplicidade, flexibilidade e sobretudo porque o processamento do sinal de fala é feito na própria forma de onda, mantendo-se assim as características originais desse sinal.

Capítulo 6

Desenvolvimento do Dicionário de Unidades Acústicas

O modelo prosódico proposto neste trabalho é aplicado a um conversor texto-fala utilizando síntese concatenativa e, para tal, é necessário obter os segmentos de fala, correspondentes às unidades acústicas, que devem ser previamente gravados e armazenados em uma base de dados (dicionário). O desenvolvimento de um dicionário desse tipo é realizado basicamente a partir das seguintes etapas [8, 10]:

1. Escolha do tipo de segmento ou de unidade acústica;
2. Elaboração de um *corpus* para extração das unidades;
3. Gravação do *corpus*;
4. Segmentação de cada unidade a partir da fala gravada.

Além disso, é importante que as unidades do dicionário contenham características articulatórias correspondentes aos fenômenos de coarticulação na geração da cadeia fonética, para a obtenção de uma fala sintetizada de melhor qualidade.

Assim, neste capítulo, são apresentadas as etapas básicas a serem consideradas na concepção de um dicionário de unidades acústicas. Também é apresentada a metodologia usada no desenvolvimento de um dicionário para um conversor texto-fala concatenativo para o Português Brasileiro, usado em testes no modelo prosódico proposto.

6.1 Escolha do Tipo de Unidade

Em um sistema de síntese concatenativa podem ser usados os seguintes tipos de unidades: fones, difones, polifones, demissílabas, sílabas, palavras ou frases. As unidades devem ser escolhidas de modo que permitam [8, 87]:

1. pequenas distorções na concatenação;
2. captura da maior quantidade possível de efeitos coarticulatórios na transição entre os fonemas;
3. número e tamanho reduzidos o máximo possível.

Na prática, estes critérios são conflitantes de modo que é necessário estabelecer um compromisso entre o tamanho do segmento que contemple o maior número de efeitos coarticulatórios e o número total de unidades do dicionário. A Tabela 6.1, por exemplo, apresenta uma relação entre os tipos de unidades mais comuns, juntamente com informações sobre tamanho do segmento, número de unidades necessárias e qualidade obtida [147]. Uma descrição mais completa de cada tipo de unidade é apresentada nas subseções seguintes.

Tabela 6.1: Relação entre tamanho, número e qualidade dos diferentes tipos de unidades (adaptado de [147])

Tamanho do Segmento	Tipo de Unidade	Número de Unidades (aprox.)	Qualidade
Curto	Fonema	35	Baixa
↓	Difone	1.500	↓
↓	Demissílaba	2.000	↓
↓	Trifone	30.000	↓
↓	Sílaba	11.000	↓
↓	Palavra	100.000 - 1.500.000	↓
Longo	Frase	∞	Alta

6.1.1 Frases e Palavras

As frases e palavras atendem aos critérios de escolha 1 e 2, porém não se aplicam a sistemas de conversão texto-fala, pois torna-se impraticável gravar e armazenar todas

as frases ou palavras de uma língua e suas variações prosódicas. Normalmente, tais unidades são usadas em sistemas de reprodução vocal com vocabulário limitado.

6.1.2 Fonemas Independentes do Contexto

Os fonemas são segmentos mínimos da fala e, se forem considerados independentes do contexto em que se inserem, tem-se de 33 a 34 unidades para a Língua Portuguesa [70, 107]. Nesse caso, os fones, correspondentes aos fonemas, atendem ao critério de escolha número 3, porém não produzem um resultado aceitável na síntese por concatenação devido aos efeitos coarticulatórios e à quantidade de descontinuidades relativamente grande.

6.1.3 Difones

O difone pode ser definido como sendo o segmento do sinal de fala que inicia na região espectralmente estável de um fone e termina na região estável do próximo fone contendo, portanto, uma transição completa entre dois fones [87]. Normalmente o difone tem sido usado como unidade básica na síntese concatenativa e nesse caso tem-se um valor em torno de N^2 unidades, onde N é o número de fonemas. Para a Língua Portuguesa tem-se aproximadamente 1200 difones.

6.1.4 Polifones

Os polifones são segmentos que contêm informações de três ou mais fones na sua composição. Um destaque é feito aos trifones. O trifone pode ser considerado como o segmento correspondente a um fonema com um contexto específico anterior e posterior, ou seja, ele se inicia na região estável de um fone, inclui o próximo fone e termina na região estável do fone seguinte [87]. Embora o número teórico de trifones seja N^3, onde N é o número de fonemas em determinada língua, muitas combinações não ocorrem na prática, sendo necessário realizar uma triagem nesse caso.

6.1.5 Demissílabas e Sílabas

As demissílabas, por sua vez, representam a metade inicial e final de sílabas [124]. Assim, essas unidades incluem um número maior de coarticulações do que os difones e, conseqüentemente, tem-se menos pontos de concatenação durante a síntese [87]. O

uso de demissílabas em lugar de sílabas reduz significativamente o número de unidades no dicionário. Por exemplo, são necessárias 1000 demissílabas para representar aproximadamente 10.000 sílabas na Língua Inglesa [124].

A sílaba é um fonema ou um grupo de fonemas emitido em um só impulso expiatório [80]. O número de sílabas em uma língua é menor do que o número de palavras, porém, é maior com relação ao número de difones, como mostrado no exemplo da Tabela 6.1. Apesar de ser necessário maior quantidade de memória para o dicionário, tem-se maior número de efeitos coarticulatórios incluidos nesse tipo de unidade e, para a Língua Portuguesa, torna-se uma boa opção devido à estrutura silábica na produção de palavras e frases.

Portanto, no dicionário desenvolvido neste trabalho são consideradas demissílabas e sílabas correspondentes às unidades silábico-fonéticas das palavras. Assim, tem-se um número maior de unidades com relação ao uso de difones (acima de 1200), porém, tem-se um número menor de unidades em relação ao uso exclusivo de sílabas e também uma maior representitatividade da coarticulação para a obtenção de uma fala mais natural possível.

Um exemplo de unidade usada no dicionário é apresentado na Figura 6.1. Nesta figura observa-se que é difícil definir exatamente onde fica a parte final do fonema /b/ e o início do fonema /a/ devido à coarticulação. Essa dificuldade torna-se evidente no estágio de segmentação das unidades a partir do *corpus* estabelecido.

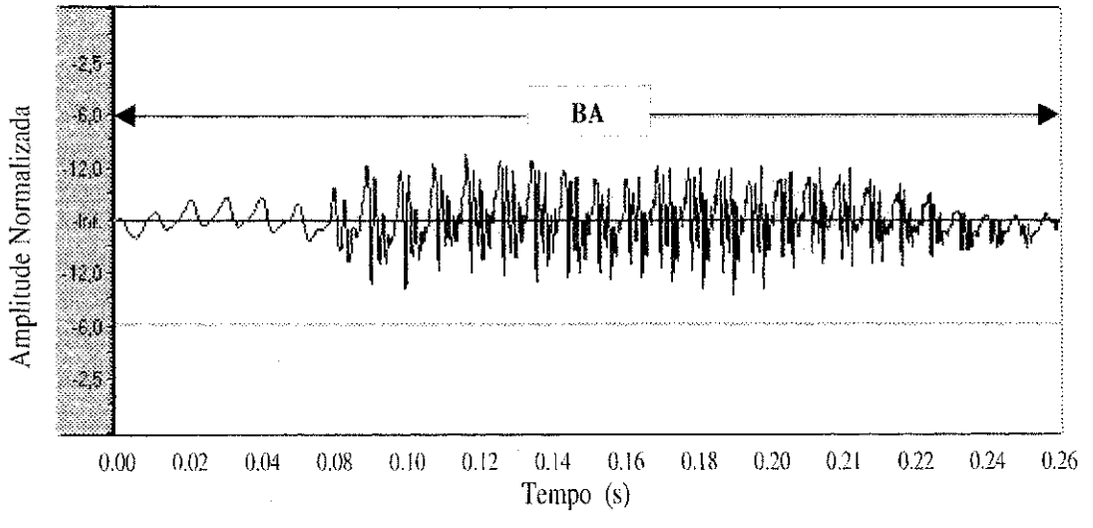

Figura 6.1: - Forma de onda da unidade acústica BA.

6.2 Determinação das Unidades

A determinação das unidades acústicas do dicionário foi realizada em duas etapas [148, 149]:

1. Identificação dos fonemas que fazem parte da língua;

2. Obtenção das possíveis combinações dos fonemas dentro da língua, considerando o tipo de unidade previamente escolhida.

Na primeira etapa, foi estabelecido um alfabeto fonético para a Língua Portuguesa contendo 34 fonemas, sendo vinte consoantes, doze vogais e duas semivogais [70, 107]. O código deste alfabeto foi denominado de AFLAPS (Alfabeto Fonético definido no Laboratório de Automação e Processamento de Sinais do DEE/UFCG), e é apresentado na Tabela 6.2, juntamente com os fonemas correspondentes ao alfabeto fonético internacional (IPA) [69] e ao alfabeto fonético SAMPA (*Speech Assessment Methods Phonetic Alphabet*), que é um alfabeto fonético compatível com o computador [150].

Na segunda etapa, foram obtidas as possíveis combinações das vogais e consoantes do alfabeto da Língua Portuguesa, considerando-se a constituição fonético-silábica das palavras [26, 107, 151]. Assim, a partir dos padrões: V, VC, VV, VVC, CV, CVC, CVV, CCV e CCVC, em que C é consoante e V é vogal, foram obtidas 1994 unidades. Um exemplo, com o padrão CV, é apresentado na Tabela 6.3. As unidades correspondentes aos demais padrões são apresentadas no Apêndice A.

6.3 Elaboração do Corpus

Um *corpus* pode ser constituído por um conjunto de textos, frases ou mesmo palavras isoladas, dos quais podem ser obtidas as unidades acústicas formadoras do dicionário. Devido à grande variação de freqüência de ocorrência das unidades da fala natural, tentar obter as unidades acústicas a partir da gravação de um texto qualquer é praticamente enviável. Algumas seriam repetidas várias vezes e outras não apareceriam. Isso pode ser observado, por exemplo, para a Língua Francesa, em que um dado conjunto de 100 frases foneticamente balanceadas cobre apenas 43% dos 1200 difones necessários, com uma redundância em torno de 80% [10]. Além disso, pode haver uma grande influência das unidades adjacentes sobre a unidade desejada com relação à prosódia, devido ao fenômeno de coarticulação. Esse fenômeno também está presente

Tabela 6.2: Alfabeto fonético AFLAPS, associado ao IPA e ao SAMPA

IPA	AFLAPS	SAMPA	Exemplos	IPA	AFLAPS	SAMPA	Exemplos
b	b	b	Bala	r	r	r	Puro
k	k	k	Casa		r2		Arpa
d	d	d	Dado	R	rr	R	Torre
g	g	g	Galo	a	a	a	Vale
p	p	p	Pato	â	am	a ~	Campanha
t	t	t	Tato	ɛ	ê	e	Pêra
f	f	f	Farofa	E	é	E	Quero
v	v	v	Vaca	ê	em	e ~	Quente
ʒ	j	ʒ	Janela	i	i	i	Pico
s	s	s	Sapo	î	im	i ~	Brinco
ʂ	x	ʂ	Xadrez	o	o	o	Tolo
z	z	z	Zebra	O	ó	O	Bola
m	m	m	Mala	ô	om	o ~	Ombro
n	n	n	Nariz	u	u	u	Duro
ɲ	nh	ɲ	Nhoque	û	um	u ~	Algum
l	l	l	Lata	y	y	y	Mais
ʎ	lh	ʎ	Alho	w	w	w	Mau

Tabela 6.3: Matriz referente às combinações consoante-vogal (CV)

C/V	a	am	é	ê	em	i	im	ó	ô	om	u	um
b	ba	bam	bé	bê	bem	bi	bim	bó	bô	bom	bu	bum
k	ka	kam	ké	kê	kem	ki	kim	kó	kô	kom	ku	kum
d	da	dam	dé	dê	dem	di	dim	dó	dô	dom	du	dum
g	ga	gam	gé	gê	gem	gi	gim	gó	gô	gom	gu	gum
p	pa	pam	pé	pê	pem	pi	pim	pó	pô	pom	pu	pum
t	ta	tam	té	tê	tem	ti	tim	tó	tô	tom	tu	tum
f	fa	fam	fé	fê	fem	fi	fim	fó	fô	fom	fu	fum
v	va	vam	vé	vê	vem	vi	vim	vó	vô	vom	vu	vum
j	ja	jam	jé	jê	jem	ji	jim	jó	jó	jom	ju	jum
s	sa	sam	sé	sê	sem	si	sim	só	sô	som	su	sum
x	xa	xam	xé	xê	xem	xi	xim	xó	xô	xom	xu	xum
z	za	zam	zé	zê	zem	zi	zim	zó	zô	zom	zu	zum
m	ma	mam	mé	mê	mem	mi	mim	mó	mô	mom	mu	mum
n	na	nam	né	nê	nem	ni	nim	nó	nô	nom	nu	num
nh	nha	nham	nhé	nhê	nhem	nhi	nhim	nhó	nhô	nhom	nhu	nhum
l	la	lam	lé	lê	lem	li	lim	ló	lô	lom	lu	lum
lh	lha	lham	lhé	lhê	lhem	lhi	lhim	lhó	lhô	lhom	lhu	lhum
r	ra	ram	ré	rê	rem	ri	rim	ró	rô	rom	ru	rum
rr	rra	rram	rré	rrê	rrem	rri	rrim	rró	rrô	rrom	rru	rrum

em palavras, de modo que o uso de unidades extraídas dessas palavras também pode produzir uma fala sintetizada com efeitos de descontinuidade, quando uma unidade é concatenada com outras em um contexto diferente, do qual foi obtida.

Uma opção bastante interesssante para se obter um *corpus*, e utilizada neste trabalho, é através de *logatomos*. O *logatomo* é constituído por uma sílaba ou por uma seqüência de sílabas que pertencem a uma língua, mas que não formam uma palavra ou um sintagma significativo [69]. Nesse caso, tem-se uma redução significativa dos efeitos de coarticulação e podem ser obtidas todas as unidades requeridas no dicionário de um conversor texto-fala concatenativo.

Para facilitar o processo de segmentação das unidades são utilizados *logatomos* com três sílabas, nos quais as unidades de interesse são inseridas na sílaba central, como, por exemplo, a unidade [**ka**] inserida no logatomo *pa***ka***pa*. Observa-se também que a unidade na posição central do *logatomo* apresenta uma maior estabilidade do que as demais quando o vocábulo é enunciado. As consoantes mais "neutras" são as bilabiais, mais precisamente, a bilabial surda [p], e a vogal cujas transições menos afetam os demais segmentos corresponde ao [a] [146]. Assim, a maioria das sílabas laterais são constituídas pela consoante bilabial [**p**] e a vogal [**a**]. Quando as unidades de interesse são iniciadas por vogais (tipo VV ou VC), devem ser usados *logatomos* iniciados por vogais, e quando forem iniciadas por consoante (tipo CV ou CCV), os *logatomos* devem ser iniciados por consoante, como, por exemplo, em *ap***cu***pa* e *pa***ka***pa*.

Considerando-se as unidades relacionadas na Tabela 6.3 e nas tabelas do Apêndice A, foi obtido um *corpus* com 1994 logatomos, conforme os critérios aqui apresentados.

6.4 Gravação do Corpus

Após estabelecido o *corpus*, ele foi lido da forma neutra (monótona), gravado e digitalizado utilizando-se uma placa SOUND BLASTER da CREATIVE LABS. Na digitalização foi utilizada uma taxa de amostragem de 16 kHz, uma resolução de 16 *bits* por amostra, em formato mono e com a técnica de codificação PCM (*Pulse Code Modulation*). A codificação PCM é realizada diretamente na forma de onda, e a taxa de 16 kHz aliada à resolução de 16 bits têm sido usadas como padrão em conversores texto-fala de qualidade, como, por exemplo, o conversor MBROLA [11, 139, 36].

Assim, foram gravados blocos de logatomos do *corpus* elaborado na seção anterior, conforme a semelhança das sílabas adjacentes a cada unidade de interesse e a

capacidade do locutor, havendo um espaçamento de tempo entre um logatomo e o subseqüente. Cada bloco de logatomos foi armazenado com o nome do grupo das unidades constituintes e as numerações correspondentes. A Figura 6.2 ilustra a forma de onda do logatomo *pa***ka***pa* gravado.

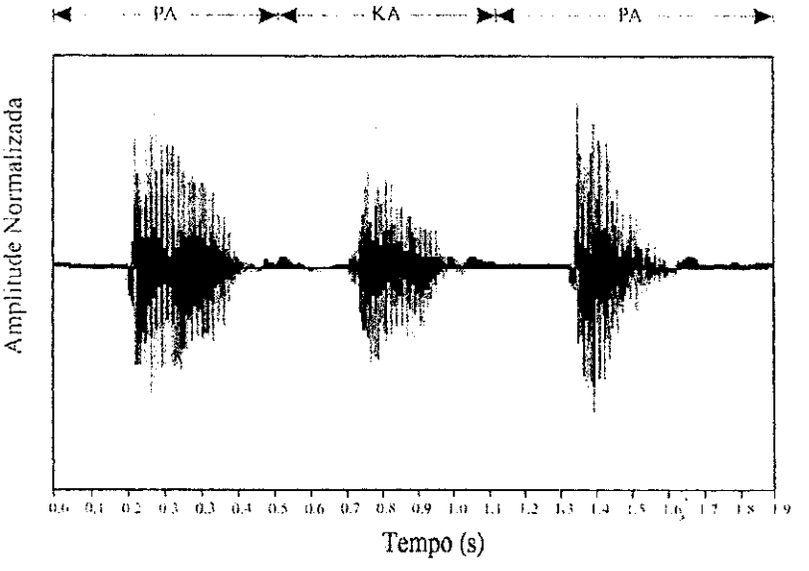

Figura 6.2: - Forma de onda do logatomo *pa***ka***pa*.

A gravação foi realizada em estúdio profissional, para evitar o ruído ambiental, e com equipamentos de áudio de alta qualidade (microfone, pré-amplificador e conversor A/D), com o objetivo de eliminar, ao máximo, os ruídos que tais dispositivos possam produzir no sinal da fala. Além de um locutor, foi também necessário um operador para a seleção e armazenagem da informação na memória do computador.

Durante a gravação, procurou-se manter de forma homogênea o intervalo de tempo entre a unidade a ser isolada e as adjacentes no logatomo para facilitar a análise e segmentação das unidades. Após uma avaliação de escuta informal, o processo de gravação foi repetido para as unidades que não foram pronunciadas de forma correta.

6.5 Segmentação das Unidades

Normalmente, o processo de segmentação de unidades acústicas pode ser realizado de duas formas: através de algoritmos baseados em HMMs (Modelos de Markov Escondidos) [152, 153], ou com o auxílio de ferramentas de visualização gráfica de sinais [10, 154, 155], como, por exemplo, *Sound Forge*. O uso de algoritmos baseados em

HMMs torna o processo de segmentação automatizado e, portanto, mais rápido, porém é necessário realizar a segmentação manual de determinada quantidade de *logatomos* do *corpus* considerado no momento, para treinar os HMMs; e alguns erros podem ocorrer durante o processo [152]. Por outro lado, o uso de ferramentas de visualização gráfica na segmentação requer um tempo bem maior do que o uso de algoritmos com HMMs, porém, é possível identificar cada segmento, não só de de forma gráfica como também através um teste de escuta, acionando-se uma função tipo *play* para a reprodução do sinal.

Neste trabalho foi usado o *Sound Forge* como ferramenta de visualização de forma de onda, como mostrado na Figura 6.2, e de espectro como mostrado na Figura 6.3. O *Sound Forge* é um programa de edição de áudio, usado em estúdios de gravação, e apresenta recursos necessários à análise da forma de onda e espectro, além de outras opções, dentre as quais se destacam: ajustes de duração, *pitch* e volume (amplitude), das unidades acústicas.

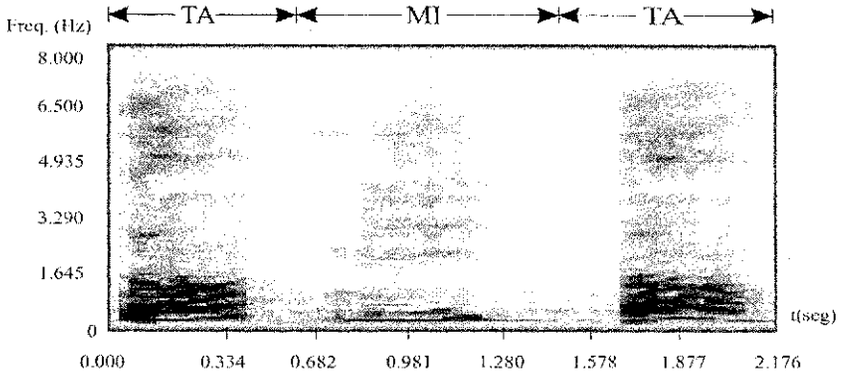

Figura 6.3: - Espectro do logatomo *tamita*.

Assim, foram segmentadas as unidades do dicionário, através da visualização da forma de onda e do espectro de freqüência, observando-se o início e o final de cada unidade, como também realizando-se um teste de escuta do sinal.

6.6 Rotulamento das Unidades

As unidades acústicas segmentadas (arquivos **.WAV**) foram rotuladas com o nome da sílaba ou demissílaba correspondente e com os valores de freqüência fundamental, para efeito de seleção no processamento de síntese. A freqüência fundamental foi determinada através de um detetor de *pitch*. Na prática são encontrados vários detetores

de *pitch*, os quais são implementados em função do algoritmo correspondente, como, por exemplo, os detetores desenvolvidos por Fechine, Quast, Wang e Deshmukh em [71, 95, 96, 97]. Neste trabalho, foi utilizado um detetor de *pitch* desenvolvido por Fechine em [71], pelo fato de ter sido disponibilizado pela autora e por ter sido usado com sucesso no seu trabalho de doutorado na UFCG. Nesse detetor, é utilizado inicialmente um estágio para identificar e separar os intervalos de silêncio, sons surdos e sonoros com base nos parâmetros temporais (Energia do Sinal, Taxa de Cruzamento por Zero, Coeficiente de Correlação Normalizado e Número Total de Picos) [73]. Posteriormente, é feita a estimativa da freqüência fundamental do segmento sonoro, utilizando um algoritmo específico e a função média de diferenças de amplitudes (AMDF - *Average Magnitude Difference Function*) [71].

6.7 Conclusão

O dicionário de unidades acústicas é um estágio fundamental para um sistema de conversão texto-fala usando a síntese concatenativa. O desenvolvimento de um dicionário desse tipo requer a execução de várias etapas, tais como: escolha do tipo de unidade, elaboração e gravação do *corpus* e segmentação das unidades [8, 10].

As unidades acústicas escolhidas para o dicionário, aqui apresentado, foram as sílabas e demissílabas, conforme relacionado na Tabela 6.3 e no Apêndice A. A escolha foi realizada, considerando-se a redução do número de unidades em relação ao uso exclusivo de sílabas e a inclusão de um número maior de coarticulações comparando-se aos difones. Assim, foram determinadas 1994 unidades, com base na identificação dos fonemas e nas possíveis combinações desses, dentro da Língua Portuguesa falada no Brasil .

Após a escolha e determinação das unidades, foi elaborado um *corpus* com 1994 *logatomos*. Para facilitar o processo de segmentação, foram usados *logatomos* com três sílabas, nos quais as unidades de interesse foram inseridas na sílaba central.

Na etapa seguinte, o *corpus* foi gravado e as unidades acústicas foram segmentadas e rotuladas. A gravação foi realizada em um estúdio profissional, visando reduzir a interferência de ruídos e obter uma fala de melhor qualidade. Para a segmentação foi utilizado o *Sound Forge* como ferramenta de visualização de forma de onda do sinal e teste de escuta.

Para o modelo prosódico apresentado no capítulo seguinte, torna-se necessário fazer

ajustes de *pitch* e duração em algumas unidades do dicionário aqui desenvolvido, dependendo da posição e da tonicidade das sílabas dentro de cada palavra. Teoricamente o dicionário teria seu número de unidades aumentado no mínimo de cinco vezes, pois no modelo são consideradas palavras com até cinco sílabas. Observa-se, porém, que as sílabas com o padrão V e CV, nas quais V é uma vogal oral aberta (**/a/, /é/, /i/, /ó/, /u/**), ocorrem com predominância na língua portuguesa, sendo **/i/ e /u/** bastante usadas em ditongos e tritongos, como, por exemplo, nas palavras: *história*, *roupa* e *Paraguai* [26, 151]. As sílabas terminadas por consoante, tipo (VC), são menos freqüentes e algumas provavelmente nunca ocorrem fora de determinada posição, como, por exemplo, a sílaba '**ab**' da palavra '*absoluto*'. Assim, tem-se, na prática, um número bem menor de variações prosódicas do que o número teórico de unidades acústicas idealizado, resultando em um dicionário com menos de 10.000 unidades.

Capítulo 7

Modelo Proposto e Resultados

O desenvolvimento de um modelo prosódico a ser aplicado em um conversor texto-fala implica no conhecimento do comportamento prosódico de pelo menos um falante, ou seja, no controle que o falante exerce sobre os parâmetros prosódicos (duração, *pitch* e energia) enquanto está falando [4]. Também é necessário o conhecimento fonético e fonológico associado às unidades acústicas a serem sintetizadas no conversor.

A aplicação de um modelo prosódico na síntese da fala pode ser realizada fazendo-se o ajuste de *pitch* e/ou duração do sinal, no estágio de síntese do conversor, utilizando-se um algoritmo tipo PSOLA [29, 43, 132, 133], ou então, criando-se antecipadamente as unidades acústicas com determinados valores de *pitch* e duração as quais são devidamente selecionadas, *a posteriori*, para a síntese da fala [27, 42, 43, 44, 86].

Assim, neste capítulo, é apresentada a proposta de um modelo para a obtenção automática da prosódia em um conversor texto-fala para a Língua Portuguesa falada no Brasil, baseado na tonicidade de palavras, em regras fonéticas e fonológicas e nas unidades acústicas obtidas no dicionário descrito no capítulo anterior. Também são apresentados os resultados obtidos com o modelo e um ambiente para análise da transcrição fonética, transcrição prosódica e avaliação qualitativa da fala sintetizada.

7.1 Geração Automática da Prosódia

Para a geração automática da prosódia em um sistema texto-fala, além do desenvolvimento de um modelo prosódico, é necessário definir as etapas de processamento lingüístico sobre o texto, a técnica de síntese utilizada e determinar as fronteiras prosódicas sobre as quais será aplicado o modelo [27, 42, 86].

O modelo prosódico proposto neste trabalho é baseado em regras e na tonicidade de palavras para determinar os contornos de entonação. A aplicação deste modelo em um sistema texto-fala, utilizando a síntese concatenativa, resulta em uma simplificação nas etapas de processamento do texto e de sinal (necessita de concatenação simples), como mostrado na Figura 4.7. Assim, é realizado um processamento lingüístico mais simples, contemplando os estágios de pré-processamento e transcrição fonética, os quais são implementados com base nos algoritmos de normalização e transcrição fonética, descritos nas Subseções 3.1.1 e 3.2.2 do Capítulo 3. A síntese concatenativa é realizada através da junção das unidades acústicas selecionadas no dicionário, de forma seqüencial, conforme os comandos provenientes do estágio de processamento prosódico. Esse tipo de síntese é bastante usado atualmente, considerando-se as vantagens de simplicidade, flexibilidade e sobretudo porque o processamento do sinal de fala é feito na própria forma de onda, mantendo-se assim as características originais desse sinal. A seleção das unidades é realizada através de uma estrutura de pesos atribuídos às sílabas tônicas, pretônicas e postônicas, considerando as curvas de entonação analisadas em um *corpus* de palavras, contendo as mais diversas combinações de fonemas para as sílabas, e de frases foneticamente balanceadas, para identificação do comportamento da duração e sobretudo do *pitch*, ao longo de palavras oxítonas, paroxítonas e proparoxítonas com até cinco sílabas. Relativamente a determinação das fronteiras prosódicas é usada a pontuação (.), (?) e (!), para a identificação e separação de frases em determinado texto e, espaços em branco para identificação e separação de palavras. Outros tipos de pontuação no interior de frases podem ser considerados, tais como: (,), (:), (...), (-), (( )) e ([ ]). Nas próximas seções são apresentados os algoritmos para a determinação da tonicidade de palavras e para a geração dos segmentos fonéticos, como também as diretrizes para o modelo de entonação e o rotulamento dos segmentos fonético-prosódicos.

7.2 Determinação da Tonicidade das Palavras

Para a obtenção da prosódia baseada na tonicidade das palavras é fundamental desenvolver um procedimento para marcar a acentuação no âmbito da palavra, pois quando os vocábulos são proferidos, torna-se evidente a presença de uma sílaba mais forte.

Em palavras dotadas de sinal gráfico (proparoxítonas, paroxítonas acentuadas e oxítonas acentuadas), como, por exemplo: parâmetro, rádio e você, pode ser mantido o sinal para identificar a sílaba acentuada com o som aberto ou fechado. Para as demais

palavras a tonicidade pode ser determinada através de regras, conforme descrito nas seguintes etapas [56]:

1ª Etapa:

Na primeira etapa são consideradas as palavras não acentuadas, com duas ou mais sílabas e que apresentam maior ênfase na última sílaba (oxítonas não acentuadas graficamente). Nesse caso, coloca-se um acento simbólico na última vogal das palavras terminadas em "l", "r" e "z". Para as palavras terminadas pelas vogais "i" e "u" o acento simbólico é colocado nessa vogal e para as terminadas em ditongo decrescente o acento é colocado na vogal do ditongo. Também as palavras terminadas em "im" e "um" são oxítonas. Essa etapa é exemplificada na Tabela 7.1.

Tabela 7.1: Exemplos de palavras com maior ênfase na última sílaba

Palavras	Fonemas Acentuados	Palavras	Fonemas Acentuados
dedal	ded-*á*-u	samurai	samur-*á*-i
ureter	uret-*é*-r	museu	muz-*ê*-u
rapaz	rap-*á*-z	pudim	pud-*í*-m
colibri	kolibr-*í*	atum	at-*ú*-m
caramuru	karamur-*ú*		

2ª Etapa:

Na segunda etapa são consideradas as palavras monossílabas não acentuadas e terminadas em "*i*" e "*u*". O acento simbólico é colocado nas referidas vogais e, quando fizerem parte de um ditongo decrescente, o acento é colocado na vogal do ditongo. Essa etapa é exemplificada na Tabela 7.2.

Tabela 7.2: Exemplos de monossílabos não acentuados

Palavras Monossílabas	Fonemas Acentuados
vi	v-*í*
tu	t-*ú*
pai	p-*á*-i
mau	m-*á*-u

Nos demais casos, as palavras monossílabas são consideradas átonas.

3ª Etapa:

Na terceira etapa são consideradas as palavras que não se enquadram nas regras anteriores, mais precisamente as paroxítonas não acentuadas graficamente. O acento simbólico é colocado na segunda vogal (direita para a esquerda), exceto as palavras terminadas em "*que*" e "*gue*" que são acentuadas na terceira vogal (direita para a esquerda). Essa etapa é exemplificada na Tabela 7.3.

Tabela 7.3: Exemplos de palavras paroxítonas sem acento gráfico

Palavras Paroxítonas	Fonemas Acentuados
sapato	sap-*á*-tu
cachorro	kax-*ô*-ru
almanaque	auman-*á*-qe
albergue	aub-*é*-rge

Assim, a identificação da tonicidade das palavras é realizada, conforme o fluxograma apresentado na Figura 7.1 e, implementada logo após o estágio de transcrição fonética, considerando a relação entre as sílabas e os segmentos fonéticos correspondentes.

Além disso, é observado em dicionários para o Português Brasileiro que a maioria das palavras têm no máximo cinco sílabas e, à medida que esse número vai crescendo, o número de palavras correspondentes vai diminuindo significativamente. Assim, para o modelo prosódico proposto foram consideradas efetivamente palavras com até cinco sílabas e realizada uma aproximação para as demais. Os modelos de palavras oxítonas, paroxítonas e proparoxítonas com até cinco sílabas, em função da posição da sílaba tônica e do número de sílabas nas palavras são apresentadas nas Tabelas 7.4, 7.5 e 7.6, respectivamente.

Tabela 7.4: Modelos de palavras oxítonas com até cinco sílabas

N.º de Sílabas	Palavras Oxítonas
Uma	Tônica ou Átona
Duas	(Pre1) + (Tônica)
Três	(Pre2) + (Pre1) + (Tônica)
Quatro	(Pre3) + (Pre2) + (Pre1) + (Tônica)
Cinco	(Pre4) + (Pre3) + (Pre2) + (Pre1) + (Tônica)

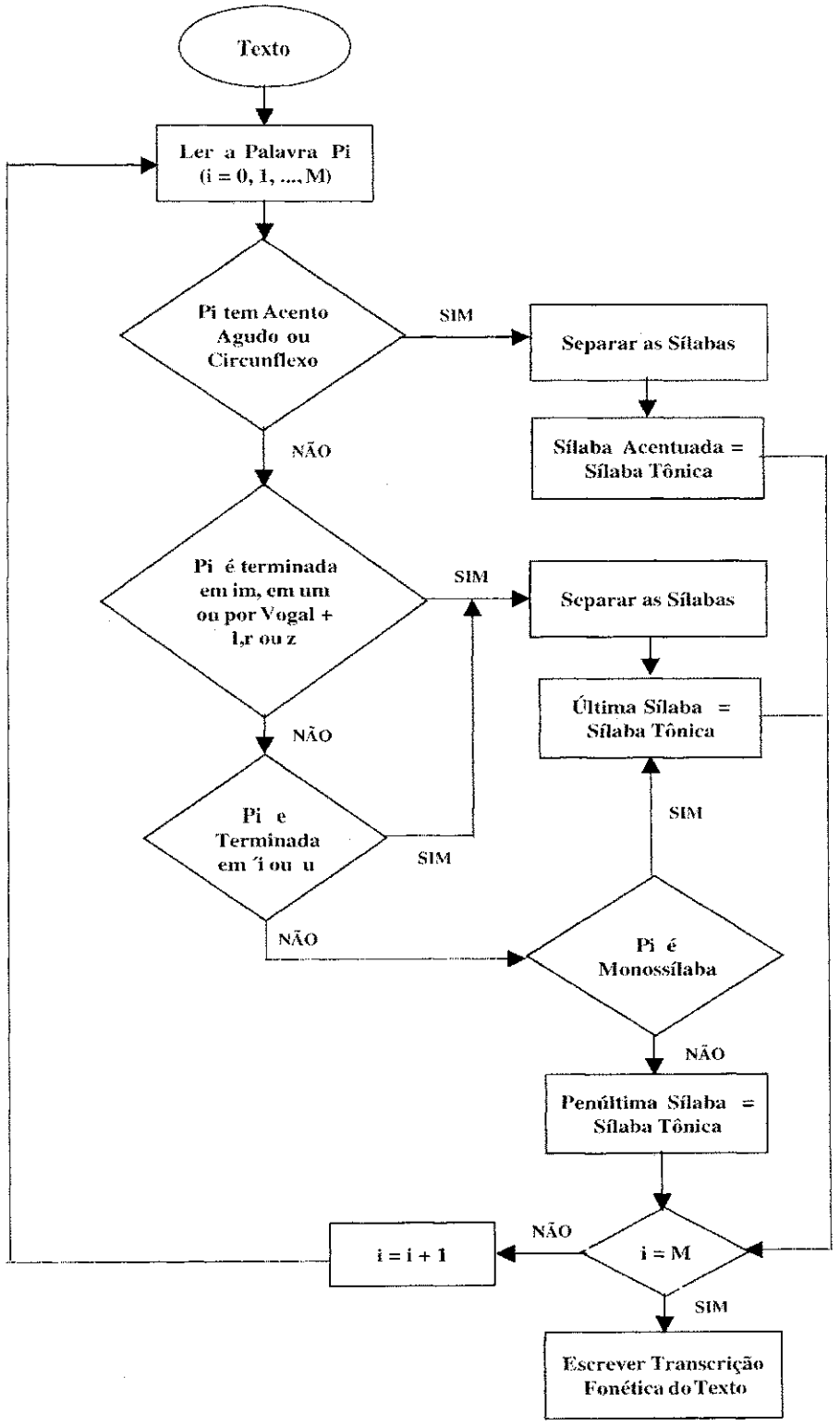

Figura 7.1: - Fluxograma para a identificação da tonicidade das palavras.

Tabela 7.5: Modelos de palavras paroxítonas com até cinco sílabas

N.º de Sílabas	Palavras Paroxítonas
Duas	(Tônica) + (Pos1)
Três	(Pre1) + (Tônica) + (Pos1)
Quatro	(Pre2) + (Pre1) + (Tônica) + (Pos1)
Cinco	(Pre3) + (Pre2) + (Pre1) + (Tônica) + (Pos1)

Tabela 7.6: Modelos de palavras proparoxítonas com até cinco sílabas

N.º de Sílabas	Palavras Proparoxítonas
Três	(Tônica) + (Pos1) + (Pos2)
Quatro	(Pre1) + (Tônica) + (Pos1) + (Pos2)
Cinco	(Pre2) + (Pre1) + (Tônica) + (Pos1) + (Pos2)

Nas Tabelas 7.4, 7.5 e 7.6, tem-se que:

- Tônica ou Átona significam sílaba tônica ou átona, respectivamente.
- Pre1, Pre2, Pre3 e Pre4 significam sílabas pretônicas nas posições 1, 2, 3 e 4, respectivamente. A numeração é crescente conforme o distanciamento da sílaba tônica.
- Pos1 e Pos2 significam sílabas postônicas nas posições 1 e 2, respectivamente. A numeração é crescente conforme o distanciamento da sílaba tônica.

Para as sílabas pretônicas a partir da posição 5 (Pre5, Pre6, ...) são considerados os mesmos valores de duração e *pitch* de sílabas semelhantes e correspondentes às pretônicas na posição 4.

7.3 Geração dos Segmentos Fonéticos

A geração dos segmentos fonéticos é realizada a partir das informações recebidas do estágio de transcrição fonética e a partir dos padrões silábico-fonéticos das palavras. Para tal, são separados os segmentos fonéticos correspondentes às sílabas das palavras e às unidades acústicas do dicionário. O processo de separação é realizado em quatro etapas, conforme descrito a seguir:

1. São separados os grupos de fonemas compostos por CV (consoante + vogal). Nessa etapa, os fonemas que não compõem grupos CV são separados isoladamente.

2. São verificadas se as consoantes que não foram incluídas nos grupos de fonemas (CV) irão permanecer separadas ou deverão ser unidas ao grupo anterior ou posterior.

3. São verificadas se as vogais que não foram incluídas nos grupos de fonemas (CV) irão permanecer separadas ou deverão ser unidas ao grupo anterior ou posterior.

4. São verificados se os grupos de fonemas terminados por vogais (CV, CVV, etc.) são nasalizados pelo grupo seguinte.

A escolha do padrão (CV) aplicado na primeira etapa de segmentação fonética prende-se ao fato de que a maioria das sílabas da Língua Portuguesa tem essa estrutura reduzindo-se, assim, o processo de segmentação nas etapas seguintes.

O processo de geração de segmentos, pode ser ilustrado através de exemplos, com as palavras: *substantivo* e *banana*. A palavra *substantivo* assume as seguintes formas durante o processo: - transcrição fonética: substantivu; - 1ª Etapa: su/b/s/ta/m/ti/vu; - 2ª Etapa: su/bs/ta/m/ti/vu; - 3ª Etapa: subs/tam/ti/vu; - 4ª Etapa: subs/tam/ti/vu. A palavra *banana* assume as seguintes formas: - transcrição fonética: banana; - 1ª Etapa: ba/na/na; - 2ª Etapa: ba/na/na; - 3ª Etapa: ba/na/na; - 4ª Etapa: bam/nam/na.

Portanto, foi determinado um conjunto de regras para a segmentação fonética, apresentado no Apêndice B, que foram implementadas no estágio de processamento prosódico.

7.4 Modelo Proposto

O modelo de prosódia proposto neste trabalho, é baseado em regras e na tonicidade de palavras para determinar os contornos de entonação. Assim, são estabelecidas regras para a determinação da freqüência fundamental das unidades acústicas correspondentes às sílabas, em função do contexto em que se inserem. São consideradas as seguintes características em nível de palavra:

1. Número de sílabas (monossílabas, dissílabas, trissílabas e polissílabas com até cinco sílabas);

2. Tonicidade: sílaba tônica e sílaba átona (postônica e pretônica);

3. Posição da sílaba tônica (oxítonas, paroxítonas, proparoxítonas);

4. Número de caracteres do segmento fonético correspondente a cada sílaba;

5. Estrutura do segmento fonético correspondente a cada sílaba no que se refere à combinação das vogais com as consoantes (CV, CVC, CVV,...).

7.4.1 Padrões de *Pitch* das Unidades

Para a determinação dos padrões de *pitch*, das unidades acústicas a serem usadas no modelo, foi inicialmente elaborado e gravado um *corpus* constituído por palavras oxítonas, paroxítonas e proparoxítonas com até cinco sílabas, conforme apresentado no Apêndice C, contendo as mais diversas combinações de fonemas para as sílabas. Para manter a neutralidade das palavras e dar um sentido ao que estava sendo lido na etapa de gravação, foram usadas frases veículo do tipo '*digo* (**palavra**) *baixinho*' [146]. Também foi gravado um *corpus* de 200 frases foneticamente balanceadas elaboradas por Alcaim *et al.* [68]. As unidades acústicas correspondentes às sílabas foram segmentadas utilizando-se o editor de áudio *Sound Forge* e determinados os valores de *pitch*, usando-se o detetor de *pitch* desenvolvido por Fechine em [71]. Analisando-se os valores de *pitch* das unidades acústicas, correspondentes às sílabas do *corpus* estabelecido, são encontrados alguns aspectos importantes, os quais foram incorporados no modelo, tais como:

- em palavras oxítonas, as sílabas pretônicas apresentam valores de F_0 inferiores aos das sílabas tônicas. Um exemplo é observado na palavra *abafar*, onde os valores de F_0 medidos para as sílabas pretônicas **a** e **ba** são 135 e 140 Hz, respectivamente, enquanto que para a sílaba tônica **far** tem-se um valor de 144 Hz. Essa tendência ocorre para a maioria das palavras oxítonas relacionadas no *corpus* apresentado no Apêndice C.

- em palavras paroxítonas, as sílabas postônicas têm valor de F_0 superior a F_0 das tônicas, e as sílabas pretônicas apresentam valores de F_0 inferiores aos das

sílabas tônicas. Um exemplo é observado na palavra *discurso*, onde os valores de F_0 medidos para as sílabas: pretônica **dis**, tônica **cur** e postônica **so**, são 110, 122 e 130 Hz, respectivamente. Essa tendência ocorre para a maioria das palavras paroxítonas relacionadas no *corpus* apresentado no Apêndice C.

- em palavras proparoxítonas polissílabas, tem-se um crescimento de F_0 das pretônicas para as tônicas e das tônicas para as postônicas. Um exemplo é observado na palavra *matemática*, onde os valores de F_0 medidos para as sílabas: pretônicas **ma** e **te**, para a tônica **má** e para as postônicas **ti** e **ka**, são 114, 128, 131, 137 e 138 Hz, respectivamente. Essa tendência ocorre para a maioria das palavras proparoxítonas relacionadas no *corpus* apresentado no Apêndice C.

- as sílabas das palavras oxítonas, paroxítonas e proparoxítonas tem valores de F_0 variando entre 110 e 160 Hz (características de um locutor masculino, usado na gravação), e uma diferença máxima de 30 Hz entre uma sílaba e a sua anterior.

Os três primeiros aspectos também foram observados em estudos realizados por Madureira em [63], para trissílabos.

Portanto, a partir das observações realizadas, foi estabelecida uma relação de freqüência fundamental entre as sílabas tônicas, pretônicas e postônicas de palavras com até cinco sílabas, conforme mostrado na Tabela 7.7, a serem usadas no modelo prosódico.

Tabela 7.7: Modelo geral de *pitch* das unidades correspondentes às sílabas

Relação de Freqüência Fundamental entre Sílabas
(F_0 da sílaba tônica) < (F_0 da sílaba Pos_1) < ($1,10.F_0$ da sílaba tônica)
(F_0 da sílaba tônica) < (F_0 da sílaba Pos_2) < ($1,15.F_0$ da sílaba tônica)
($0,96.F_0$ da sílaba tônica) < (F_0 da sílaba Pre_1) < (F_0 da sílaba tônica)
($0,92.F_0$ da sílaba tônica) < (F_0 da sílaba Pre_2) < (F_0 da sílaba tônica)
($0,88.F_0$ da sílaba tônica) < (F_0 da sílaba Pre_3) < (F_0 da sílaba tônica)
($0,84.F_0$ da sílaba tônica) < (F_0 da sílaba Pre_4) < (F_0 da sílaba tônica)

A incorporação dos valores de *pitch*, para as unidades acústicas usadas no modelo, é realizada através ajustes de *pitch* feitos nessas unidades, usando-se o editor de áudio *Sound Forge*.

7.4.2 Duração das Unidades

Apesar do modelo prosódico proposto ser baseado na entonação das palavras, de modo que a seleção de unidades acústicas no dicionário é realizada com base em valores de *pitch*, determinadas características de duração relativas à tonicidade também são incorporadas nessas unidades. Assim, a partir de uma análise nos valores de duração das unidades acústicas, correspondentes às sílabas do *corpus* de palavras e frases estabelecido anteriormente, são encontrados alguns aspectos importantes, tais como:

- sílabas tônicas têm uma tendência de possuir duração maior do que sílabas pretônicas e sílabas postônicas. Isso pode ser observado no exemplo da forma de onda da palavra *benéfico*, apresentada na Figura 7.2. Neste caso a sílaba tônica *né* tem duração de 254 milisegundos, enquanto que as postônicas *fi* e *co* têm durações de 148 e 198 milisegundos, e a pretônica *be* tem duração de 138 milisegundos. Esse fato é corroborado em estudos realizados por Massini-Cagliari em [61], que conclui que a maioria das sílabas tônicas tem duração maior que as sílabas átonas, para um determinado conjunto de palavras. Em um total de 626 palavras analisadas, observou-se que mais de 80% mantêm-se nessa regra.

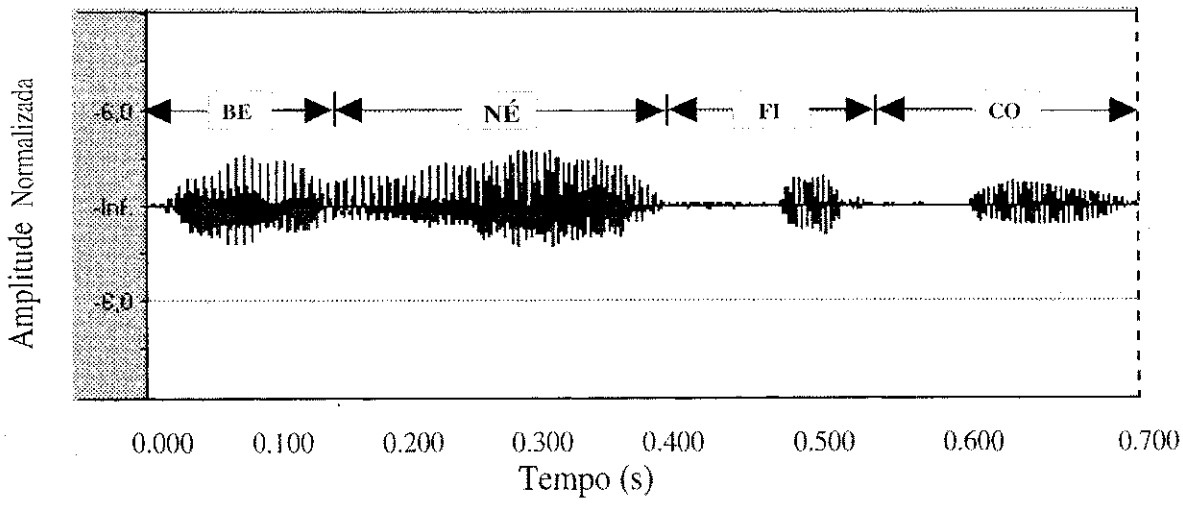

Figura 7.2: - Forma de onda no tempo da palavra *benéfico*.

- sílabas tônicas com vogais abertas (tipo /a/) têm uma tendência de possuir uma duração maior do que tônicas com vogais fechadas (tipo /ê/). Isso pode ser observado, por exemplo, na forma de onda das palavras *página* e *pêsames*, mostradas

na Figura 7.3. A sílaba *pá* da palavra *página* tem uma duração de 204 milisegundos e a sílaba *pê* da palavra *pêsames* tem uma duração de 182 milisegundos. Esse fenômeno também ocorre com sílabas postônicas e pretônicas em posições similares nas palavras, e é confirmado em estudos realizados por Massini-Cagliari em [61], a qual conclui que vogais abertas têm uma duração maior do que as fechadas.

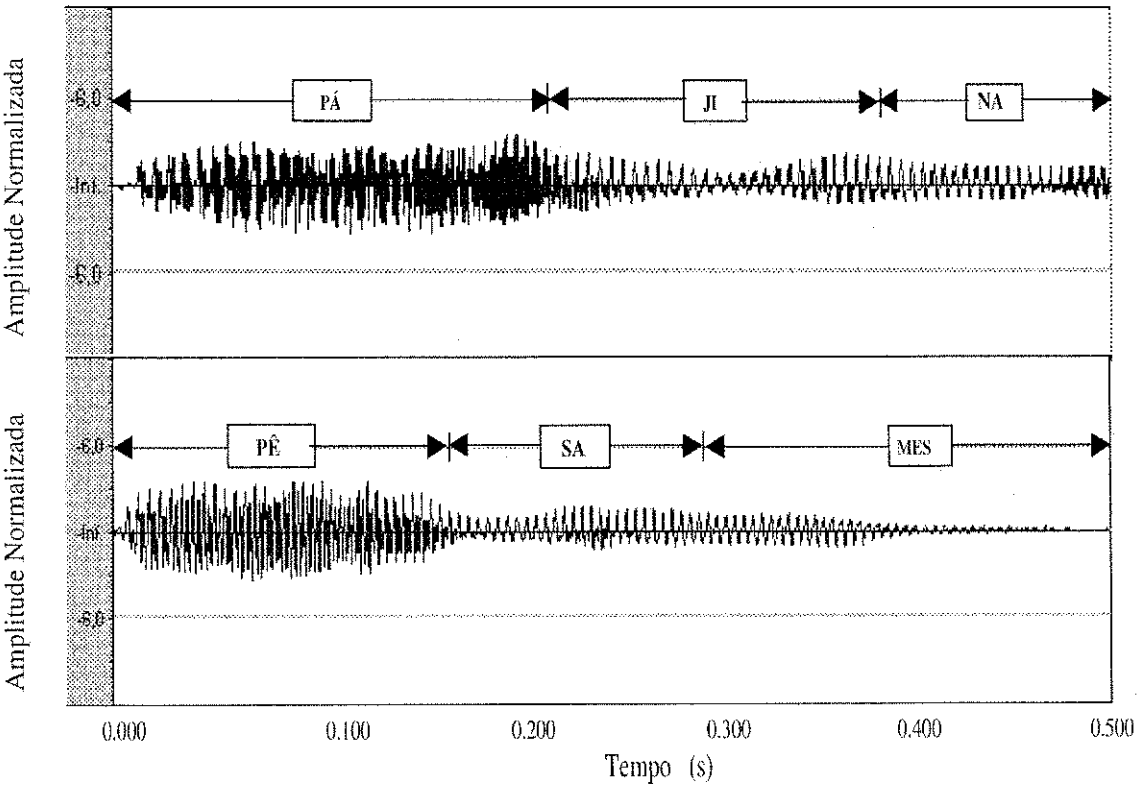

Figura 7.3: - Formas de onda no tempo das palavras *página* e *pêsames*.

- sílabas com vogais nasais ou nasalisadas (tipo /am/) têm uma tendência de possuir duração maior do que sílabas com vogais orais (tipo /a/). Isso pode ser observado, por exemplo, nas formas de onda das palavras *abater* e *abanar*, apresentadas na Figura 7.4. A sílaba *ba* da palavra *abater* tem uma duração de 130 milisegundos e a sílaba *ba* da palavra *abanar* é nasalisada pela sílaba *nar*, e tem uma duração de 175 milisegundos. Esse fato também é corroborado em estudos realizados por Massini-Cagliari em [61] e por Moraes em [156].

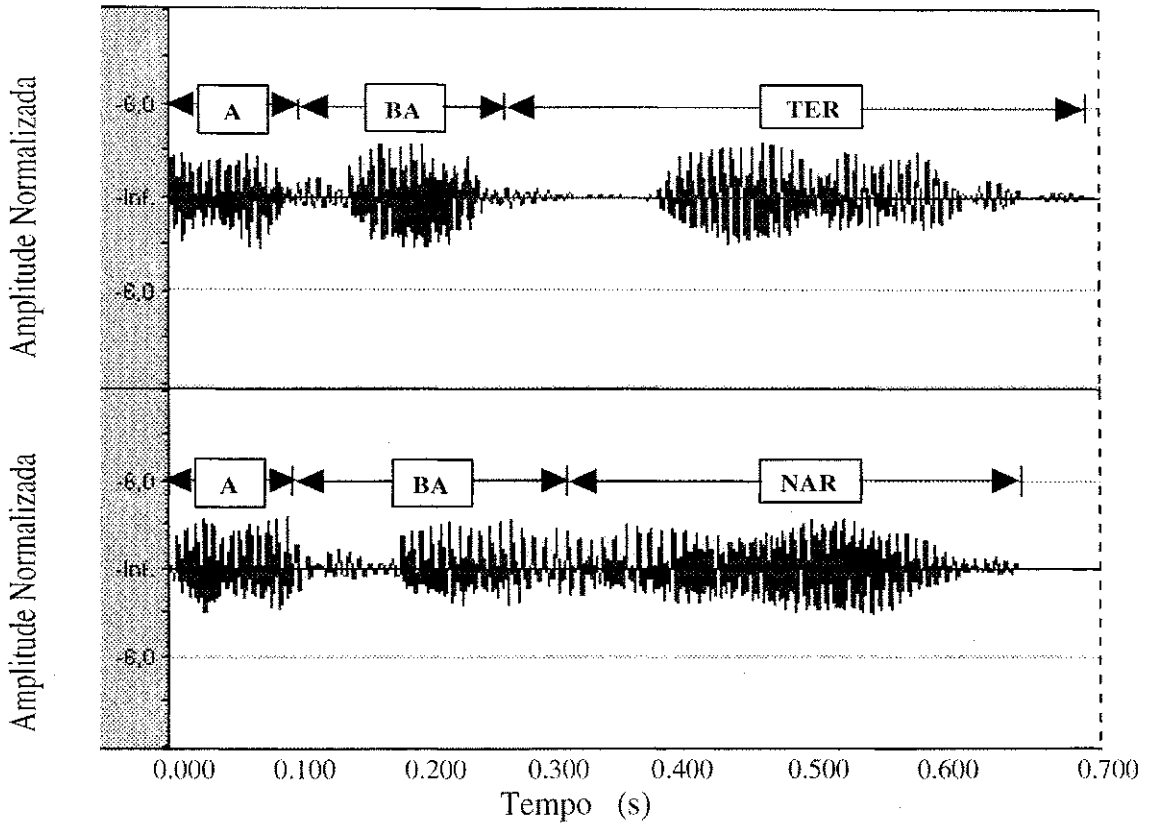

Figura 7.4: - Formas de onda no tempo das palavras *abater* e *abanar*.

Portanto, além de padrões de *pitch* também são criados padrões de duração, para as unidades acústicas do dicionário, para atender a tonicidade de sílabas em palavras oxítonas, paroxítonas e proparoxítonas com até cinco sílabas.

7.4.3 Seleção das Unidades

Para a seleção das unidades acústicas no dicionário, os segmentos fonéticos obtidos na etapa inicial do processamento prosódico são classificados e rotulados com um nome específico e com um valor de F_0, função do contexto em que se inserem, conforme a unidade acústica correspondente no dicionário. Como, por exemplo, a unidade [ba] passa a ter a forma 'ba 127', onde 127 Hz corresponde ao *pitch* dessa unidade. Para a classificação e rotulamento dos segmentos é usada uma estrutura de pesos baseada na tonicidade das palavras, conforme apresentado na Tabela 7.8, na qual Pre1, Pre2, Pre3 e Pre4 significam sílabas pretônicas nas posições 1, 2, 3 e 4, e Pos1 e Pos2 significam

sílabas postônicas nas posições 1 e 2, respectivamente.

Tabela 7.8: Valores arbitrados aos pesos das sílabas no modelo prosódico

Sílaba	Pre4	Pre3	Pre2	Pre1	Tônica	Pos1	Pos2
Peso	1	2	3	4	5	6	7

A estrutura de pesos é implementada conforme as seguintes etapas:

1. todos os segmentos da palavra têm peso inicial igual a 1, independente da acentuação, para que nenhum segmento fique sem classificação;

2. é feita a classificação dos segmentos segundo a estrutura silábica (V, CV, etc.). Nessa etapa se a vogal do segmento for acentuada, ele recebe peso 5;

3. caso os segmentos da palavra não sejam acentuados (palavra oxítona ou proparoxítona não acentuada), é determinada qual é a sílaba tônica , com base nas regras de tonicidade apresentadas na Seção 7.2 e, assim, a sílaba determinada como tônica recebe peso 5;

4. após a determinação da sílaba tônica com o peso 5, são determinadas as sílabas pretônicas e postônicas com os pesos especificados na Tabela 7.8.

5. é feito um mapeamento dos pesos sobre os padrões de F_0 de cada segmento, de modo que cada segmento é rotulado com um valor de *pitch* específico.

Um exemplo de classificação para as sílabas da palavra *paroxítona* com o valor de F_0 (Hz) correspondente é apresentado na Tabela 7.9. Após a classificação e rotulamento tem-se a seguinte transcrição fonético-prosódica: [pa 117 / ro 121 / xí 127 / tom 132 / na 135].

Tabela 7.9: Classificação das sílabas da palavra *paroxítona* com o valor de F_0 (Hz) correspondente

Sílaba	pa	ro	xí	to	na
Peso	3	4	5	6	7
F_0 (Hz)	117	121	127	132	135

Portanto, a partir da implementação do modelo prosódico apresentado neste capítulo tem-se a geração automática da prosódia de palavras que pode ser inserida em um conversor texto-fala concatenativo para a Língua Portuguesa falada no Brasil.

7.5 Ambiente de Testes

Para análise da transcrição fonética, transcrição prosódica e sobretudo para a avaliação da fala sintetizada utilizando-se o modelo prosódico proposto foi implementado um ambiente utilizando compilador C++ *Builder* 5 sob *Windows* 98, conforme mostrado na Figura 7.5.

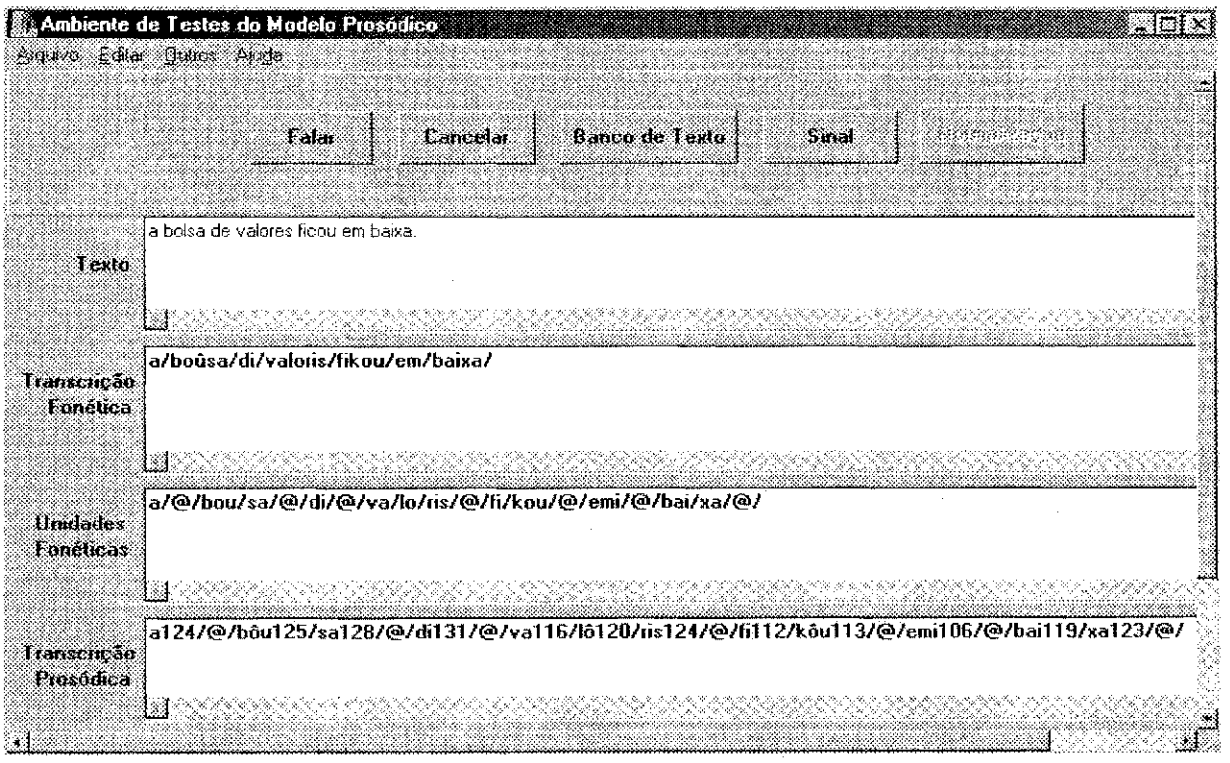

Figura 7.5: Ambiente de testes para o modelo prosódico.

No ambiente da Figura 7.5 tem-se quatro janelas:

- na Janela 'Texto', é apresentado um texto, que foi digitado ou selecionado em um banco de dados ativado pelo botão 'Banco de Texto';
- na Janela 'Transcrição Fonética', é apresentado o resultado da transcrição fonética

do texto, após a separação das frases pela pontuação e das palavras pelos espaços em branco;

- na Janela 'Unidades Fonéticas', é apresentado o resultado da separação dos segmentos fonéticos, a partir da transcrição fonética do texto, conforme as considerações feitas na Seção 7.3;

- na Janela 'Transcrição Prosódica', os segmentos fonéticos são rotulados com os valores de F_0 (Hz) função do contexto em que se inserem e do modelo prosódico proposto.

Além disso:

- Pode-se processar o texto automaticamente e escutar a fala sintetizada correspondente, clicando o botão 'Falar'.

- Pode-se eliminar um texto digitado ou copiado na tela 'Texto', clicando-se o botão 'Cancelar'.

- Pode-se acrescentar ou retirar textos a serem lidos automaticamente a partir de um banco de dados, clicando o botão 'Banco de Texto'.

O banco de dados foi estruturado com o *Database Desktop*, incluído no C++ *Builder* 5, de forma que armazene um texto em cada entrada. Ele é acionado clicando-se o botão 'Banco de Texto' no ambiente da Figura 7.5. Um exemplo do ambiente do banco de dados com seis frases é apresentado na Figura 7.6.

No ambiente da Figura 7.6, cada texto é digitado em uma linha e posteriormente acrescentado ao banco de dados clicando-se os botões (+) e ($\surd$). Para excluir o texto do banco deve-se selecionar a linha correspondente e clicar o botão (-). Logo após surge uma janela com uma mensagem de advertência questionando se realmente deseja excluir a linha. Em caso positivo clica-se o botão 'OK' e em caso negativo clica-se o botão 'Cancel'. Existem outros botões na parte superior desta tela que podem conduzir o usuário à primeira ou à última linha que consta no banco.

Para testes realizados com os ambientes mostrados na Figuras 7.5 e 7.6 foram usadas frases foneticamente balanceadas obtidas de um trabalho anterior desenvolvido por Alcaim *et al* em [68].

Conclui-se então que o ambiente de testes é de extrema importância para a análise do processamento do texto e da fala sintetizada, como também para a validação

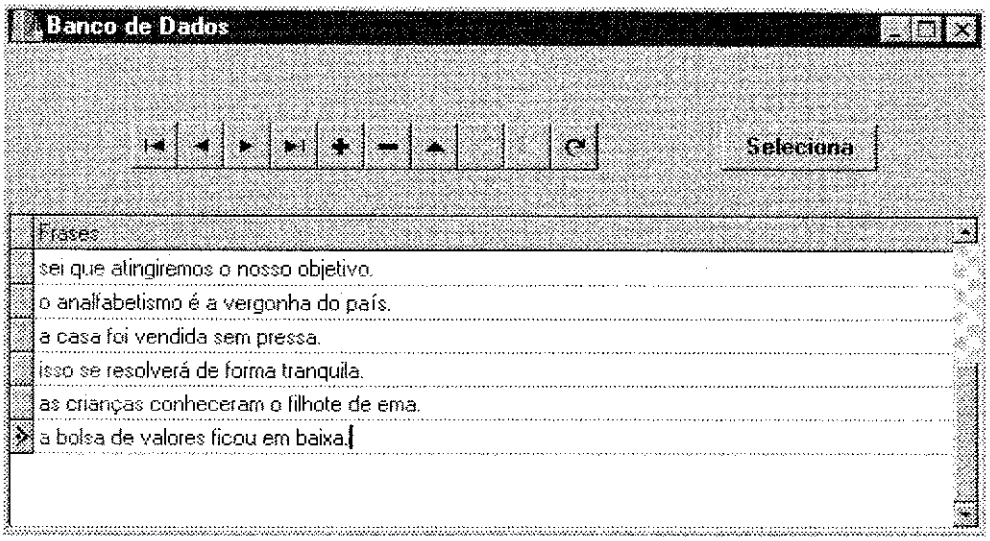

Figura 7.6: Interface para armazenar o texto no banco de dados.

da modelagem prosódica, podendo ser utilizado em sistemas similares de conversão texto-fala.

7.6 Testes Realizados com Palavras

Para verificar a eficiência do modelo prosódico proposto, foram realizados testes com o objetivo de analisar a tonicidade das palavras, em termos de freqüência fundamental e duração. Assim, foi observado inicialmente o efeito das sílabas postônicas em relação às tônicas para um *corpus* de palavras paroxítonas dissílabas, paroxítonas trissílabas, proparoxítonas trissílabas, apresentadas nas tabelas do Apêndice C. Posteriormente, foi observado o efeito das pretônicas em relação às tônicas para palavras paroxítonas trissílabas, proparoxítonas tretassílabas e pentasílabas e oxítonas pentassílabas.

Através das análises realizadas foram observadas as tendências do comportamento da fala em termos de ritmo e entonação e deduzidos os valores de duração e freqüência fundamental usados no modelo prosódico.

Na Tabela 7.10 são apresentados os valores de duração e freqüência fundamental dos segmentos correspondentes às sílabas da palavra '*fonética*' original, e sintetizada pelo conversor texto-fala, como também a relação entre esses parâmetros.

Assim, tem-se que:

- D_0Nat é a duração do segmento fonético correspondente à sílaba da palavra

Tabela 7.10: Duração e freqüência fundamental dos segmentos correspondentes às sílabas da palavra *'fonética'*

Segmento	$D_0Nat(ms)$	$D_0Sint(ms)$	K_{D_0}	$F_0Nat(Hz)$	$F_0Sint(Hz)$	K_{F_0}
/fom/	165	247	1,50	117	121	1,03
/né/	176	320	1,82	135	138	1,02
/ti/	128	257	2,0	152	140	0,92
/ka/	163	246	1,50	150	147	0,98

natural;

- D_0Sint é a duração do segmento fonético correspondente à unidade do dicionário;
- $FoNat$ é a freqüência fundamental do segmento fonético correspondente à sílaba da palavra natural;
- $FoSint$ é a freqüência fundamental do segmento fonético correspondente à unidade do dicionário;
- K_{D_0} é a relação entre a duração do segmento correspondente à unidade do dicionário e a duração do segmento correspondente à sílaba da palavra natural, expresso pela equação:

$$K_{D_0} = \frac{D_0Sint}{D_0Nat} \tag{7.1}$$

- K_{F_0} é a relação entre a freqüência fundamental do segmento correspondente à unidade do dicionário e a freqüência fundamental do segmento correspondente à sílaba da palavra natural, e pode ser expresso pela equação:

$$K_{F_0} = \frac{F_0Sint}{F_0Nat} \tag{7.2}$$

Observa-se na Tabela 7.10 que K_{F_0} fica em torno de 1 (100%). Assim, os segmentos fonéticos do dicionário e os segmentos correspondentes da fala natural têm valores de F_0 bastante próximos, apesar da seqüência lógica do modelo de $FoSint$ não ser exatamente igual ao de $FoNat$. Por outro lado, K_{D_0} tem valores variando de 1,5 à 2,0 e, assim, alguns segmentos do dicionário têm valores de duração quase duas vezes o valor dos segmentos correspondentes nas palavras naturais. Esse fato deve-se inicialmente à forma que a gravação das unidades, definidas no Capítulo 6 e no

Apêndice A, foi realizada. Ocorre uma perda considerável na informação transmitida pela fala sintetizada caso sejam realizados grandes ajustes de duração nos segmentos, principalmente, abaixo de 50% do seu valor original, o que vem a corroborar com o modelo de Klatt para duração, descrito no Capítulo 4. Verifica-se também que a duração de cada segmento fonético depende não só da tonicidade como também do número e da combinação de fonemas que o compõem. Assim, a diferença na duração não constitui um efeito crítico na palavra produzida pelo conversor texto-fala quando comparada à palavra natural, o que pode ser observado através de avaliação subjetiva.

Normalmente, a duração é associada ao ritmo (repetição dos sons em intervalos regulares), e a freqüência fundamental é associada à entonação (melodia). Apesar desses dois parâmetros serem importantes na produção da fala, tem-se observado, na maioria dos trabalhos desenvolvidos relativos ao assunto, que o modelo de freqüência fundamental tem prevalecido sobre o modelo de duração, considerando-se que na prática uma variação de duração em um segmento fonético não produz grande alteração na informação produzida pela fala quanto à variação na mesma proporção da freqüência fundamental. Assim, os valores de freqüência fundamental e duração para os segmentos fonéticos do dicionário são aceitáveis na síntese da fala desde que estejam dentro de certos limites, quando comparados com a fala natural e respeitando os critérios adotados no modelo prosódico.

Nas Figuras 7.7 e 7.8 são apresentadas as formas de onda correspodentes à palavra '*fonética*' original e sintetizada. Observa-se que a forma de onda da palavra '*fonética*' sintetizada tem uma duração maior do que a da palavra '*fonética*' natural, pelas razões apresentadas anteriormente e pelo efeito de coarticulação na composição da palavra. Também os níveis de amplitude são diferentes devido à forma de gravação e equalização das unidades do dicionário e das palavras.

A Tabela 7.10 juntamente com as formas de onda das Figuras 7.7 e 7.8 representam uma amostra que pode ser usada na avaliação do modelo prosódico, pois tem-se uma palavra com quatro sílabas incluindo a tônica, duas postônicas e uma pretônica, em que é possível estabelecer uma relação de freqüência e duração entre segmentos fonéticos correspondentes.

Para se obter uma amostra mais significativa da fala produzida no sistema de conversão texto-fala e avaliar o modelo prosódico proposto, foi utilizado um *corpus* de 20 frases conforme descrito na seção seguinte.

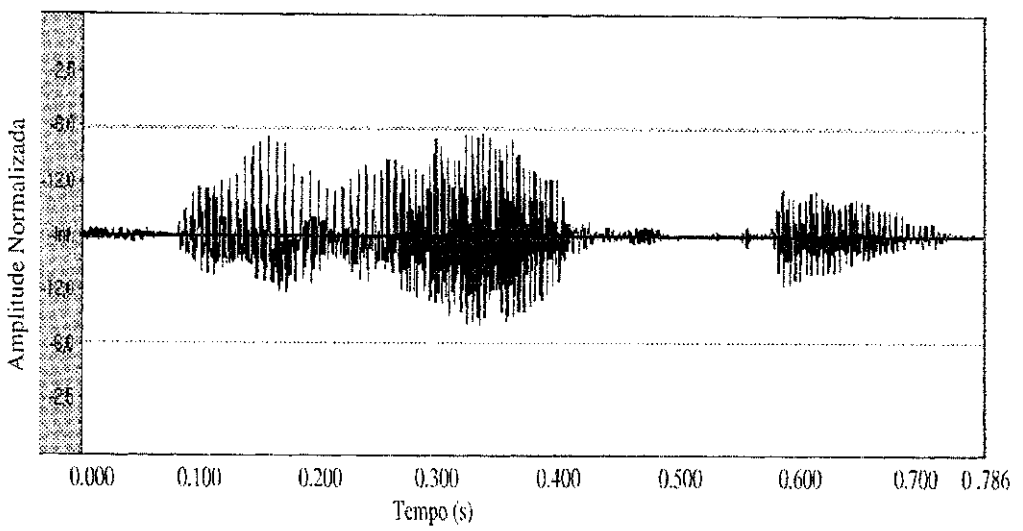

Figura 7.7: Forma de onda da palavra *fonética* produzida de forma natural.

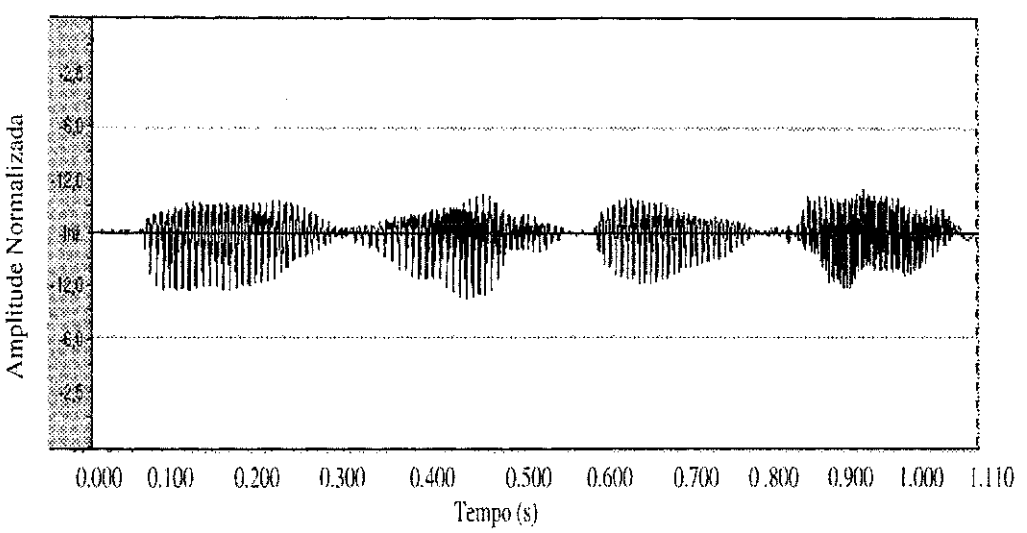

Figura 7.8: Forma de onda da palavra *fonética* produzida de forma sintetizada.

7.7 Avaliação do Modelo com Frases

O objetivo principal da avaliação da modelagem prosódica em um conversor texto-fala é determinar a diferença entre a fala natural e a fala sintetizada, em termos qualitativos através de testes de escuta e/ou em termos quantitativos através dos valores de duração e principalmente de freqüência fundamental.

Duas formas são comuns na avaliação da qualidade do sinal da fala: avaliação subjetiva e avaliação objetiva [47].

Na avaliação subjetiva, determinados textos, frases ou palavras representativos de um *corpus* da fala são submetidos a um conversor texto-fala e um grupo de ouvintes classificam a qualidade da fala produzida pelo conversor. Para quantificar tal informação pode ser usada a escala de opinião média MOS (*Mean Opinion Score*) [87], com os seguintes valores de classificação: 5 - Ótimo; 4 - Bom; 3 - Regular; 2 - Ruim; 1 - Péssimo.

Na avaliação objetiva, a fala sintetizada é comparada com a original de forma quantitativa, ou seja, são observados parâmetros tais como: duração e/ou freqüência fundamental e tiradas conclusões sobre a semelhança entre os valores nos dois casos. Um método bastante usado nessa avaliação é o da correlação linear de Pearson [157, 158]. Nesse método, um coeficiente de correlação próximo de um indica, por exemplo, que duas curvas de *pitch* estão bastante próximas (expressão natural e a sintetizada), enquanto que um coeficiente próximo de zero indica que essas duas curvas são bastante diferentes [47].

Neste trabalho foram realizados testes MOS para avaliar a qualidade da fala sintetizada em um conversor texto-fala concatenativo, usando-se o modelo de prosódia apresentado no Capítulo 7. Para tal, foi estabelecido um *corpus* de vinte frases a partir das vinte listas de dez frases foneticamente balanceadas desenvolvidas por Alcain *et al* em [68]. As vinte frases juntamente com os segmentos fonéticos e os valores correspondentes de freqüência fundamental e duração são relacionados no Apêndice E. No *corpus* das vinte frases são observadas as seguintes características:

- Estão contidos todos os fonemas definidos no alfabeto AFLAPS apresentado na Tabela 6.2 do Capítulo 6.

- São incluidos segmentos fonéticos importantes para análise da inteligibilidade dos sons como: ficativos sonoros (exemplo: ji, jé, vu, va, vi, za, etc.), fricativos surdos

(exemplo: fa, tis, is, foi, sa, su, si, sas, sê, fi, ris, xa, etc.), nasalisadas (exemplo: gom, vem, semi, tram, tim, rem, am, ramo, em, etc.), encontros consonantais (exemplo: pré, kri, prê, brou, tru, grau, kro, prim, kla, etc) e ditongos (exemplo: nau, kui, sei, bou, kou, bai, foi, mui, kau, brou, zei, rei, etc.). Na síntese da fala as consoantes nasais, fricativas e encontros consonantais são menos inteligíveis do que as vogais [87].

- Tem-se um total de 500 fones, sendo a maior ocorrência para [a] - 63 vezes, [i] - 48 vezes, [u] - 41 vezes, [s] - 34 vezes, [m] - 33 vezes e /r/ - 37 vezes. Comparando-se esses valores ao número de fones apresentado na Tabela 2 do trabalho de Alcaim *et al.* em [68], no qual para um total de 10147 fones tem-se uma ocorrência de [a] - 1313 vezes, [i] - 870 vezes, [u] - 557 vezes, [s] - 424 vezes, [m] - 418 vezes e [r] - 363 vezes, obtém-se uma correlação de 0,983, de modo que os valores obtidos nos dois casos estão fortemente correlacionados.

- Tem-se um total de 83 palavras com o número de silabas variado (de uma a seis), com as três classes de tonicidade (oxítonas, paroxítonas e proparoxítonas) e com silabas pretônicas nas posições 1, 2, 3 e 4, e postônicas nas posições 1 e 2, conforme mostrado na Tabela 7.8.

Assim, as vinte frases, na forma de texto, foram convertidas em fala utilizando-se o ambiente apresentado na Seção 7.5, e os arquivos **.WAV** resultantes foram submetidos à escuta de 40 ouvintes, sendo 8 deficientes visuais que tem usado outros sistemas como o DOSVOX desenvolvido pela UFRJ (*www.nce.ufrj.br*) [12, 13] e o Virtual Vision desenvolvido pela Micropower (*www.micropower.com.br*) [14], 8 professores universitários pós-graduados, 2 professores do ensino fundamental e 22 estudantes universitários, com idades acima de 18 anos. Cada ouvinte atribuiu uma nota a cada frase escutada, conforme a escala e os critérios de classificação definidos na Tabela 7.11.

Tabela 7.11: Escalas usadas na avaliação MOS das 20 frases

Escala	Classificação	Critérios de Avaliação
5	Excelente	**sem falhas notáveis** (frase com palavras inteligíveis e com naturalidade)
4	Bom	**com poucas falhas** (frase com no máximo duas palavras sem inteligibilidade e sem naturalidade)
3	Regular	**com determinada quantidade de falhas, porém aceitável** (frase com metade de palavras com pouca inteligibilidade e sem naturalidade)
2	Pobre	**com muitas falhas** (frase com a maioria das palavras sem inteligibilidade e sem naturalidade)
1	Péssimo	**alta degradação da fala** (frase em que todas as palavras não foram entendidas)

7.8 Resultados Obtidos com Frases

Na Tabela 7.12, tem-se os resultados obtidos com os testes MOS conforme os critérios descritos na seção anterior. Na parte superior da tabela tem-se a freqüência de ocorrência do escore na avaliação das 20 frases por 40 ouvintes, para o qual são computados os resultados por escore e por frase (E_i Total), como também no conjunto das 20 frases (MOS Total). Na parte inferior da tabela tem-se a freqüência relativa (em porcentagem - R_i % Total) de cada escore, perfazendo um total de 100% para cada frase individual como também um total de 100% para o conjunto das 20 frases (MOS % Total).

Observa-se na Tabela 7.12 que no escore da primeira frase (E_1) a classificação bom e regular prevalece sobre excelente em oposição ao escore da última frase (E_{20}). Esse fato ocorre não só pela existência de alguma palavra sem inteligibibilidade ou naturalidade na primeira frase, como também pelo fato de ser a primeira a ser submetida aos testes.

O teste MOS tem a vantagem de realizar a avaliação da fala de forma global, de ser simples e de serem identificados os segmentos fonéticos que precisam ser substituídos, devido a péssima qualidade.

Tabela 7.12: Freqüência de ocorrência e freqüência relativa das vinte frases usadas nos testes MOS.

- Escala -	Freqüência de Ocorrência do Escore na Avaliação das 20 Frases por 40 Ouvintes																				
Classificação	E_1	E_2	E_3	E_4	E_5	E_6	E_7	E_8	E_9	E_{10}	E_{11}	E_{12}	E_{13}	E_{14}	E_{15}	E_{16}	E_{17}	E_{18}	F_{19}	E_{20}	MOS Total
5. Excelente	6	22	20	20	10	19	18	17	13	25	24	23	19	13	21	17	22	16	12	25	362
4. Bom	17	15	17	17	19	11	16	16	17	13	13	15	17	16	16	13	10	13	14	13	298
3. Regular	13	2	3	2	10	9	4	7	9	1	3	1	4	10	2	9	6	11	11	2	119
2. Pobre	4	1	0	1	1	1	2	0	1	1	0	1	0	1	1	1	2	0	3	0	21
1. Péssimo	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
E_i Total	40	40	40	40	40	40	40	40	40	40	40	40	40	40	40	40	40	40	40	40	800

- Escala -	Freqüência Relativa (em porcentagem) do Escore na Avaliação das 20 Frases por 40 Ouvintes																				
Classificação	R_1	R_2	R_3	R_4	R_5	R_6	R_7	R_8	R_9	R_{10}	R_{11}	R_{12}	R_{13}	R_{14}	R_{15}	R_{16}	R_{17}	R_{18}	R_{19}	R_{20}	MOS % Total
5. Excelente	15	55	50	50	25	47,5	45	42,5	32,5	62,5	60	57,5	47,5	32,5	52,5	42,5	55	40	30	62,5	45,25
4. Bom	42,5	37,5	42,5	42,5	47,5	27,5	40	42,5	42,5	32,5	32,5	37,5	42,5	40	40	32,5	25	32,5	35	32,5	37,25
3. Regular	32,5	5	7,5	5	25	22,5	10	40	22,5	2,5	7,5	2,5	10	25	5	22,5	15	27,5	27,5	5	14,875
2. Pobre	10	2,5	0	2,5	2,5	2,5	5	17,5	2,5	2,5	0	2,5	0	2,5	2,5	2,5	5	0	7,5	0	2,625
1. Péssimo	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
R_i % Total	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100

A partir dos dados obtidos na coluna MOS % Total da Tabela 7.12 é traçado um gráfico conforme mostrado na Figura 7.9. Neste gráfico observa-se que os percentuais de excelente (45%) e bom (37%) prevalecem sobre regular (15%) e pobre (3%).

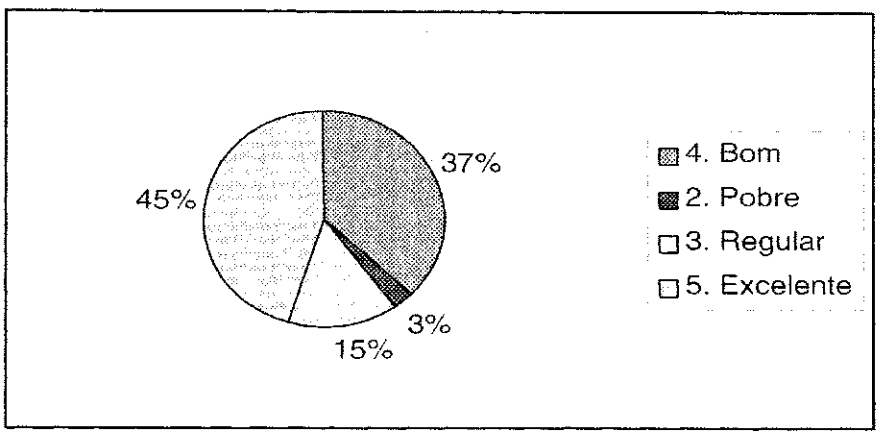

Figura 7.9: - Avaliação MOS das 20 Frases.

Também é possível determinar a avaliação global calculando-se a média ponderada da freqüência de ocorrência do escore na avaliação das 20 frases, multiplicando-se cada número de ocorrências da coluna MOS Total, da parte superior da Tabela 7.12, pelo valor do seu respectivo escore, somando-se os resultados parciais obtidos e dividindo-se o sub-total pelo número total de ocorrências (800). O valor encontrado neste caso é 4,25, ou seja, um pouco superior ao escore 4 (Bom).

Assim, os resultados obtidos demonstram o bom desempenho do modelo de prosódia de palavras proposto, considerando que as 20 frases utilizadas constituem uma amostra representativa do universo de todos os fonemas e unidades do dicionário apresentado no Capítulo 6.

Capítulo 8

Conclusões e Sugestões

Neste capítulo são apresentadas algumas conclusões importantes sobre o modelo prosódico proposto. Também são apresentadas as contribuições e sugestões sobre o desenvolvimento de trabalhos futuros que podem complementar ou aperfeiçoar a proposta de modelo aqui apresentada.

8.1 Sumário da Pesquisa

O trabalho aqui descrito apresenta uma proposta de modelo prosódico para um conversor texto-fala, usando a síntese concatenativa, para a Língua Portuguesa falada no Brasil. O modelo é baseado em regras e na tonicidade de palavras oxítonas, paroxítonas e proparoxítonas com até cinco sílabas.

Para a elaboração do modelo proposto foi feito inicialmente um estudo sobre a forma natural de produção da fala.

Posteriormente, foi realizado um levantamento das propostas de conversores desenvolvidos em outras línguas, como, por exemplo, para a Língua Inglesa e para a Língua Francesa, incluindo os estágios de processamento do texto, processamento prosódico e as técnicas de síntese do sinal da fala. Foi observado que o conversor texto-fala utilizando a síntese concatenativa tem sido bastante usado atualmente, considerando-se as vantagens de simplicidade, flexibilidade e sobretudo a naturalidade da fala.

No processamento do texto, foi observado que existe uma série de etapas fundamentais, tais como o analisador de texto e a transcrição letra-fonema, que são dependentes do tipo de idioma. Por outro lado, dentre as técnicas de síntese do sinal, foi observado que o algoritmo PSOLA tem servido atualmente como base para o desenvolvimento de

grande parte dos sintetizadores concatenativos.

Assim, partiu-se para o desenvolvimento de um modelo de prosódia para um conversor texto-fala concatenativo tendo como ponto inicial a concepção de um dicionário de unidades acústicas, conforme os procedimentos adotados no Capítulo 6, considerando que o resultado final da síntese da fala é dependente da qualidade das unidades e da necessidade de se definir o tipo de unidade usada no modelo.

Para a determinação do modelo de entonação, foram medidos os valores de F_0 das unidades acústicas correspondentes às sílabas do *corpus*, usando-se o detetor de *pitch* desenvolvido por Fechine em [71]. Foram analisados os valores de freqüência fundamental correspondente ao *pitch* de sílabas em palavras oxítonas, paroxítonas e proparoxítonas com até cinco sílabas. Foram identificados aspectos importantes, como, por exemplo: tendência de crescimento do valor de F_0 da sílaba pretônica para a postônica. Também foram medidos e analisados os valores de duração da sílabas, utilizando-se o editor de áudio *Sound Forge*. Assim, foram realizados ajustes nos valores de duração e F_0 das unidades acústicas do dicionário de modo a atender o modelo proposto.

Para a análise subjetiva da qualidade da fala produzida na síntese utilizando o modelo, foi desenvolvido um conversor texto-fala, implementado em *software*, que utiliza os conceitos de processamento lingüístico e de sinal, apresentados nos Capítulos 3 e 5. Posteriormente foi criado um ambiente de testes usando o compilador C++ *Builder* 5, conforme descrito na Seção 7.5 do Capítulo 7, para facilitar o manuseio dos dados no estágio de análise.

Portanto, o modelo prosódico proposto, foi obtido a partir de uma análise dos sinais da fala, como também de uma análise estrutural de palavras com até cinco sílabas, levando-se em consideração o tipo de sílaba (CV, CVC, ...), a sua tonicidade e a sua posição relativa dentro da palavra.

8.2 Conclusões

A fala tem sido o meio natural e preferido na comunicação humana, considerando-se a rapidez com que se processa a informação com relação a outros meios, como, por exemplo, a escrita. Dentro desse contexto foram desenvolvidos vários sistemas de conversão texto-fala nos últimos anos, considerando-se sobretudo as aplicações descritas no Capítulo 1 e a disponibilidade de recursos computacionais.

O desenvolvimento e implementação de um conversor texto-fala constituem-se em uma tarefa complexa e um grande desafio, tendo em vista a necessidade de conhecimentos nas áreas de Processamento Digital de Sinais e Lingüística Computacional. A qualidade da fala produzida por um conversor concatenativo está relacionada com a inteligibilidade, naturalidade e reconhecibilidade da mesma, o que requer um processamento do texto adequado, sobretudo um estágio de processamento prosódico, e um processamento de sinal que proporcione o devido ajuste e concatenação das unidades acústicas.

Algumas dificuldades iniciais são identificadas na conversão texto-fala, como, por exemplo, a diversidade de estruturas oracionais e dos sotaques regionais. Além disso, o desenvolvimento dos vários estágios do conversor texto-fala envolve um certo grau de complexidade, dependendo do tipo de idioma e do tipo de síntese, destacando-se os seguintes itens:

- criação de um banco de dados com siglas, números e abreviaturas e de um algoritmo para a conversão desses caracteres em palavras por extenso;
- criação de um banco de dados com as palavras do idioma e de algoritmos para a análise do texto (análise morfológica, semântica e sintática);
- criação de regras baseadas na composição fonética das palavras e de um algoritmo para a transcrição fonética do texto;
- criação de um modelo prosódico e de um algoritmo para processamento prosódico a partir da transcrição fonética do texto;
- determinação de um método de síntese do sinal da fala.

Dentro deste contexto, está sendo proposto um modelo para a geração de prosódia de palavras em um conversor texto-fala para a Língua Portuguesa falada no Brasil. O modelo é baseado em regras e na tonicidade das palavras e em um dicionário contendo unidades acústicas contemplando regras fonéticas e fonológicas, como também transições entre os fonemas, correspondentes ao fenômeno de coarticulação.

O dicionário de unidades acústicas foi desenvolvido conforme os procedimentos apresentados no Capítulo 6 e corresponde a uma base de dados contendo o máximo de informações prosódicas para que se obtenha uma fala mais natural possível. Inicialmente foi desenvolvida uma metodologia, da qual o estágio mais complexo foi a

elaboração do *corpus*, e o estágio mais laborioso, a segmentação das unidades através do editor de áudio *Sound Forge*, tendo sido necessário além de visualizar o sinal de cada unidade, avaliá-lo através de testes de escuta informais.

A gravação do *corpus* foi realizada em um estúdio profissional com isolação acústica e com equipamentos de alta qualidade para a digitalização do sinal.

Foi desenvolvido um ambiente de testes para a validação do modelo prosódico contendo um conversor texto-fala conforme descrito no Capítulo 7, o qual foi usado para a avaliação qualitativa do sinal de fala sintetizado, e também para a análise a e correção das várias etapas do processamento do texto.

Também foram feitas medições de duração e freqüência fundamental de sílabas de palavras e a partir desses resultados e dos critérios estabelecidos no modelo prosódico, foram feitos ajustes nas unidades do dicionário como mostrado no Apêndice C. A partir da variação de duração e freqüência fundamental foram rotuladas as unidades do dicionário com o nome da unidade e o seu valor de duração e freqüência. Finalmente, foram estabelecidas regras para a busca automática das unidades e realizada a síntese da fala.

8.3 Contribuições

A partir dos resultados obtidos neste trabalho, podem ser destacadas algumas contribuições importantes:

1. Concepção de um dicionário de unidades acústicas, baseado em sílabas e demissílabas, para a extração da prosódia em um conversor texto-fala para utilização de síntese concatenativa;

2. Elaboração de regras para o pré-processamento do texto, para a transcrição letra-fonema e para a separação dos segmentos fonéticos a serem rotulados no estágio de modelagem prosódica;

3. Desenvolvimento de uma proposta de modelo prosódico baseado na tonicidade de palavras, a partir de considerações fonético-fonológicas e da análise de um *corpus* de palavras oxítonas, paroxítonas e proparoxítonas com até cinco sílabas, como também de um *corpus* de frases foneticamente balanceadas.

4. Elaboração de regras para geração automática da prosódia para um conversor texto-fala para a Língua Portuguesa, utilizando síntese concatenativa.

5. Desenvolvimento de um ambiente de testes, para análise de cada etapa da conversão texto-fala, como também para avaliação do modelo prosódico.

8.4 Sugestões para Trabalhos Futuros

O modelo prosódico proposto neste trabalho foi aplicado a um sistema utilizando a técnica síntese concatenativa simples com bons resultados, para um *corpus* de palavras e frases declarativas. Assim, sugere-se que, em trabalhos futuros, o modelo possa ser ampliado para outros contornos de entonação como, por exemplo, contornos para frases interrogativas e exclamativas. Também sugere-se que seja analisada a acentuação das sílabas em nível de frases, que seja realizado um estudo sobre a aplicação do modelo prosódico em um sistema texto-fala usando outras técnicas de síntese, como, por exemplo, a técnica TD-PSOLA e, um estudo sobre como a redução do número de unidades do dicionário pode afetar a qualidade da fala sintetizada, considerando-se a existência de dicionários com número reduzido de unidades [58, 93].

Apêndice A

Unidades Acústicas Usadas no Modelo Prosódico

Neste apêndice são apresentadas as matrizes contendo as unidades acústicas usadas no modelo prosódico que foram previamente definidas com base nos critérios relacionados no Capítulo 6 e que constituem o dicionário. Na composição das unidades também foram considerados os fonemas /é/, /ê/, /ó/ e /ô/. A matriz referente à composição CV já foi apresentada como exemplo no Capítulo 6 (Tabela 6.3) e, portanto, não se encontra neste apêndice com as demais.

Tabela A.1: Matriz referente às vogais (V)

Vogais (V)	a	am	é	ê	em	i	im	ó	ô	om	u	um

Tabela A.2: Matriz referente às combinações vogal-vogal (VV)

V/V	a	am	é	ê	em	i	im	ó	ô	om	u	um
a	aa	aam	aé	aê	aem	ai	aim	aó	aô	aom	au	aum
é	éa	éam	éé	éê	éem	éi	éim	éó	éô	éom	éu	éum
ê	êa	êam	êé	êê	êem	êi	êim	êó	êô	êom	êu	êum
i	ia	iam	ié	iê	iem	ii	iim	ió	iô	iom	iu	ium
ó	óa	óam	óé	óê	óem	ói	óim	óó	óô	óom	óu	óum
ô	ôa	ôam	ôé	ôê	ôem	ôi	ôim	ôó	ôô	ôom	ôu	ôum
u	ua	uam	ué	uê	uem	ui	uim	uó	uô	uom	uu	uum

Tabela A.3: Matriz referente às combinações vogal-semivogal (V-SV)

V/SV	a	am	é	ê	em	i	im	ó	ô	om	u	um
y(i)	ay	amy	éy	êy	emy	iy	imy	óy	ôy	omy	uy	umy
y(i)	ya	yam	yé	yê	yem	yi	yim	yó	yô	yom	yu	yum
w(u)	aw	amw	éw	êw	emw	iw	imw	ów	ôw	omw	uw	umw
w(u)	wa	wam	wé	wê	wem	wi	wim	wó	wô	wom	wu	wum

Tabela A.4: Matriz referente às combinações vogal-consoante (VC)

V/C	a	am	é	ê	em	i	im	ó	ô	om	u	um
b	ab	amb	éb	êb	emb	ib	imb	ób	ôb	omb	ub	umb
k	ak	amk	ék	êk	emk	ik	imk	ók	ôk	omk	uk	umk
d	ad	amd	éd	êd	emd	id	imd	ód	ôd	omd	ud	umd
g	ag	amg	ég	êg	emg	ig	img	óg	ôg	omg	ug	umg
p	ap	amp	ép	êp	emp	ip	imp	óp	ôp	omp	up	ump
t	at	amt	ét	êt	emt	it	imt	ót	ôt	om	ut	umt
f	af	amf	éf	êf	emf	if	imf	óf	ôf	omf	uf	umf
v	av	amv	év	êv	emv	iv	imv	óv	ôv	omv	uv	umv
j	aj	amj	éj	êj	emj	ij	imj	ój	ôj	omj	uj	umj
s	as	ams	és	ês	ems	is	ims	ós	ôs	oms	us	ums
x	ax	amx	éx	êx	emx	ix	imx	óx	ôx	omx	ux	umx
z	az	amz	éz	êz	emz	iz	imz	óz	ôz	omz	uz	umz
m	am	amm	ém	êm	emm	im	imm	óm	ôm	omm	um	umm
n	an	amn	én	ên	emn	in	imn	ón	ôn	omn	un	umn
nh	anh	amnh	énh	ênh	emnh	inh	imnh	ónh	ônh	omnh	unh	umnh
l	al	aml	él	êl	eml	il	iml	ól	ôl	oml	ul	uml
lh	alh	amlh	élh	êlh	emlh	ilh	imlh	ólh	ôlh	omlh	ulh	umlh
r	ar	amr	ér	êr	emr	ir	imr	ór	ôr	omr	ur	umr
rr	arr	amrr	érr	êrr	emrr	irr	imrr	órr	ôrr	omrr	urr	umrr

Tabela A.5: Matriz referente às combinações (CCV)

CC/V	a	am	é	ê	em	i	im	ó	ô	om	u	um
bl	bla	blam	blé	blê	blem	bli	blim	bló	blô	blom	blu	blum
br	bra	bram	bré	brê	brem	bri	brim	bró	brô	brom	bru	brum
kl	kla	klam	klé	klê	klem	kli	klim	kló	klô	klom	klu	klum
kr	kra	kram	kré	krê	krem	kri	krim	kró	krô	krom	kru	krum
ks	ksa	ksam	ksé	ksê	ksem	ksi	ksim	ksó	ksô	ksom	ksu	ksum
dr	dra	dram	dré	drê	drem	dri	drim	dró	drô	drom	dru	drum
fl	fla	flam	flé	flê	flem	fli	flim	fló	flô	flom	flu	flum
fr	fra	fram	fré	frê	frem	fri	frim	fró	frô	from	fru	frum
gl	gla	glam	glé	glé	glem	gli	glim	gló	glô	glom	glu	glum
gr	gra	gram	gré	grê	grem	gri	grim	gró	grô	grom	gru	grum
pl	pla	plam	plé	plê	plem	pli	plim	pló	plô	plom	plu	plum
pr	pra	pram	pré	prê	prem	pri	prim	pró	prô	prom	pru	prum
tr	tra	tram	tré	trê	trem	tri	trim	tró	trô	trom	tru	trum
vr	vra	vram	vré	vrê	vrem	vri	vrim	vró	vrô	vrom	vru	vrum

CCV	mna	mné	mnê	mni	mnó	mnô	mnu	pna	pné	pnê	pni	pnó
CCV	pnô	pnu	gna	gné	gnê	gni	gnó	gnô	gnu	fta	tmo	vla

Tabela A.6: Matriz referente às combinações consoantes-vogais-consoantes

C-V-C	a	é	ê	i	ó	ô	u
b-r	bar	bér	bêr	bir	bór	bôr	bur
b-s	bas	bés	bês	bis	bós	bôs	bus
k-r	kar	kér	kêr	kir	kór	kôr	kur
k-s	kas	kés	kês	kis	kós	kôs	kus
d-r	dar	dér	dêr	dir	dór	dôr	dur
d-s	das	dés	dês	dis	dós	dôs	dus
g-r	gar	gér	gêr	gir	gór	gôr	gur
g-s	gas	gés	gês	gis	gós	gôs	gus
p-r	par	pér	pêr	pir	pór	pôr	pur
p-s	pas	pés	pês	pis	pós	pôs	pus
t-r	tar	tér	têr	tir	tór	tór	tur
t-s	tas	tés	tês	tis	tós	tôs	tus
f-r	far	fér	fêr	fir	fór	fôr	fur
f-s	fas	fés	fês	fis	fós	fôs	fus
v-r	var	vér	vêr	vir	vór	vór	vur
v-s	vas	vés	vês	vis	vós	vós	vus
j-r	jar	jér	jêr	jir	jór	jôr	jur
j-s	jas	jés	jês	jis	jós	jôs	jus
s-r	sar	sér	sêr	sir	sór	sôr	sur
s-s	sas	sés	sês	sis	sós	sós	sus
x-r	xar	xér	xêr	xir	xór	xôr	xur
x-s	xas	xés	xês	xis	xós	xôs	xus
z-r	zar	zér	zêr	zir	zór	zôr	zur
z-s	zas	zés	zês	zis	zós	zôs	zus
m-r	mar	mér	mêr	mir	mór	mór	mur
m-s	mas	més	mês	mis	mós	môs	mus
n-r	nar	nér	nêr	nir	nór	nôr	nur
n-s	nas	nés	nês	nis	nós	nôs	nus
nh-r	nhar	nhér	nhèr	nhir	nhór	nhór	nhur
nh-s	nhas	nhés	nhês	nhis	nhós	nhôs	nhus

Tabela A.7: Cont. da matriz referente às combinações consoantes-vogais-consoantes

C-V-C	a	é	ê	i	ó	ô	u
l-r	lar	lér	lêr	lir	lór	lôr	lur
l-s	las	lés	lés	lis	lós	lôs	lus
lh-r	lhar	lhér	lhêr	lhir	lhór	lhôr	lhur
lh-s	lhas	lhés	lhês	lhis	lhós	lhôs	lhus
r-r	rar	rér	rêr	rir	rór	rôr	rur
r-s	ras	rés	rês	ris	rós	rôs	rus
rr-r	rrar	rrér	rrêr	rrir	rrór	rrôr	rrur
rr-s	rras	rrés	rrês	rris	rrós	rrôs	rrus

Tabela A.8: Matriz referente às combinações consoantes-vogais-vogais

b + VV(o)	béo	bêo	bio	béa	bêa	bia	bie	bóa	bôa	bua	bue	buo
b + VV(o)	bai	bau	béi	bêi	béu	bêu	biu	bói	bói	bóu	bôu	bui
b + VV(n)	buam	buem	buim	bami	bame	bemi	bome	bumi	bamo	-	-	-
k + VV(o)	kéo	kêo	kio	kéa	kêa	kia	kie	kóa	kôa	kua	kue	kuo
k + VV(o)	kai	kau	kéi	kêi	kéu	kêu	kiu	kói	kói	kóu	kóu	kui
k + VV(n)	kuam	kuem	kuim	kami	kame	kemi	kome	kumi	kamo	-	-	-
d + VV(o)	déo	dêo	dio	déa	dêa	dia	die	dóa	dôa	dua	due	duo
d + VV(o)	dai	dau	déi	dêi	déu	dêu	diu	dói	dói	dóu	dôu	dui
d + VV(n)	duam	duem	duim	dami	dame	demi	dome	dumi	damo	-	-	-
g + VV(o)	géo	gèo	gio	géa	gêa	gia	gie	góa	gôa	gua	gue	guo
g + VV(o)	gai	gau	géi	gêi	géu	gêu	giu	gói	gói	góu	gôu	gui
g + VV(n)	guam	guem	guim	gami	game	gemi	gome	gumi	gamo	-	-	-
p + VV(o)	péo	pêo	pio	péa	pêa	pia	pie	póa	pôa	pua	pue	puo
p + VV(o)	pai	pau	péi	pêi	péu	pêu	piu	pói	pói	póu	pôu	pui
p + VV(n)	puam	puem	puim	pami	pame	pemi	pome	pumi	pamo	-	-	-
t + VV(o)	téo	têo	tio	téa	têa	tia	tie	tóa	tôa	tua	tue	tuo
t + VV(o)	tai	tau	téi	têi	téu	têu	tiu	tói	tói	tóu	tôu	tui
t + VV(n)	tuam	tuem	tuim	tami	tame	temi	tome	tumi	tamo	-	-	-
f + VV(o)	féo	fêo	fio	féa	fêa	fia	fie	fóa	fôa	fua	fue	fuo
f + VV(o)	fai	fau	féi	fêi	féu	fêu	fiu	fói	fói	fóu	fôu	fui
f + VV(n)	fuam	fuem	fuim	fami	fame	femi	fome	fumi	famo	-	-	-
v + VV(o)	véo	vêo	vio	véa	vêa	via	vie	vóa	vôa	vua	vue	vuo
v + VV(o)	vai	vau	véi	vêi	véu	vêu	viu	vói	vói	vóu	vôu	vui
v + VV(n)	vuam	vuem	vuim	vami	vame	vemi	vome	vumi	vamo	-	-	-
j + VV(o)	jéo	jêo	jio	jéa	jêa	jia	jie	jóa	jôa	jua	jue	juo
j + VV(o)	jai	jau	jéi	jêi	jéu	jêu	jiu	jói	jói	jóu	jôu	jui
j + VV(n)	juam	juem	juim	jami	jame	jemi	jome	jumi	jamo	-	-	-
s + VV(o)	séo	sêo	sio	séa	sêa	sia	sie	sóa	sôa	sua	sue	suo
s + VV(o)	sai	sau	séi	sêi	séu	sêu	siu	sói	sói	sóu	sôu	sui
s + VV(n)	suam	suem	suim	sami	same	semi	some	sumi	samo	-	-	-

Tabela A.9: Cont. da matriz referente às combinações consoantes-vogais-vogais

x + VV(o)	xéo	xéo	xio	xéa	xêa	xía	xie	xóa	xôa
x + VV(o)	xua	xue	xuo	xóu	xóu	xui	-	-	-
x + VV(o)	xai	xau	xéi	xêi	xéu	xêu	xiu	xói	xói
x + VV(n)	xuam	xuem	xuim	xami	xame	xemi	xome	xumi	xamo
z + VV(o)	zéo	zêo	zio	zéa	zéa	zia	zie	zóa	zôa
z + VV(o)	zua	zue	zuo	zóu	zôu	zui	-	-	-
z + VV(o)	zai	zau	zéi	zêi	zéu	zêu	ziu	zói	zôi
z + VV(n)	zuam	zuem	zuim	zami	zame	zemi	zome	zumi	zamo
m + VV(o)	méo	mêo	mio	méa	méa	mia	mie	móa	môa
m + VV(o)	mua	mue	muo	móu	môu	mui	-	-	-
m + VV(o)	mai	mau	méi	mêi	méu	mêu	miu	mói	môi
m + VV(n)	muam	muem	muim	mami	mame	memi	mome	mumi	mamo
n + VV(o)	néo	nêo	nio	néa	nêa	nia	nie	nóa	nôa
n + VV(o)	nua	nue	nuo	nóu	nôu	nui	-	-	-
n + VV(o)	nai	nau	néi	nêi	néu	nêu	niu	nói	nôi
n + VV(n)	nuam	nuem	nuim	nami	name	nemi	nome	numi	namo
nh + VV(o)	nhéo	nhêo	nhio	nhéa	nhêa	nhia	nhie	nhóa	nhóa
nh + VV(o)	nhua	nhue	nhuo	nhóu	nhôu	nhui	-	-	-
nh + VV(o)	nhai	nhau	nhéi	nhêi	nhéu	nhêu	nhiu	nhói	nhôi
nh + VV(n)	nhuam	nhuem	nhuim	nhami	nhame	nhemi	nhome	nhumi	nhamo
l + VV(o)	léo	lêo	lio	léa	lêa	lia	lie	lóa	lóa
l + VV(o)	lua	lue	luo	lóu	lôu	lui	-	-	-
l + VV(o)	lai	lau	léi	lêi	léu	lêu	liu	lói	lôi
l + VV(n)	luam	luem	luim	lami	lame	lemi	lome	lumi	lamo
lh + VV(o)	lhéo	lhêo	lhio	lhéa	lhêa	lhia	lhie	lhóa	lhôa
lh + VV(o)	lhua	lhue	lhuo	lhóu	lhôu	lhui	-	-	-
lh + VV(o)	lhai	lhau	lhéi	lhêi	lhéu	lhêu	lhiu	lhói	lhôi
lh + VV(n)	lhuam	lhuem	lhuim	lhami	lhame	lhemi	lhome	lhumi	lhamo

Tabela A.10: Cont. da matriz referente às combinações consoantes-vogais-vogais

r + VV(o)	réo	rêo	rio	réa	rêa	ria	rie	róa	rôa
r + VV(o)	rua	rue	ruo	róu	rôu	rui	-	-	-
r + VV(o)	rai	rau	réi	rêi	réu	rêu	riu	rói	rôi
r + VV(n)	ruam	ruem	ruim	rami	rame	remi	rome	rumi	ramo
rr + VV(o)	rréo	rrêo	rrio	rréa	rrêa	rria	rrie	rróa	rrôa
rr + VV(o)	rrua	rrue	rruo	rróu	rrôu	rrui	-	-	-
rr + VV(o)	rrai	rrau	rréi	rrêi	rréu	rrêu	rriu	rrói	rrôi
rr + VV(n)	rruam	rruem	rruim	rrami	rrame	rremi	rrome	rrumi	rramo

Tabela A.11: Matriz referente às combinações consoantes-consoantes-vogais-consoantes (CCVC)

CC/VC	**as**	**ams**	**és**	**ês**	**ems**	**is**	**ims**	**ós**	**ôs**	**oms**	**us**	**ums**
bl	blas	blams	blés	blês	blems	blis	blims	blós	blôs	bloms	blus	blums
br	bras	brams	brés	brês	brems	bris	brims	brós	brôs	broms	brus	brums
kl	klas	klams	klés	klês	klems	klis	klims	klós	klôs	kloms	klus	klums
kr	kras	krams	krés	krês	krems	kris	krims	krós	krôs	kroms	krus	krums
dr	dras	drams	drés	drês	drems	dris	drims	drós	drôs	droms	drus	drums
fl	flas	flams	flés	flês	flems	flis	flims	flós	flôs	floms	flus	flums
fr	fras	frams	frés	frês	frems	fris	frims	frós	frôs	froms	frus	frums
gl	glas	glams	glés	glês	glems	glis	glims	glós	glôs	gloms	glus	glums
gr	gras	grams	grés	grês	grems	gris	grims	grós	grôs	groms	grus	grums
gn	gnas	gnams	gnés	gnês	gnems	gnis	gnims	gnós	gnôs	gnoms	gnus	gnums
lh	lhas	lhams	lhés	lhês	lhems	lhis	lhims	lhós	lhôs	lhoms	lhus	lhums
mn	-	-	mnés	mnês	-	mnis	-	mnós	mnôs	-	-	-
pl	plas	plams	plés	plês	plems	plis	plims	plós	plôs	ploms	plus	plums
pn	-	-	pnés	pnês	-	pnis	-	pnós	pnôs	-	-	-
pr	pras	prams	prés	prês	prems	pris	prims	prós	prôs	proms	prus	prums
tr	tras	trams	trés	três	trems	tris	trims	trós	trôs	troms	trus	trums
vr	vras	vrams	vrés	vrês	vrems	vris	vrims	vrós	vrós	vroms	vrus	vrums

Tabela A.12: Cont. da matriz referente às combinações consoantes-consoantes-vogais-consoantes (CCVC)

CC/VC	**ar**	**ér**	**êr**	**ir**	**ór**	**ôr**	**ur**
bl	blar	blér	blêr	blir	blór	blôr	blur
br	brar	brér	brêr	brir	bror	brôr	brur
kl	klar	klér	klêr	klir	klór	klôr	klur
kr	krar	krér	krêr	krir	kror	krór	krur
dr	drar	drér	drêr	drir	dror	drôr	drur
fl	flar	flér	flêr	flir	flór	flôr	flur
fr	frar	frér	frêr	frir	fror	frôr	frur
gl	glar	glér	glêr	glir	glór	glôr	glur
gr	grar	grér	grêr	grir	gror	grôr	grur
gn	gnar	gnér	gnêr	gnir	gnór	gnôr	gnur
lh	lhar	lhér	lhêr	lhir	lhór	lhôr	lhur
mn	mnar	mnér	mnêr	mnir	mnór	mnôr	mnur
pl	plar	plér	plêr	plir	plór	plôr	plur
pn	pnar	pnér	pnêr	pnir	pnór	pnôr	pnur
pr	prar	prér	prêr	prir	pror	prôr	prur
tr	trar	trér	trêr	trir	tror	trôr	trur
vr	vrar	vrér	vrêr	vrir	vror	vrôr	vrur

Apêndice B

Regras para Geração dos Segmentos Fonéticos

Para a definição das regras de geração dos segmentos fonéticos a partir da transcrição fonética de um texto, aplicado na entrada do conversor texto-fala desenvolvido neste trabalho, são consideradas as unidades já existentes no dicionário, como também as formas de combinações das letras para gerar as sílabas das palavras. Também é considerada a questão da nasalidade de uma sílaba sobre a antecedente, conforme mostrado a seguir.

B.1 Composição das Unidades

Inicialmente, a composição das unidades no processo de separação é realizada em três etapas:

- **Etapa 1**:

  São separados os grupos de fonemas compostos por CV (consoante + vogal) e os fonemas que não compõem tais grupos são separados (isoladamente), como mostrado nos exemplos da Tabela B.1.

Tabela B.1: Separação inicial dos grupos CV

Transcrição	batata	bluza	substamtivu	martíriu	abacati
Separação	ba/ta/ta	b/lu/za	su/b/s/ta/m/ti/vu	ma/r/tí/ri/u	a/ba/ca/ti

- **Etapa 2**:

  São verificadas se as consoantes que não foram incluídas nos grupos de fonemas (CV) irão permanecer separadas ou deverão ser unidas ao grupo anterior ou posterior.

  Consoante R:

- Se o R for sucedido por um grupo (CV) em que a consoante é um outro R, então, o R que está separado, deve se unir ao grupo que o sucede, como mostrado nos exemplos da Tabela B.2.

Tabela B.2: Composição da consoante R com grupos CV posteriores

Transcrição	carru	rrei	rrurau	carrosséu
Separação	ca/r/ru	r/re/i	r/ru/ra/u	ca/r/ro/sé/u
Composição	ca/rru	rre/i	rru/ra/u	ca/rro/sé/u

- Se não ocorreu o caso anterior, então este R, que está separado, deve se unir ao grupo que o antecede, como mostrado nos exemplos da Tabela B.3.

Tabela B.3: Composição da consoante R com grupos CV anteriores

Transcrição	martíriu	mar	morti
Separação	ma/r/tí/ri/u	ma/r	mo/r/ti
Composição	mar/tí/ri/u	mar	mor/ti

Consoante S:

- Se a consoante for um S, ela deve se unir ao grupo que a antecede, como mostrado nos exemplos da Tabela B.4.

Tabela B.4: Composição da consoante S com grupos CV anteriores

Transcrição	esta	mesmu	coizas	substamtivu
Separação	e/s/ta	me/s/mu	co/i/za/s	su/b/s/ta/m/ti/vu
Composição	es/ta	mes/mu	co/i/zas	su/bs/ta/m/ti/vu

Consoante M:

- Se a consoante for um M, ela deve se unir ao grupo que a antecede, como nos exemplos mostrados na Tabela B.5.

Tabela B.5: Composição da consoante M com grupos CV anteriores

Transcrição	camtu	anju	pemti	substamtivu
Separação	ca/m/tu	a/m/ju	pe/m/ti	su/bs/ta/m/ti/vu
Composição	cam/tu	am/ju	pem/ti	su/bs/tam/ti/vu

Consoantes diferentes de R, S e M:

- Se o grupo (CV) que sucede a consoante começa por R ou L, então a consoante que está separada deve se unir a este grupo, como mostrado nos exemplos da Tabela B.6.

Tabela B.6: Composição de consoantes diferentes de R, S e M com os grupos CV posteriores

Transcrição	bluza	pratu	crimi	flauta
Separação	b/lu/za	p/ra/tu	c/ri/mi	f/la/u/ta
Composição	blu/za	pra/tu	cri/mi	fla/u/ta

- Se não ocorrer o caso anterior então a consoante, que está separada, deve se unir ao grupo que a antecede, como nos exemplos da Tabela B.7.

Tabela B.7: Composição de consoantes diferentes de R, S e M com os grupos CV anteriores

Transcrição	signu	réptiu
Separação	si/g/nu	ré/p/ti/u
Composição	sig/nu	rép/ti/u

- Se o grupo é formado por duas consoantes, sendo a segunda S, esse grupo deve ser unido ao grupo que o antecede, como nos exemplos da Tabela B.8.

Tabela B.8: Composição de duas consoantes com os grupos CV anteriores

Transcrição	substantivu	abstratu	tóraks
Separação	su/bs/ta/n/ti/vu	a/bs/t/ra/tu	tó/ra/ks
Composição	subs/tan/ti/vu	abs/tra/tu	tó/raks/

- **Etapa 3**

São verificadas se as vogais que não foram incluídas nos grupos de fonemas (CV) irão permanecer separadas ou deverão ser unidas ao grupo anterior ou posterior. Neste caso a verificação é feita com o uso das regras de ditongo, tritongo e hiato.

Regra 3.1

Se for encontrada uma vogal sozinha no final de uma palavra e o segundo grupo antes dessa vogal for acentuado, então essa vogal irá formar um ditongo crescente se unindo ao grupo que a antecede, como nos exemplos da Tabela B.9.

Tabela B.9: Composição de vogais isoladas em final de palavra com grupos CV

Transcrição	relójio	vácuo	farmásia	série
Separação	re/ló/ji/o	vá/cu/o	far/má/si/a	sé/ri/e
Composição	re/ló/jio	vá/cuo	far/má/sia	sé/rie

Regra 3.2

Se for encontrada a vogal O antecedida por um grupo (CV) em que a vogal é um I e que é sucedida por um grupo iniciado com as consoantes N, M ou Z, então esse O deverá se unir ao grupo que o antecede, como mostrado nos exemplos da Tabela B.10.

Tabela B.10: Composição da Vogal O no meio da palavra com Grupos CV

Transcrição	ansiozo	frasionário
Separação	an/si/o/zo	fra/si/o/ná/rio
Composição	an/sio/zo	fra/sio/ná/rio

Regra 3.3

Se for encontrada uma vogal sozinha ou acompanhada por um S, sucedendo um grupo terminado com uma vogal com o sinal diacrítico til, então ela deverá se unir ao grupo que a antecede, como nos exemplos mostrados na Tabela B.11.

Tabela B.11: Composição de vogais antecedidas por um grupo CV com vogal e sinal diacrítico til

Transcrição	grãos	limão	limões	pães
Separação	g/rã/os	li/mã/o	li/mõ/es	pã/es
Composição	grãos	li/mão	li/mões	pães

Regra 3.4

Se for encontrada uma vogal isolada e acentuada, esta deverá permanecer separada, como nos exemplos mostrados na Tabela B.12.

Tabela B.12: Composição das vogais acentuadas.

Transcrição	viúva	siúmi	saúdi
Separação	vi/ú/va	si/ú/mi	sa/ú/di
Composição	vi/ú/va	si/ú/mi	sa/ú/di

Regra 3.5

Se forem encontradas as vogais O ou A sozinhas ou iniciando um grupo e se elas forem antecedidas por um grupo terminado em I, então elas deverão permanecer separadas. Isso só não ocorrerá com palavras monossílabas, como nos exemplos mostrados na Tabela B.13.

Tabela B.13: Composição das vogais O ou A antecedidas por grupos terminados em I

Transcrição	iatu	navio	apoio	tiu	pia
Separação	i/a/tu	na/vi/o	a/po/i/o	ti/o	pi/a
Composição	i/a/tu	na/vi/o	a/poi/o	tio	pia

Regra 3.6

Se forem encontradas as vogais I ou U, acompanhadas ou não por S, sucedendo grupos terminados com É ou Ó, então elas devem se unir a estes grupos para formar os ditongos abertos, como nos exemplos mostrados na Tabela B.14.

Tabela B.14: Composição das vogais I e U sucedendo a grupos terminados em É ou Ó

Transcrição	platéia	pastéu	destrói
Separação	pla/té/i/a	pas/té/u	des/tró/i
Composição	pla/téi/a	pas/téu	des/trói

Regra 3.7

Se for encontrada uma vogal isolada, ou iniciando um grupo, que é antecedido por um outro grupo terminado com a mesma vogal, esses grupos devem permanecer isolados, como nos exemplos mostrados na Tabela B.15.

Tabela B.15: Composição de vogais isoladas antecedidas por grupos terminados com a mesma vogal.

Transcrição	áukoou	kooperar	rreemkomtru
Separação	á/u/ko/o/u	ko/o/pe/rar	rre/em/kom/tru
Composição	á/u/ko/o/u	ko/o/pe/rar	rre/em/kom/tru

Regra 3.8

A vogal U resultante da mudança fonética do L deverá se unir ao grupo que a antecede, como nos exemplos mostrados na Tabela B.16.

Tabela B.16: Composição da vogal U correspondente ao L com grupos CV

Transcrição	áukoou	auvu	pouvu
Separação	á/u/ko/o/u	a/u/vu	po/u/vu
Composição	áu/ko/ou	au/vu	pou/vu

Regra 3.9

Se for encontrada uma vogal isolada ou iniciando um grupo, que é antecedido por um grupo GU, então a vogal deve se juntar ao grupo GU. Inclusive se houver mais de uma vogal isolada, todas elas deverão se unir formando um único grupo com o GU, como nos exemplos mostrados na Tabela B.17.

Tabela B.17: Composição de vogais isoladas com grupos GU

Transcrição	paraguai	uruguai	água
Separação	pa/ra/gu/a/i	u/ru/gu/a/i	á/gu/a
Composição	pa/ra/guai	u/ru/guai	á/gua

Regra 3.10

Se ainda houver vogal I ou U isolada, ou iniciando um grupo, que não tenha sido testada por nenhuma regra anterior, ela deverá se unir ao grupo que a antecede, formando ditongos ou tritongos, como nos exemplos mostrados na Tabela B.18.

Tabela B.18: Composição das vogais I e U sozinhas com os grupos CV

Transcrição	kouru	muitu	praia
Separação	ko/u/ru	mu/i/tu	p/ra/i/a
Composição	kou/ru	mui/tu	prai/a

B.2 Regras de Nasalização

Na geração dos segmentos fonéticos observa-se que alguns segmentos são nasalisados, quando o segmento seguinte é iniciado por **m** ou **n**. Para contornar tal problema foi definido um conjunto de regras a partir de uma análise feita em segmentos obtidos de 200 frases foneticamente balanceadas desenvolvidas por Alcaim *et al* em [68], e nas 1994 unidades obtidas nos 1994 logatomos do dicionário, conforme apresentado a seguir.

B.2.1 Casos em que Ocorre Nasalização

1. O segmento **ma** nasaliza a vogal do segmento (CV) ou (CCV) anterior, como se observa, por exemplo, em: camada (kam/ma/da) e drama (dram/ma).

2. O segmento **me** nasaliza a vogal do segmento (V) ou (CCV) anterior, como se observa, por exemplo, em: ciúme (si/um/me) e crime (krim/me).

3. O segmento **mi** nasaliza a vogal dos segmentos (CV) e (CCV) anteriores, quando nesses últimos a vogal for '**o**' ou '**e**' , como se observa, por exemplo, em: comia (kom/mi/a) e leucemia (leu/sem/mi/a).

4. O segmento **mus** nasaliza a vogal do segmento (CV) anterior, quando nesse último a vogal for '**e**', como se observa, por exemplo, em: nascemos (nas/cem/mus)

5. O segmento **na** nasaliza a vogal do segmento (CV) anterior, como se observa, por exemplo, em: canário (kam/na/ri/o) e dezena(de/zem/na).

6. O segmento **ne** nasaliza a vogal dos segmentos (CV) ou (CCV) anteriores, como se observa, por exemplo, em: planejo (plam/ne/jo) e telefone (te/le/fom/ne).

7. Os segmentos **nel** e **ner** nasalizam a vogal do segmento (V) anterior no início de palavra, como se observa, por exemplo, em: anel (am/nel), inerte (im/ner/te).

8. O segmento **ni** nasaliza a vogal do segmento (V) anterior no início de palavra ou (CV) anterior, em que V deve ser '**a**' . como se observa, por exemplo, em: animais (am/nim/ma/is), única (um/ni/ca), canibal (cam/ni/bal).

9. O segmento **no** nasaliza a vogal do segmento (CV) anterior, como se observa, por exemplo, em: duodeno(du/o/dem/no).

10. O segmento **nu** nasaliza a vogal do segmento (CV) anterior, como se observa, por exemplo, em: menu (mem/nu), tenue (tem/nue).

11. Os segmentos **nha**, **nhe** e **nho** nasalizam a vogal do segmento (CV) ou (CCV) anterior, como se observa, por exemplo, em: minha (mim/nha), tenho (tem/nho), sonhei (som/nhei).

B.2.2 Casos em que não Ocorre Nasalização

1. O segmento **men** não nasaliza a vogal do segmento (CV), (CCV) ou (VV) anterior, como se observa, por exemplo, em: fundamental (fun/da/men/tal), momento (mo/men/to), aumentou (au/men/tou) e blumenau (Blu/men/nau).

2. O segmento **mi** não nasaliza a vogal do segmento (CV) anterior, quando nessa última a vogal deve ser '**u**', ou da unidade (V) no início da palavra, como se observa, por exemplo, em: ilumina (i/lu/mi/na), humilhante (hu/mi/lhan/te) e emitido (e/mi/ti/do).

3. O segmento **mo** não nasaliza a vogal do segmento (VV) anterior, como se observa, em: almoço (au/mo/so).

4. O segmento **mom** não nasaliza a vogal do segmento (CCV) anterior, como se observa, por exemplo, em: matrimônio (ma/tri/mom/ni/o),

5. O segmento **mu** não nasaliza a vogal dos segmentos (CVV) e (CV) anteriores, quando essa vogal deve ser 'l' ou '**u**', como se observa, por exemplo, em: calmo(cal/mo), caramujo (ca/ra/mu/jo).

6. O segmento **na** não nasaliza a vogal do segmento (CVV) anterior, como se observa, por exemplo, em: reinado (rei/na/do).

7. O segmento **nen** não nasaliza a vogal do segmento (CV) ou anterior, como se observa, por exemplo, em: impertinente (im/per/ti/nen/te).

8. O segmento **ni** não nasaliza a vogal dos segmentos (CV) e (V) anteriores, quando a vogal for '**o**' , como se observa, por exemplo, em: bonito (bo/ni/to), acionista (a/ci/o/nis/ta).

Apêndice C

Palavras Usadas no Modelo Prosódico

Neste apêndice são apresentadas as tabelas com palavras oxítonas, paroxítonas e proparoxítonas utilizadas na determinação do modelo de prosódia para palavras para um conversor texto-fala para a Língua Portuguesa. Estas palavras foram obtidas do dicionário Aurélio [1] e gravadas utilizando-se logatomos conforme descrito no Capítulo 6. Observou-se, na pesquisa realizada no dicionário, que o número de palavras diminui consideravelmente à medida que o número de sílabas aumenta. Isso contribuiu para que a tabela contendo as paroxítonas pentassílabas (Tabela C.1) seja menor do que as demais.

Tabela C.1: Palavras proparoxítonas pentassílabas usadas no modelo prosódico

Alcebíades	mimeógrafo	biológico	estrambótico	supersônico	despropósito	onomatopéia
higiênico	aerólito	edifício	fotogênico	catastrófico	diabólico	ornitólogo
econômico	aborígene	eletrônico	ilegítimo	cancerígeno	hidrelétrico	ortográfico
filarmônica	alcoólico	equilátero	inequívoco	científico	hipódromo	oceânico
matemática	alegórico	específico	informática	cronológico	hipótese	melancólico
paroxítona	ambulância	espetáculo	infrutífero	demoníaco	humorístico	monossílabo
aeródomo	anacrônico	esporático	jornalístico	desestímulo	neurastênico	zoológico
bibliófilo	primogênito	patriótico	paisagística	paralítico	tetrassílaba	telepático

Tabela C.2: Palavras proparoxítonas tetrassílabas usadas no modelo prosódico

abóbada	cerâmica	dissílabo	esquálido	exército	gramática	incrédulo	marítimo	protótipo	simbólico
acréscimo	cleópatra	distância	estático	exótico	hepático	indígena	mecânico	pseudônimo	simpático
acústico	climático	doméstico	estímulo	explícito	heráclito	inédito	minúsculo	psicólogo	sinônimo
adversário	coágulo	dramático	estúpido	estático	herbívoro	inóspito	monólogo	quadrúmano	solícito
agrícola	colérico	efêmero	eufórico	facínora	hermético	inquérito	monótono	quadrúpede	sonâmbulo
alcântara	congênito	elétrico	eurípedes	famélico	hidráulica	insípido	mortífero	quilômetro	termômetro
alérgico	crepúsculo	eletrônica	exâmine	fanático	hidrômetro	insólito	niágrara	raquítico	amálgma
alfândega	crisântemo	emérito	excêntrico	fantástico	hipérbole	intrépido	nipônico	recíproco	amêndoa
aloísio	cronômetro	empréstimo	exército	farândola	hipódromo	inúmero	obstáculo	república	Américo
aricte	cubículo	enérgico	êxito	fatídico	Hipólito	inválido	oxítona	ridículo	Angélico
arquétipo	cutícula	espécime	êxodo	fenômeno	hipótese	legítimo	pacífico	romântico	antártico
autêntico	debênture	epílogo	exótico	filósofo	histórica	leucócito	parêntese	sarcástico	turístico
azêmola	decrépito	epístoloa	explícito	fôlego	horóscopo	limítrofe	parênteses	sarcófago	utópico
benéfico	decúbito	epístolo	estático	fonética	idêntica	lunático	patético	satânico	veículo
binóculo	desânimo	equívoco	estímulo	frenético	idólatra	magnânimo	penitência	satélite	verídico
botânica	devêramos	erótico	estúpido	frutífero	ilícito	magnífico	pernóstico	satírico	vernáculo
cardíaco	didático	escândalo	eufórico	gasômetro	ilógico	maiúsculo	poética	semáforo	vinícola
carnívoro	dinâmico	esférico	Eurípedes	gastrônomo	incógnita	maléfico	político	semítico	vocábulo
catástrofe	Diógenes	espírito	exâmine	ginástica	incógnito	malévolo	pretérito	sepúlveda	vulcânico
cenáculo	discípulo	esplêndido	excêntrico	girândola	incólume	mamífero	propósito	silvícola	zodíaco

Tabela C.3: Palavras proparoxítonas trissílabas usadas no modelo prosódico

ânfora	bússola	cínico	fábrica	hálito	lânguido	módulo	pânico	pródigo	sábado	válido
ângelus	cálculo	dádiva	fábula	hégira	lápide	músculo	pântano	pródomo	sádico	válvula
ângulo	cálice	dálmata	fáceis	hélice	lástima	música	pároco	prógnato	sátira	vándalo
ânimo	cálido	débito	física	híbrido	lépido	náutico	pássaro	próspero	século	vértebra
ântipas	câmara	década	fórmica	hípico	léxico	nítido	pátima	prótese	séquito	vértice
ápice	câmera	déficit	fórmula	hóspede	línguido	nódulo	pégaso	próximo	sésamo	véspera
árbitro	cândido	déspota	frêmito	húngaro	límpido	nômade	pélvico	público	sífilis	víbora
áspero	cânone	dívida	frígido	idêntico	líquido	número	pênalti	púrpura	sílica	vínculo
áspide	cântico	dízima	frívolo	ídolo	mágico	núpcias	pêndulo	pústula	símbolo	vírgula
átomo	cápsula	dízimo	fúlgido	ímprobo	máquina	óbito	pêsames	quádruplo	síndrome	vísceras
bárbaro	cárcere	drástico	fúnebre	índice	mármore	óbolo	pêssego	química	síntese	vítima
básico	cédula	dúplice	gélido	íngrime	máximo	ômega	péssimo	réplica	sólido	víveres
bátega	célebre	dúvida	gênero	íntegro	médico	ônibus	pílula	réprobo	súbito	vívido
bêbabo	célere	êmbolo	gênese	ínterim	mérito	ópera	plácido	réquiem	súdito	vômito
bélico	céltico	ênfase	glândula	íntimo	método	órbita	pólvora	rígido	trânsito	xácara
bíblico	célula	época	glóbulo	ípsilon	métrico	ósculo	póstuma	ríspido	último	xícara
bígamo	cênico	éramos	gôndola	Ítala	mímica	óxido	póstumo	rítmico	úmido	zéfiro
biótipo	cérebro	êxito	grânulo	lábaro	mínimo	página	prática	rótula	único	zênite
bípede	cético	êxodo	hábitat	lâmina	místico	pálido	préstimo	rótulo	úrsula	zíngaro
búfalo	chácara	êxtase	hábito	lâmpada	mítico	pândega	príncipe	rústico	úvula	tática

Tabela C.4: Palavras paroxítonas trissílabas usadas no modelo prosódico

abismo	cacique	doutrina	enfermo	grevista	irado	migalha	obscuro	quadrado	safado	utopia
acaso	caçula	eclipse	engano	grisalho	ironia	milagre	obséquio	quadrilha	safári	vacina
acento	cadeia	efeito	façanha	gritante	jeitoso	milênio	ocaso	quarteto	sagrado	vaguear
acervo	caduco	Egídio	faceiro	grosseiro	joalheiro	mimoso	oculto	querido	saída	valente
acesso	caipira	Egrégio	fachada	grotesco	jornada	mineiro	ofensa	quieto	salário	valentia
açouge	calçado	eleito	facial	guaíba	juízo	ministro	oferta	quilombo	salobro	vampiro
açúcar	calote	elite	fadiga	guerreiro	jumento	minúcia	ofício	quinteto	tristeza	vantagem
baderna	calouro	elogio	falência	guloso	jurado	miragem	olaria	quinzena	tristonho	velhaco
bagulho	discurso	embrulho	fulano	harmonia	jazida	narina	olfato	quitute	triunfo	xarope
bagunça	dispensa	empenho	fulgente	herbáceo	lacuna	nascença	opaco	romeiro	trombada	xereta
balança	distante	emprego	fumante	herdeiro	lamentar	nascente	pacato	rosado	tumulto	zabumba
balido	ditado	empresa	fundura	herança	lamúria	nativo	padeiro	roteiro	turismo	zangado
banguela	diurno	encalço	funesto	história	larápio	naufrágo	pagode	rotina	turísta	zarolho
banquete	divino	encanto	furioso	honesto	largada	neblina	paisagem	rubrica	uísque	zeloso
barato	docente	encargo	furtivo	honrado	largura	negócio	palácio	ruído	ultraje	zombaria
barulho	doença	encontro	futuro	inútil	lavagem	negreiro	palavra	ruidoso	unido	zumbido
batalha	doente	encosta	goteira	inveja	leitura	nervoso	palerma	rodovia	urbano	zunido
cabana	doido	encosto	gozado	invento	mesura	nevasca	palhaço	rupestre	urgência	sujeito
cabeça	donzela	encrenca	gratuito	inverno	metade	ninhada	palpite	ruptura	urgente	supremo
cachaça	dourado	energia	graúdo	invicto	micróbio	obeso	pancada	sadismo	usado	tabefe

Tabela C.5: Palavras paroxítonas dissílabas usadas no modelo prosódico

ágio	cáqui	Dóris	fútil	índio	mágoa	Nínjer	pêlo	sêmen	térreo
ágil	cárie	drágea	gêiser	íngua	martir	Nílson	pênis	sépia	téxtil
água	Cármem	dúbio	gênio	íon	médium	Nilton	pénsil	série	tórax
álbum	Célio	dúctil	gérsei	íris	méier	níquel	pêra	sério	trégua
álbuns	Sézar	dúzia	gíria	ísis	mícron	nível	plágio	símio	tróia
âmbar	cílio	ébrio	glútem	jérsei	mídia	nódoa	Plínio	sírio	túnel
ânsia	círio	édem	glúteo	jóia	Míltom	ócio	pôde	sóbrio	túneis
ânus	cíveu	égua	grácil	jókei	míngua	óleo	pólen	sócio	útil
área	Cleber	Ênio	grátis	Júnior	míope	ônus	pólo	sódio	vácuo
ária	Clóvis	Éster	hábil	júri	míssil	orfã	pônei	Sólon	Valter
Ásia	cólon	éter	Hélder	lábia	móvel	orfão	pôker	sóror	várzea
áudio	cônsio	fáseis	hérnia	lábio	náilon	orgão	próton	sócia	vêem
áureo	cônsul	fácil	hífen	lápis	náusea	ósseo	púbis	sotão	vênus
benção	kôo	fêmea	hímem	lêem	néctar	pâncreas	rádio	súcia	vício
bílis	creme	fêmur	hókei	líder	Nélson	pára	rédea	tábua	vídeo
bônus	cútis	fértil	hóstia	Lídia	néon	páreo	régua	tátil	Wilson
cádis	dêem	flúor	Húdson	língua	Néri	páscoa	réptil	táxi	vírus
cálcio	díspar	fórum	humus	Lóide	néscio	pátio	réstia	tédio	sítio
câncer	dólar	fóssil	imã	lótus	nêutron	péla	rímel	tênis	zangão
kânom	dôo	fóton	ímpio	Lúcia	névoa	pélo	róseo	tênue	zíper

Tabela C.6: Palavras oxítonas pentassílabas usadas no modelo prosódico

abandonar	concorrência	desencadear	ensurrecedor	familiarizar	inusitado	presidencial
acariciar	confeccionar	desencorajar	entristecedor	generalizar	investigador	providenciar
admoestar	confidencial	desestimular	entusiasmar	heterogênea	manifestação	realização
alfabetizar	confidenciar	desmoralizar	esbofetar	heterogêneo	modificação	recaptular
amadurecer	consideração	desobedecer	escandalizar	historiador	notabilizar	recomendação
amaldiçoar	conscientizar	desocupado	especificar	homenagear	organização	reconciliação
ameaçador	datilografar	desorientar	espetacular	hospitalizar	organizador	reconciliar
anoitecer	decepcionar	desvalorizar	espiritual	imaginação	ornamentação	reflorestar
beneficiar	democratizar	dicionário	esquematizar	individual	parabenizar	regularizar
civilização	desacomodar	diversificar	estabelecer	influenciar	paralisação	reiniciar
colaboração	desaconselhar	economizar	estabilizar	inocentar	peregrinação	reivindicar
colaborador	desacostumar	elaboração	exemplificar	insubordinação	possibilitar	rejuvenescer
colecionar	desaparecer	empalidecer	experimentar	internacional	preocupação	telespectador
comunicação	desclassificar	encolerizar	exteriorizar	inutilizar	superficial	uniformizar

Tabela C.7: Palavras oxítonas trissílabas usadas no modelo prosódico

abafar	abalar	abanar	abater	abolir	abortar	acabar	acatar	adorar	ajeitar	alegar
badalar	bajular	batalhar	bimensal	bimestral	blasfemar	bloquear	bocejar	bofetão	buzinar	caçador
caducar	cafundó	caminhar	camponês	canjerê	cativar	debandar	debater	decepar	decidir	decolar
delatar	depenar	depender	derivar	desatar	desfilar	editar	educar	eficaz	eleger	emitir
ensinar	entalar	escrever	esfriar	espetar	evitar	fabricar	farejar	fatigar	festejar	figurar
flexionar	flutuar	fracassar	folhear	fraturar	fundador	galopar	garantir	garimpar	genial	genitor
glacial	gorgear	grajaú	gratidão	guardar	guerrear	habitar	hastear	Havaí	hesitar	hibernar
hospedar	hospital	humilhar	ignorar	iludir	imbecil	imitar	imoral	imortal	incapaz	incolor
indicar	induzir	informar	ingerir	inserir	intervir	jataí	jejuar	Jericó	Josafá	jovial
judiar	justapor	justiça	juvenil	labutar	lastimar	lateral	lavrador	lesionar	levantar	liberar
libertar	limitar	liquidar	litoral	madrugar	magoar	maquinar	mastigar	matinal	meditar	melhorar
merecer	manual	modelar	obrigar	obstruir	obturar	ocorrer	ocular	ocultar	ofertar	operar
orador	oscilar	Pajeú	paladar	paletó	patoá	pecador	pedalar	perecer	picolé	polegar
pontapé	quarteirão	quinzenal	rasurar	realçar	rebaixar	receptor	reclamar	recolher	recortar	redentor
refazer	reforçar	saciar	sacristão	saltitar	salutar	salvador	secador	secular	sedutor	segredar
segurar	semestral	simular	temporal	tentação	tentador	terminar	tolerar	torturar	tosquiar	trabalhar
tradição	traição	traidor	transferir	transmitir	união	usual	usurpar	ultimar	vacilar	vadiar
vaguear	valentão	validar	vastidão	vegetal	vencedor	viajar	viciar	violar	vitimar	vigiar
ventilar	vendedor	vaiar	sacudir	tapear	supetão	subtrair	submerso	solução	soterrar	tapear
sucessor	tremular	triplicar	triturar	xilindró	vocação	voador	zelador	vomitar	xerocar	zombador

Apêndice D

Frases Usadas nos Testes MOS

Neste apêndice são relacionadas as 20 frases usadas nos testes MOS, para a avaliação qualitativa do modelo prosódico descrito no Capítulo 7, com base nas considerações feitas no Capítulo 8. Também são relacionados os valores de freqüência fundamental e duração das unidades acústicas usadas na síntese da fala.

Tabela D.1: Relação das 20 frases usadas nos testes MOS

1	sei que atingiremos o nosso objetivo	11	parece que nascemos ontem
2	o analfabetismo é a vergonha do país	12	hoje eu acordei muito calmo
3	a casa foi vendida sem pressa	13	nosso telefone quebrou
4	isso se resolverá de forma tranqüila	14	queremos discutir o orçamento
5	as crianças conheceram o filhote de ema	15	ela tem muita fome
6	a bolsa de valores ficou em baixa	16	hoje dormirei bem
7	uma garota foi presa ontem à noite	17	o termômetro marcava um grau
8	essa medida foi devidamente alterada	18	o discurso de abertura é bem longo
9	a mudança é lenta porém duradoura	19	eu precisei de microfone na conferência
10	muito prazer em conhecê-lo	20	nossa filha é a primeira aluna da classe

Tabela D.2: Freqüência fundamental e duração das unidades acústicas de 20 frases usadas nos testes MOS

Frase 1	Unidades	sei	ki	a	tim	gi	rem	mus	o	no	su	ob	je	ti	vu	
	Freqüência	108	140	124	118	116	128	130	120	117	118	126	126	129	135	
	Duração	364	321	239	272	256	315	270	265	347	203	257	239	326	239	
Frase 2	Unidades	o	am	nau	fa	be	tis	mu	é	a	ver	gom	nha	du	pa	ís
	Freqüência	120	124	115	113	128	132	138	133	124	105	123	126	135	115	124
	Duração	265	329	261	243	269	307	285	265	239	267	302	283	272	253	270
Frase 3	Unidades	a	ka	za	foi	vem	di	da	semi	pre	sa					
	Freqüência	124	124	126	131	105	119	122	107	114	121					
	Duração	239	309	230	291	269	296	261	316	278	241					
Frase 4	Unidades	i	su	si	rre	zou	ve	ra	di	for	ma	tram	kui	la		
	Freqüência	123	128	106	122	129	126	137	131	121	127	127	122	131		
	Duração	281	283	246	245	254	269	324	301	330	256	270	338	251		
Frase 5	Unidades	as	kri	am	sas	kom	nhe	se	ramo	o	fi	lho	ti	di	em	ma
	Freqüência	124	121	124	129	123	119	117	126	120	112	131	134	131	131	137
	Duração	364	251	298	395	288	270	300	271	265	233	288	266	301	290	256
Frase 6	Unidades	a	bou	sa	di	va	lo	ris	fi	kou	emi	bai	xa			
	Freqüência	124	125	128	131	116	120	124	112	113	106	119	123			
	Duração	239	277	264	301	260	316	208	321	288	316	284	267			
Frase 7	Unidades	um	ma	ga	ro	ta	foi	pre	za	om	temi	a	noi	ti		
	Freqüência	121	127	126	119	127	131	121	122	126	134	124	121	128		
	Duração	310	256	283	305	254	291	266	230	302	263	239	282	257		

Tabela D.3: Freqüência fundamental e duração das unidades acústicas de 20 frases usadas nos testes MOS

Frase 8	Unidades	e	sa	me	di	da	foi	de	vi	da	mem	ti	al	te	ra	da
	Freqüência	127	131	112	119	122	131	131	112	122	122	135	127	124	127	132
	Duração	305	241	241	296	261	291	257	240	261	302	257	252	285	324	261
Frase 9	Unidades	a	mu	dam	sa	é	lem	ta	po	rem	du	ra	do	ra		
	Freqüência	124	128	129	131	133	127	131	109	119	120	112	125	126		
	Duração	239	285	337	241	265	304	231	257	326	261	276	409	252		
Frase 10	Unidades	mui	to	pra	zer	emi	kom	nhe	sê	lu						
	Freqüência	125	126	121	129	106	123	119	127	130						
	Duração	336	274	274	253	316	288	270	300	261						
Frase 11	Unidades	pa	re	si	ki	nas	sem	mus	om	temi						
	Freqüência	115	124	126	140	122	128	130	126	131						
	Duração	253	309	246	321	277	380	270	302	263						
Frase 12	Unidades	o	ji	eu	a	kor	dei	mui	tu	kaum	mu					
	Freqüência	126	132	123	124	125	128	127	129	126	128					
	Duração	278	256	279	234	291	271	336	247	318	285					
Frase 13	Unidades	no	su	te	le	fom	ni	ke	brou							
	Freqüência	124	131	107	128	123	129	130	137							
	Duração	347	203	243	320	332	259	321	309							
Frase 14	Unidades	ke	rem	mus	dis	ku	tir	o	or	sa	mem	tu				
	Freqüência	122	127	130	109	114	119	120	120	123	125	126				
	Duração	221	315	270	272	254	333	265	232	264	302	274				

Tabela D.4: Freqüência fundamental e duração das unidades acústicas de 20 frases usadas nos testes MOS

Frase 15	Unidades	e	la	tem	mui	ta	fom	mi								
	Freqüência	127	133	129	127	131	133	139								
	Duração	305	251	334	336	231	332	262								
Frase 16	Unidades	o	ji	dor	mi	rei	bem									
	Freqüência	121	127	116	123	128	108									
	Duração	278	256	259	333	246	375									
Frase 17	Unidades	o	ter	mom	me	tru	mar	ka	va	um	grau					
	Freqüência	120	117	126	133	138	122	124	126	133	123					
	Duração	265	240	315	241	242	278	309	260	291	399					
Frase 18	Unidades	o	dis	kur	su	di	a	ber	tu	ra	é	bem	lom	gu		
	Freqüência	120	110	125	129	131	124	122	122	126	133	108	131	134		
	Duração	265	254	319	283	301	239	250	305	252	265	375	348	294		
Frase 19	Unidades	eu	pre	si	zei	di	mi	kro	fom	ni	na	kom	fe	rem	sia	
	Freqüência	123	142	147	159	131	111	127	123	128	116	123	136	127	134	
	Duração	279	250	343	256	301	260	275	332	259	348	288	262	315	271	
Frase 20	Unidades	no	sa	fi	lha	é	a	prim	mei	ra	a	lum	na	da	kla	si
	Freqüência	116	121	124	129	133	124	111	129	133	124	130	141	122	128	136
	Duração	347	241	321	268	265	239	266	326	262	239	337	237	261	339	246

Bibliografia

[1] Ferreira, A. B. de H. *Dicionário Aurélio Básico da Língua Portuguesa.* Editora Nova Fronteira, Rio de Janeiro, 2003.

[2] Gomes, L. de C. T. *Sistema de Conversão Texto-Fala para a Língua Portuguesa Utilizando a Abordagem de Síntese por Regras.* Dissertação de Mestrado, Faculdade de Engenharia Elétrica, Universidade Estadual de Campinas, Julho. 1998.

[3] Sproat, R. W.; Olive, J. P. Text-to-Speech Synthesis. *AT&T Technical Journal*, pp. 35–44, March/April, 1995.

[4] Silva, C. H. da. *Modelamento Prosódico para Conversão Texto-Fala do Português Falado no Brasil.* Dissertação de Mestrado. Faculdade de Engenharia Elétrica, Universidade Estadual de Campinas, Dezembro. 1995.

[5] Egashira, F.; Violaro, F. Conversor Texto-Fala para a Língua Portuguesa. *Anais do Simpósio Brasileiro de Telecomunicações 1995 - SBT'95*, pp. 71–76, Águas de Lindóia, Setembro, 1995.

[6] Rabiner, L. R. Applications of Voice Processing to Telecommunications. *Proceedings of the IEEE*, 2(82):199–228, February, 1994.

[7] Ribeiro, C. M.; Trancoso, I. Compressão de Sinais de Fala Baseada em Segmentos Etiquetados Foneticamente. *Proc. JETC'99 - Jornadas de Engenharia de Telecomunicações e Computadores*, ISEL, Lisbon, 1999.

[8] Pacheco, F. S. *Técnicas de Processamento de Sinais para Alteração de Parâmetros Prosódicos Aplicadas a um Sistema de Conversão Texto-Fala para a Língua Portuguesa Falada no Brasil.* Dissertação de Mestrado, Departamento de Engenharia Elétrica, Universidade Federal de Santa Catarina, Abril, 2001.

[9] Deketelaere, S.; Dutoit, T.; DEROO, O. Speech Processing for Communications: what's new?. *Revue HF*, pp. 5-24, March, 2001.

[10] Dutoit, T. *An Introduction to Text-to-Speech Synthesis*. Kluwer Academic Publishers, Dordrecht, The Netherlands, 1997.

[11] Dutoit, T.; Ricco, X. Towards a Free Multilingual Speech Synthesis Software for the Vocally Handicapped. *Österreichische Gesellschaft für Artificial Intelligence*, (20):36–38, July, 2001.

[12] Borges, J.A. *DOSVOX - um novo horizonte para deficientes visuais*. Revista Técnica do Instituto Benjamin Constant, nº 3, 1997.

[13] Borges, J.A. *Access and Technology: The DOSVOX Project - Changing the Lives of Thousands of Blind Brazilians*. Desability World, nº 4, August/September, 2000.

[14] Alves. J. B. M.; Miranda, A. S.; Torres, E. F. Análise Ergonômica dos Programas DOSVOX e VIRTUAL VISION. *I Seminário de Acessibilidade, Tecnologia da Informação e Inclusão Digital - I ATIID*, Vol. 1, pp. 1–2, São Paulo, SP, 2001.

[15] Araújo, A. M. de L. *Jogos Computacionais Fonoarticulatórios para Crianças com Deficiência Auditiva*. Tese de Doutorado, Faculdade de Engenharia Elétrica, Universidade Estadual de Campinas, Julho, 2000.

[16] Jianhua, T.; Xing, N. Auditive learning based Chinese F0 prediction. *Proceedings of Acoustics, Speech, and Signal Processing*, pp. 500–503, USA, April, 2003.

[17] Serralheiro, A.; Trancoso, I.; Caseiro, D.; Chambel, T.; Carrio, L.; Guimares, N. Towards a Repository of Digital Talking Books. *Proc. EUROSPEECH'2003 - 8th European Conference on Speech Communication and Technology (Interspeech'2003)*, pp. 1605–1608, Genve, Switzerland, September 2003.

[18] Serralheiro, A.; Caseiro, D.; Meinedo, H.; Trancoso, I.; Word Alignment in Digital Talking Books Using WFSTs. *Proc. ECDL'2002 - 6th European Conference on Digital Libraries*, pp. 508–515, Roma, Italy, September 2002.

[19] Y. Xu.; Hao, T.; Peiren, Z. Hadim, M.; Bagein, M.; Mannebach, P.; Maon, P. Load Balancing Voice Applications with Piranha. *The 2003 International*

Conference on Parallel and Distributed, pp. 1070–1082, Las Vegas, Nevada, June, 2003.

[20] Y. Xu.; Hao, T.; Peiren, Z. An advanced text-to-speech server system based on SOAP protocol. *Proceedings of Acoustics, Speech, and Signal Processing*, pp. 728-731, USA, April, 2003.

[21] J. Neto, H. Meinedo, R. Amaral, I. Trancoso. The development of an automatic system for selective dissemination of multimedia information. *Proc. CBMI'03 - Third International Workshop on Content-Based Multimedia*, Rennes, France, September 2003.

[22] Trancoso, I. The Alert System for Selective Dissemination of Multimedia Information. *Workshop "Past and Future of Speech Technology, Spoken Language Processing and Intelligent Multimedia", Invited Speaker*, Aalborg, Denmark, June 2003.

[23] Alves, F. O.; Osório, F. S. Integração de Regras e Exemplos para o Controle Inteligente de Robôs Autônomos. *XIII SIC - Anais do Salão de Iniciação Científica da UFRGS*. Editora da UFRGS, Vol. 1, nº 1, pp. 101-102, Porto Alegre, RS, 2001.

[24] Lopes, L. S. *et al.* Towards a Personal Robot with Language Interface. *Proc. EUROSPEECH'2003 - 8th European Conference on Speech Communication and Technology (Interspeech'2003)*, pp. 2205-2208, Genve, Switzerland, September 2003.

[25] Sallor, Ö.; Demirekler, M.; Pellom, B. A System for Voice Conversion Based on Adaptive Filtering and Line Spectral Frequency Distance Optimization for Text-to-Speech Synthesis. *Proc. EUROSPEECH'2003 - 8th European Conference on Speech Communication and Technology (Interspeech'2003)*, pp. 2205-2208, Genve, Switzerland, September 2003.

[26] De Nicola, J. e Infante, U. *Gramática Contemporânea da Língua Portuguesa*. Editora Scipione, 1997.

[27] Malfrère, F.; Dutoit, T.; Mertens, P. Automatic Prosody Generation Using Suprasegmental Unit Selection. *Proc. 3rd ESCA/COCSADA, Workshop on Speech Synthesis*, pp. 323-328, Jenolan Caves, Austrália, 1998.

[28] Madureira, S. Entoação e Síntese da Fala: Modelos e Parâmetros. *Estudos de Prosódia*, pp. 53-68, Editora da UNICAMP, Campinas - SP, 1999.

[29] Simões, F. O. *Implementação de um Sistema de Conversão Texto-Fala para o Português do Brasil*. Dissertação de Mestrado, Faculdade de Engenharia Elétrica, Universidade Estadual de Campinas, Maio, 1999.

[30] Eichner, M.; Wolff, M.; Hoffmann, R. Improved duration control for speech synthesis using a multigram language model. *Proceedings of Acoustics, Speech, and Signal Processing*, pp. 417–420, Orlando, USA, May, 2002.

[31] Yan, Q.; Vaseghi, S. Analysis, modelling and synthesis of formants of British, American and Australian accents. *Proceedings of Acoustics, Speech, and Signal Processing*, pp. 712–715, USA, April, 2003.

[32] Seresangtakul, P.; Takara, T. A generative model of fundamental frequency contours for polysyllabic words of Thai tones. *Proceedings of Acoustics, Speech, and Signal Processing*, pp. 452–455, USA, April, 2003.

[33] Sheng, Z.; Jianhua, T.; Ling, D. Chinese prosodic phrasing with extended features. *Proceedings of Acoustics, Speech, and Signal Processing*, pp. 492–495, USA, April, 2003.

[34] Rouas, J.-L.; Farinas, J.; Pellegrino, F.; Andre-Obrecht, R. Modeling prosody for language identification on read and spontaneous speech. *Proceedings of Acoustics, Speech, and Signal Processing*, pp. 40–43, USA, April, 2003.

[35] Campbell, N. Prosody and the Selection of Units for Concatenation Synthesis. *Proceedings of the Second ESCA/IEEE, Workshop on Speech Synthesis*, pp. 61-64, 1994.

[36] Dutoit, T.; Pagel, V.; Pierret, N.; Bataille, F.; Vrecken, O. van der. The MBROLA Project : Towards a Set of High Quality Speech Synthesizers Freee of Use for Non-Commercial Purposes. *Proceedings of International Conference on Spoken Language Processing*, pp. 1393–1396, Philadelphia, 1996.

[37] Chen, W.; Lin, F.; Li, J.; Zhang, B. Generation of chinese prosodic phrasing rules by a extension matrix algorithm. *Proceedings of Acoustics, Speech, and Signal Processing*, pp. 489–492, Orlando, USA, May, 2002.

[38] Prudon, R.; d'Alessandro, C.; de Mareüil, P. B. Prosody Synthesis by Unit Selection and Transplantation on Diphones. *Workshop on Speech Synthesis*, pp. 119–122, California, USA, September, 2002.

[39] Erdem, C.; Zimmermann, H.G. A data-driven method for input feature selection within neural prosody generation. *Proceedings of Acoustics, Speech, and Signal Processing*, pp. 477–480, Orlando, USA, May, 2002.

[40] Erdem, C.; Beck, F.; Hirschfeld, D.; Hoege, H.; Hoffmann, R. Robust Unit Selection Based on Syllable Prosody Parameters. *Workshop on Speech Synthesis*, pp. 159–162, California, USA, September, 2002.

[41] Jokisch, O.; Ding, H.; Kruscchke. Towards a multilingual prosody model for text-to-speech. *Proceedings of Acoustics, Speech, and Signal Processing*, pp. 421–424, Orlando, USA, May, 2002.

[42] Hu, W.; Huang, T.; Xu, B. Study on prosodic boundary location in Chinese Mandarin. *Proceedings of Acoustics, Speech, and Signal Processing*, pp. 501–504, Orlando, USA, May, 2002.

[43] Blouin, C.; Bagshaw, P.C.; Rosec, O. A method of unit preselection for speech synthesis based on acoustic clustering and decision trees. *Proceedings of Acoustics, Speech, and Signal Processing*, pp. 692–695, USA, April, 2003.

[44] Qian, Y.; Chen, F. Assigning phrase accent to Chinese text-to-speech system. *Proceedings of Acoustics, Speech, and Signal Processing*, pp. 485–488, USA, Orlando, USA, May, 2002.

[45] Venkataraman, A.; Ferrer, L.; Stolcke, A.; Shriberg, E. . Training a prosody based dialog act tagger from unlabeled data. *Proceedings of Acoustics, Speech, and Signal Processing*, pp. 272–275, USA, April, 2003.

[46] Milone, D. M.; Rubio, A. J. Prosodic and accentual information for automatic speech recognition. *Speech and Audio Processing*, 4(11):321–333, USA, July, 2003.

[47] Dusterhoff, K. E. *Synthesizing Fundamental Frequency Using Models Automatically Trained from Data.* PhD Thesis, University of Edinburgh, 2000.

[48] Zhang, J. S.; Hirose, K.; Nakamura, S. A multilevel framework to model the inherently confounding nature of sentential F0 sentential F0 contours contours for recognizing Chinese lexical tones. *Proceedings of Acoustics, Speech, and Signal Processing*, pp. 776–779, USA, April, 2003.

[49] Rama, G. L. J.; Ramakrishnan, A. G.; Muralishankar, R.; Prathibha, P. A Complete Text-to-Speech Synthesis System in Tamil. *Workshop on Speech Synthesis*, pp. 191–194, California, USA, September, 2002.

[50] Bulut, M.; Narayanan, A. G.; Syrdal, A. K. Expressive speech synthesis using a concatenative synthesizer. *Proceedings of International Conference on Spoken Language Processing*, pp. 1265–1268, Denver, Colorado, September 2002 .

[51] Solewicz, J. A. *Síntese de Voz a Partir do Texto para o Português do Brasil.* Dissertação de Mestrado, Pontifícia Universidade Católica, Rio de Janeiro, 1993.

[52] Barbosa, P. A. *et al.* Aiuruetê: A High-Quality Concatenative Text-To-Speech System for Braziliam Portuguese with Demisyllabic Analysis-Based Units and a Hierarchical Model of Rhythm Production. *European Conference on Speech Communication and Technology*, pp. 2059–2062, Sidney, Austrália, 1999.

[53] Violaro, F. Laboratório de Processamento Digital da Fala. *Seminário Conjunto UNICAMP e ITAUTEC*, Departamento de Comunicações da Faculdade de Engenharia Elétrica e Computação da Universidade Estadual de Campinas, Campinas, Brasil, 2003.

[54] Barbosa, P. A. Generating Duration from a Cognitively Plausible Model of Rhythm Production. *Proceedings of the Seventh European Conference on Speech Communication and Technology (Eurospeech 2001)*, Vol. 2, pp. 967–970, Aalborg, Dinamarca, 2001.

[55] Figueiredo, F. A. de; Costa Neto, M. L. da; Naviner, L. de B.; Azevedo, J. A. de; Aguiar Neto, B. G. Analisador de Texto para um Sistema de Conversão Texto-Fala para a Língua Portuguesa. *III Encontro para o Processamento Computacional de Português Escrito e Falado (PROPOR'98)*, pp. 17–22, Porto Alegre, RS, Novembro, 1998.

[56] Figueiredo, F. A. de. *Processamento Lingüístico para um Conversor Texto-Fala para a Língua Portuguesa.* Dissertação de Mestrado. Universidade Federal da Paraíba, Departamento de Engenharia Elétrica - UFPB, Setembro, 1998.

[57] Costa Neto, M. L. da. *Conversor Texto-Fala de Alta Qualidade para a Língua Portuguesa.* Exame de Qualificação, Departamento de Engenharia Elétrica, Universidade Federal da Paraíba, Abril, 2000.

[58] Kafka, S. G.; Pacheco, F. S.; Seara, I. C.; Klein, S.; Seara, R. Utilização de Segmentos Transicionais Homorgânicos em Síntese de Fala. *XIV Congresso Brasileiro de Automática - CBA 2002*, pp. 2742-2747, Natal, RN, Setembro 2002.

[59] Pacheco, F. S.; Seara, R. Prosodic Speech Modification Using RELP. *IEEE International Telecommunication Symposium - ITS 2002*, pp. 1-6, Natal, RN, Setembro 2002.

[60] Seara, I. C.; Kafka, S. G.; Klein, S.; Seara, R. Alternância Vocálica das Formas Verbais e Nominais do Português Brasileiro para Aplicação em Conversão Texto-fala. *Revista da Sociedade Brasileira de Telecomunicações*, Vol. 17, $n^{\underline{o}}$ 1, pp. 79-85, Junho 2002.

[61] Massini-Cagliari, G. *Acento e Ritmo.* Coleção Repensando a Língua Portuguesa, Editora Contexto, São Paulo - SP, 1992.

[62] Madureira, S. Pitch Patterns in Brazilian Portuguese: An Acoustic-Phonetic Analysis. *Proceedings of the Fifth Australian International Conference on Speech Science and Technology*, (1):156–161, Perth, Austrália, 1994.

[63] Madureira, S.; Fontes, M. A. S. Fundamental Contours in Brazilian Portuguese Words. *Proceedings of the ESCA Workshop Intonation: Teory, Models and Applications*, pp. 211–214, Athens, Greece, September 1997.

[64] Madureira, S.; Barbosa, P. A.; Fontes, M. A. S.; Crispin, K.; Spina, D. Post-Stressed Syllables in Brazilian Portuguese as Markers. *Proceedings of the XIV International Congress of Phonetic Sciences*, San Francisco, Berkeley: University of California, United States of America, 1999.

[65] Barbosa, P. A. Revelar a Estrutura Rítmica de uma Língua Construindo Máquinas Falantes: Pela Integração da Ciência e Tecnologia da Fala. *Estudos de Prosódia*, pp. 21-52, Editora da UNICAMP, Campinas - SP, 1999.

[66] Vainio, M.; Järvikivi, J.; Werner, S. Effect Prosodic Naturalness on Segmental Acceptability in Synthetic Speech. *Workshop on Speech Synthesis*, pp. 143–146, California, USA. September, 2002.

[67] Minematsu, N.; Kita, R.; Hirose, K. Automatic Estimation of Accentual Atribute Values of Words to Realize Accent Shandi in Japanese Text-to-Speech Conversion. *Workshop on Speech Synthesis*, pp. 107–110, California, USA, September, 2002.

[68] Alcaim, A.; Solewicz, J. A.; Moraes, J. A. de. Freqüência de Ocorrência dos Fones e Listas de Frases Foneticamente Balanceadas no Português Falado no Rio de Janeiro. *Revista da Sociedade Brasileira de Telecomunicações*, 7(1):23-41, Dezembro, 1992.

[69] Dubois, J.; Giacomo, M.; Guespin, L.; Marcellesi, C.; Marcellesi, J. B.; Mevel, J. P. *Dicionário de Lingüística*. Editora Cultrix, São Paulo, 1998.

[70] Silva, T. C. *Fonética e Fonologia do Português*. Editora Contexto, São Paulo - SP, 2002.

[71] Fechine, J. M. *Reconhecimento Automático de Identidade Vocal Utilizando Modelagem Híbrida: Paramétrica e Estatística*. Tese de Doutorado, Departamento de Engenharia Elétrica, Universidade Federal da Paraíba, Dezembro, 2000.

[72] Russo, I. e Behlau, M. *Percepção da Fala: Análise Acústica*. Editora Lovise, São Paulo - SP, 1993.

[73] Junior, J. R. D.; Proakis, J. G.; Hansen, J. H. L. *Discrete-Time Processing of Speech Signals*. Macmillan Publishing Company, 1993.

[74] Callou, D.; Leite, Y. *Iniciação à Fonética e à Fonologia*. Jorge Zahar Editor Ltda. Rio de Janeiro, 1999.

[75] Bechara, E. *Moderna Gramática Escolar da Língua Portuguesa*. Editora Lucerna, 2001.

[76] Mertens, P.; Beaugendre, F.; d'Alessandro, C. Comparing Approaches to Pitch Contour Stylization for Speech Synthesis. In: van Santen, J. P. H.; Sproat, R. W.; Olive, J. P.; Hirschberg, J. (Eds.). *Progress in Speech Synthesis*. Springer. pp. 347-364, 1997.

[77] Black, A. W.; Hunt, A. J. Generating F_0 Contours from ToBI Labels Using Linear Regression. *Proceedings of International Conference on Spoken Language Processing*, pp. 1385–1388, Philadelphia, Penn, 1996.

[78] Kientzle, T. *A Programmer's Guide to Sound*. Addison-Wesley Books, 2002.

[79] Souza e Silva, M. C. P. de; Koch, I. G. *Linguística Aplicada ao Português: Morfologia*. Cortez Editora, São Paulo, 1999.

[80] Cunha, C. F.; Cintra, L. F. L. *Nova Gramática do Português Contemporâneo*. 3^a Edição, Editora Nova Fronteira S.A., Rio de Janeiro. 2001.

[81] Souza e Silva, M. C. P. de; Koch, I. G. *Linguística Aplicada ao Português: Sintaxe*. Cortez Editora, São Paulo. 1998.

[82] Albano, E. C.; Moreira, A. A. Archisegment-Based Letter-to-Phone Conversion for Concatenative Speech Synthesis in Portuguese. *Proceedings of International Conference on Spoken Language Processing*. Philadelphia, Penn, pp. 1708–1711, October, 1996.

[83] Shih, C.; Ao, B. Duration Study for the Bell Laboratories Mandarin Text-to-Speech System. In: van Santen, J. P. H.; Sproat, R. W.; Olive, J. P.; Hirschberg, J. (Eds.). *Progress in Speech Synthesis*. Springer, pp. 383-399, 1997.

[84] Hoffmann, R.; Jokisch, O.; Hirschfeld, D.; Strecha, G.; Kruschke, H.; Kordon, U.; Koloska, U. A multilingual TTS system with less than 1 Mbyte footprint for embedded applications. *Proceedings of Acoustics, Speech, and Signal Processing*, pp. 532–535, USA, April, 2003.

[85] Barbosa, P. A.; Bailly, G. Generation of Pauses Within the *z-score* Model. In: van Santen, J. P. H., Sproat, R. W., Olive, J. P. and Hirschberg, J. (Eds.). *Progress in Speech Synthesis*. Springer, pp. 361-381, 1997.

[86] Black, A. W.; Taylor, P.; Caley, R. *The Festival Speech Synthesis System: System Documentation.* University of Edinburg, Edition 1.4, Version 1.4.2, July, 2001.

[87] Lemmetty, S. *Review of Speech Synthesis Technology.* Master Thesis, Helsinki University of Technology, March, 1999.

[88] Haykin, S. S. *Redes Neurais - Princípios e Prática.* Bookman Companhia Editora Ltda, Porto Alegre, RS, 2001.

[89] Klabbers,; E. A. M. *Segmental and prosodic improvements to speech generation.* PhD Thesis, Eindhovem: Technische Universiteit Eindhoven, 2000.

[90] Breiman, L.; Friedman, J. H.; Olsen, R. A.; Stone, C. J. *Classification and Regression Trees.* Chapman and Hall/CRC., Boca Raton, Florida, USA, 1998.

[91] Pessoa, L. de S. *Modelos da Língua para o Português do Brasil Aplicados ao Reconhecimento de Fala Contínua: Modelos Lineares e Modelos Hierárquicos (Parsing).* Dissertação de Mestrado, Faculdade de Engenharia Elétrica, Universidade Estadual de Campinas, Julho, 1999.

[92] Britto Jr. A. S.; Freitas, C. O. A.; Justino, E. J. R.; Borges, D. L.; Facon, F.; Bortolozzi, F.; and Sabourin R. *Técnicas em Processamento e Análise de Documentos Manuscritos, Revista de Informática Teórica e Aplicada (RITA).* Instituto de Informática, Vol. 8, $n^{\underline{o}}$ 1, pp. 47–68, UFRGS, Porto Alegre, 2001.

[93] Santos S. C.; Alcaim A. *Treinamento de Modelos de Unidades Fonéticas com Variabilidade Acústica em Reconhecedores de Voz Contínua Baseados em CDHMM.* Revista da Sociedade Brasileira de Telecomunicações, Vol. 15, $n^{\underline{o}}$ 1, pp. 34–43, Junho, 2000.

[94] Ynoguti, C. A.; Violaro, F. *Sobre a importância da Transcrição Fonética em Sistemas de Reconhecimento de Fala.* Revista da Sociedade Brasileira de Telecomunicações, Vol. 15, $n^{\underline{o}}$ 1, pp. 44–49, Junho, 2000.

[95] Quast, H.; Schreiner, O.; Schroeder, M.R. Robust pitch tracking in the car environment. *Proceedings of Acoustics, Speech, and Signal Processing,* pp. 353-356, Orlando, USA, May, 2002.

[96] Wang, Y.; Wong, I.; Tsao, T. A statistical pitch detection algorithm. *Proceedings of Acoustics, Speech, and Signal Processing*, pp. 357–360, Orlando, USA, May, 2002.

[97] Deshmukh, O.; Wilson, C.E. A measure of aperiodicity and periodicity in speech. *Proceedings of Acoustics, Speech, and Signal Processing*, pp. 448–451, USA, April, 2003.

[98] Tan, L.; Kochanski, G.; Shih, C.; Li, Y. 1. Modeling Tones in Continuous Cantonese Speech . *Proceedings of International Conference on Spoken Language Processing*, pp. 102–107, Denver, Colorado, September 2002 .

[99] Syrdal, A. K.; Hirshberg, J. Automatic ToBI Prediction and Alignment to Speed Manual Labeling of Prosody. *Speech Communication - Special Issue on Speech Annotation and Corpus Tools*, (33):135–151, USA, July, 2001.

[100] Ni, J.; Kawai, H. Tone feature extraction through parametric modeling and analysis-by-synthesis-based pattern matching. *Proceedings of Acoustics, Speech, and Signal Processing*, pp. 448–451, USA, April, 2003.

[101] Tams, A.; Tatham, M. Intonation for Synthesis of Speaking Styles. *IEE Seminar "State-Of-The-Art In Speech Synthesis"*, (Ref. No. 2000/058), 6:1–11 London, April, 2000.

[102] Mixdorff's, H. *Intonation Patterns of German - Quantitative Analysis and Synthesis of F_0 Contourns.* PhD Thesis, Technische Fachhochschule Berlin - University of Applied Sciences, Dresden, 1998.

[103] Mixdorff's, H. A Novel Approach to the Fully Automatic Extraction of Fujisaki Model Parameters. *Proceedings of International Conference on Acoustics, Speech, and Signal Processing*, 3:1281–1284, Instanbul, Turkey, 2000.

[104] Mixdorff, H.; Hu, Y.; Chen, G. Towards the Automatic Extraction of Fujisaki Model Parameters for Mandarin. *Proceedings of 8th European Conference on Speech Communication and Technology (Interspeech'2003)*, pp. 873–876, Geneva, Switzerland, 2003.

[105] Navas, E.; Hernáez, I.; Armenta, A.; Etxebarria, B.; Salaberria, J. Modelling Basque Intonation Using Fujisaki's Model and Carts. *IEE Seminar "State-Of-The-Art In Speech Synthesis"*, (Ref. No. 2000/058), 3:1–6 London, April, 2000.

[106] Wright, H. F. *Modelling Prosodic and Dialoge Information for automatic Speech Recognition.* PhD Thesis, University of Edinburg, England, 1998.

[107] Mussalim, F.; Bentes, A. C. *Introdução à Linguística: domínios e fronteiras.* Cortez Editora, São Paulo, 2001.

[108] Dusterhoff, K. E.; Black, A. W.; Taylor P. A. Using Decision Trees within the Tilt Intonation Model to Predict F0 Contours. *Proceedings of European Conference on Speech Communication and Technology*, 15:169-186, 1999.

[109] Delmonte, R.; Petrea, M.; Bacalu C. SLIM - Prosodic Module for Learning Activities. *Proceedings of European Conference on Speech Communication and Technology*, 2:669–672, Rhodes, Greece, 1997.

[110] Shih, C.; Kochanski, G. Modeling Intonation: Asking for Confirmation in English. *Proceedings of 15th International Congress of Phonetics Sciences*, Barcelona, Spain, August, 2003.

[111] Silva, C. H. da; Nagle, E. J.; Runstein, F.; Violaro, F. F_0 Generation in a Text-to-Speech System Using a Database of Natural F_0 Patterns. *Proceedings of the International Telecommunication Symposium*, 1:213–218, São Paulo, Brazil, August, 1998.

[112] Stylianou, Y.; Dutoit, T.; Schroeter, J. Diphones Concatenation Using a Harmonic plus Noise Model of Speech. *Proceedings of European Conference on Speech Communication and Technology*, pp. 613-616, Rhodes, Greece, 1997.

[113] Campbell, N.; Black, A. W. Prosody and the Selection of Source Units for Concatenative Synthesis. In: van Santen, J. P. H.; Sproat, R. W.; Olive, J. P.; Hirschberg, J. (Eds.). *Progress in Speech Synthesis.* Springer, pp. 279-292, 1997.

[114] Conkie, A. D.; Israd, S. *Optimal Coupling of Diphones.* In: van Santen, J. P. H.; Sproat, R. W.; Olive, J. P.; Hirschberg, J. (Eds.). *Progress in Speech Synthesis.* Springer, pp. 293-304, 1997.

[115] Hunt, A. J.; Black, A. W. Unit Selection in a Concatenative Speech Synthesis System Using a Large Speech Database. *Proceedings of International Conference on Acoustics, Speech, and Signal Processing*, pp. 373-376, 1996.

[116] Black, A. W.; Campbell, N. Optimising Selection of Units from Speech Databases for Concatenative Synthesis. *Proceedings of European Conference on Speech Communication and Technology*, pp. 581-584, Madrid, Spain, 1995.

[117] Narusawa, S.; Fujisaki, H.; Ohno, S. A Method for Automatic Extraction of Parameters of the Fundamental Frequency Contour. *Proceedings of International Conference on Spoken Language Processing*, Vol. 1, pp. 649–652, Beijing, China, 2000.

[118] Moulines, E.; Verhelst, W. *Time-Domain and Frequency-Domain Techniques for Prosodic Modification of Speech.* In: Kleijn, W. B.; Paliwal, K. K. (Eds.). Speech Coding and Synthesis. Amsterdam, Elsevier, pp. 519-555, 1995.

[119] Perkell, J. S.; Wilhelms-Tricarico, R. F. A Biomechanical and Physiologically Based Speech Modeling. In: van Santen, J. P. H.; Sproat, R. W.; Olive, J. P.; Hirschberg, J. (Eds.). *Progress in Speech Synthesis*. Springer, pp. 221-233, 1997.

[120] Backman, M. E. Speech Models and Speech Synthesis. In: van Santen, J. P. H.; Sproat, R. W.; Olive, J. P.; Hirschberg, J. (Eds.). *Progress in Speech Synthesis*. Springer, pp. 185-209, 1997.

[121] Prado, P. P. L. do. Sintetizador Articulatório de Voz: Mapeamento Acústico/Articulatório. *Anais do XIII Simpósio Brasileiro de Telecomunicações*, pp. 708–712, Natal, 1993.

[122] Guiard-Marigny, T.; Adjoudani, A.; Benoit, C. 3D Models of the Lips and Jaw for Visual Speech Synthesis. In: van Santen, J. P. H.; Sproat, R. W.; Olive, J. P.; Hirschberg, J. (Eds.). *Progress in Speech Synthesis*. Springer, pp. 247-258, 1997.

[123] Huang, J.; Levinson, S.; Davis, D.; Slimon, S. Articulatory speech synthesis based upon fluid dynamic principles. *Proceedings of Acoustics, Speech, and Signal Processing*, pp. 445–448, Orlando, USA, May, 2002.

[124] Donovan, R. E. *Trainable Speech Synthesis.* PhD Thesis, University of Cambridge, England, 1996.

[125] Hallahan, W. I. DECtalk Software: Text-to-Speech Technology and Implementation. *Digital Technical Journal of Digital Equipment Corporation*, 4(7):5–19, 1996.

[126] Bunnell, H. T.; Hoskins, S. R.; Yarrington, D. M. A Biphone Constrained Concatenation Method for Diphone Synthesis. *Proceedings of 5th International Conference on Spoken Language Processing*, pp. 171–176, Speech Synthesis Workshop, Sydney, Australia, 1998.

[127] Hirai, T.; Tenpaku, S.; Shikano, K. Speech Unit Selection Based on Target Values Driven by Speech Data in Concatenative Speech Synthesis. *Workshop on Speech Synthesis*, pp. 43–46, California, USA, September, 2002.

[128] Low, P. H.; Vaseghi, S. Synthesis of unseen context and spectral and pitch contour smoothing in concatenated text-to-speech synthesis. *Proceedings of Acoustics, Speech, and Signal Processing*, pp. 469–472, Orlando, USA, May, 2002.

[129] Kuo, W.; Zhong, X.; Wang, Y.; Chen, S. A Hight-Performance Min-Nan/Taiwanese TTS System. *Proceedings of Acoustics, Speech, and Signal Processing*, pp. 448–451, USA, April, 2003.

[130] Klabbers, E. A. M. *Segmental and Prosodic Improvements to Speech Generation.* PhD Thesis, Technische Universiteit Eindhoven, Netherlands, 2000.

[131] Kortekass, R. W.; Kohlrausch. Psychoacoustical Evaluation of the Pitch-Synchronous Overlap-and-Add Speech-Waveform Manipulation Technique Using Single-Format Stimuli. *Journal of the Acoustical Society of America*, 4(101):2202–2213, April, 1997.

[132] Kuo, C. C.; Kuo, C. S. Speech segment selection for concatenative synthesis based on prosody-aligned distance measure. *Proceedings of Acoustics, Speech, and Signal Processing*, pp. 473–476, Orlando, USA, May, 2002.

[133] Dorran, D.; Lawlor, R.; Coyle, E. High quality time-scale modification of speech using a peak alignment overlap-add algorithm (PAOLA). *Proceedings of Acoustics, Speech, and Signal Processing*, pp. 700–703, USA, April, 2003.

[134] Hamming, R. W. *Digital Filters*. Dover Science, 1998.

[135] Smith, S. W. *Digital Signal Processing*. California Technical Publishing. San Diego, California. Second Edition, 1999.

[136] Violaro, F. Processamento Digital de Sinais de Fala. *Anais do XV Simpósio Brasileiro de Teclecomunicações*, Minicurso, pp. 1 52, Recife, Setembro, 1997.

[137] Macon, M. W.; Clements, M. A. Speech Concatenation and Synthesis Using an Overlap-Add Sinusoidal Model. *Proceedings of International Conference on Acoustics, Speech, and Signal Processing*, pp. 361-364, 1996.

[138] Macon, M. W.; Jensen-Link, L.; Oliverio, J.; Clements, M. A.; George, E. B. A system for singing voice synthesis based on sinusoidal modeling. *Proceedings of Acoustics, Speech, and Signal Processing*, Vol. 1, pp. 435–438, 1997.

[139] Prudon, R.; d'Alessandro, C.; Mareil, P. B. Improving quality of Mbrola Synthesis for Non-Uniform Units Synthesis. *Workshop on Speech Synthesis*, California, USA, September, 2002.

[140] Toda, T.; Kawai, H.; Tsuzaki, M.; Shikano, K. Unit selection algorithm for Japanese speech synthesis based on both phoneme unit and diphone unit. *Proceedings of Acoustics, Speech, and Signal Processing*, pp. 465–468, Orlando, USA, May, 2002.

[141] Kleijn, K.; Paliwal, K. *Speech Coding and Synthesis*. Elsevier Science, The Netherlands, 1998.

[142] Aguiar Neto, B. G. *Processamento e Transmissão Digital de Voz*. Programa PCT - MOTOROLA, Centro de Ciências e Tecnologia, Universidade Federal da Paraíba, 2001.

[143] Violaro, F.; Böeffard, O. A Hybrid Model for Text-to-Speech Synthesis. *IEEE Transactions on Speech and Audio Processing*, 6(5):426–434, September, 1998.

[144] Sydral, A.; Stylianou, Y.; Garrison, L.; Conkie, A.; Schroeter, J. TD-PSOLA Versus Harmonic Plus Noise Model in Diphone Based Speech Synthesis. *Proceedings of International Conference on Acoustics, Speech, and Signal Processing*, 1:273-276, Seattle, 1998.

[145] Jilka, M.; Syrdal, A. K.; Conkie, A. D.; Kapilov, D. A. Effects on TTS quality of methods of realizing natural prosodic variations. *Proceedings of 15th International Congress of Phonetics Sciences*, pp. 2549–2552, Barcelona, Spain, August, 2003.

[146] Aquino, P. A. de. *O papel das vogais reduzidas pós-tônicas na construção de um sistema de síntese concatenativa para o português do Brasil.* Dissertação de Mestrado, Instituto de Estudos da Linguagem, Universidade Estadual de Campinas, 1997.

[147] Huang, X.; Acero, A.; Hon, H. *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development.* Prentice Hall, 2001.

[148] Costa Neto, M. L. da; Souza, M. A. T. F de; Aguiar Neto, B. G.; Figueiredo, F. A. de. Concepção de um Dicionário para um Sintetizador Texto-Fala Concatenativo. *XIII Simpósio Brasileiro de Telecomunicações (SBT'99)*, pp. 564–569, Vila Velha - ES, Setembro, 1999.

[149] Costa Neto, M. L. da; Sousa, M. A. T. F. de; Barbosa, S. G. D.; Aguiar Neto, B. G.; Bezerra, M. A. Dicionário para um Sintetizador Texto-Fala para a Língua Portuguesa. *IV Encontro para o Processamento Computacional de Português Escrito e Falado (PROPOR'99)*, pp. 155–166, Évora, Portugal, Setembro, 1999.

[150] Wells, J.; Barry, W.; Grice, M.; Fourcin, A.; Gibbon, D. Standard Computer Compatible Transcription. *SAM STAGE REPORT Sen.3, SAM-UCL-037*, ESPRIT Project 2589, February, 1992.

[151] Camara Jr., J. M. *Problemas da Lingüística Descritiva.* Editora Vozes Limitada, Petrópolis - RJ, 1997.

[152] Silva, C. H. da; Nagle, E. J.; Nunes, H. F. Automação da Construção de Dicionário de Unidades Acústicas para a Conversão Texto-Fala por Concatenação. *Anais do XV Simpósio Brasileiro de Telecomunicações*, pp. 323–327, Recife, Setembro, 1997.

[153] Toledano, D. T.; Gómez L. A. H.; Automatic Phonetic Segmentation. *IEEE Transactions on Speech and Audio Processing*, 6(11):617–624, November, 2003.

[154] Escudero, D.; Cardenõso, V. Corpus based extraction of quantitative prosodic parameters stress groups in spanish. *Proceedings of Acoustics, Speech, and Signal Processing*, pp. 481–484, Orlando, USA, May, 2002.

[155] Hwang, S.; Yei, C. The synthesis unit generation algorithm for maudarin TTS. *Proceedings of Acoustics, Speech, and Signal Processing*, pp. 457–460, Orlando, USA, May, 2002.

[156] Moraes, J. A. de. Um Algoritmo para a Correção-Simulação da Duração dos Segmentos Vocálicos em Português. *Estudos de Prosódia*, pp. 69-84. Editora da UNICAMP, Campinas - SP, 1999.

[157] Costa Neto, P. L. de O. *Estatística*. Editora Edgard Blucher, 2002.

[158] Levine, D. M.; Berenson, M. L.; Stephan, D. *Estatística: Teoria e Aplicações*. Livros Técnicos e Científicos S. A., 2000.