UNICAMP

Faculdade de Engenharia Elétrica e de Computação

DEPARTAMENTO DE COMUNICAÇÕES

LABORATÓRIO DE PROCESSAMENTO DE SINAIS

Sistema de conversão texto-fala para a língua portuguesa utilizando a abordagem de síntese por regras

Dissertação de mestrado

Autor: Leandro de Campos Teixeira Gomes

Orientador: Prof. Dr. José Geraldo Chiquito

Campinas – SP

Julho de 1998

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

G585s

Gomes, Leandro de Campos Teixeira
Sistema de conversão texto-fala para a língua portuguesa utilizando a abordagem de síntese por regras. / Leandro de Campos Teixeira Gomes.--Campinas, SP: [s.n.], 1998.

Orientador: José Geraldo Chiquito.
Dissertação (mestrado) - Universidade Estadual de Campinas, Faculdade de Engenharia Elétrica e de Computação.

1. Síntese da voz. 2. Sistemas de processamento da fala. I. Chiquito, José Geraldo. II. Universidade Estadual de Campinas. Faculdade de Engenharia Elétrica e de Computação. III. Título.

Universidade Estadual de Campinas
Faculdade de Engenharia Elétrica e de Computação

Sistema de conversão texto-fala para a língua portuguesa utilizando a abordagem de síntese por regras

Autor:

Leandro de Campos Teixeira Gomes

Banca Examinadora:

Prof. Dr. José Geraldo Chiquito – Presidente
DECOM-FEEC-UNICAMP

Prof. Dr. João Marcos Travassos Romano
DECOM-FEEC-UNICAMP

Prof. Dr. Maurílio Nunes Vieira
DF-ICEx-UFMG

Dissertação apresentada à Faculdade de Engenharia Elétrica e de Computação da Universidade Estadual de Campinas como requisito parcial para a obtenção do título de Mestre em Engenharia Elétrica.

Agradecimentos

Registro aqui meus sinceros agradecimentos às pessoas e às instituições cujo apoio foi decisivo para a concretização deste trabalho.

Ao professor José Geraldo Chiquito, sob cuja orientação venho desenvolvendo este trabalho há mais de quatro anos, agradeço pelo apoio, seriedade e amizade constantes. Seus valiosos conselhos, longe de limitarem-se a aspectos técnicos, vieram mesmo a definir a próxima etapa de minha formação acadêmica, com impacto decisivo sobre minha vida pessoal e profissional.

Não há como exagerar a importância do papel desempenhado por Edson José Nagle neste trabalho. Ao despertar meu interesse pelo campo da produção artificial de fala, convidando-me para a realização de um estágio de iniciação científica, Nagle deu início a uma gratificante colaboração que culminou nesta dissertação e, espero, está ainda longe de chegar ao fim. Verdadeiro co-autor e co-orientador deste trabalho, Nagle esteve presente em todas as suas etapas, contribuindo com idéias e sugestões, criticando as soluções adotadas e disponibilizando-se a esclarecer dúvidas. Mais do que um colega, Nagle é um amigo com o qual sempre pude contar.

Agradeço à equipe do Laboratório de Fonética Acústica e Psicolingüística Experimental (IEL/Unicamp) e em especial à professora Eleonora C. Albano, primeira orientadora deste trabalho e coordenadora do projeto temático no qual ele se insere. Agradeço também ao professor Ricardo Molina de Figueiredo, do Departamento de Medicina Legal da FCM/Unicamp, que tem utilizado os módulos de síntese do sistema de conversão texto-fala implementado e apresentou sugestões valiosas para o seu aperfeiçoamento.

Aos professores João Marcos T. Romano (FEEC/Unicamp) e Maurílio Nunes Vieira (ICEx/UFMG), agradeço por terem aceito o convite para participação na Banca Examinadora desta dissertação e pelas várias sugestões para o aprimoramento do texto final. Registro ainda meus agradecimentos à Coordenação de Pós-Graduação da FEEC/Unicamp, em especial ao professor Christiano Lyra Filho, pelo auxílio na solução das dificuldades burocráticas surgidas na fase final deste mestrado.

Aos meus pais, Vera e Eustáquio, e à minha irmã, Lalinka, agradeço pelo apoio constante ao longo de toda a minha vida, sem o qual esta e muitas outras realizações não teriam sido possíveis.

Finalmente, agradeço ao CNPq, ao FAEP, à FAPESP e à CAPES pelo suporte financeiro às várias fases deste trabalho.

Resumo

Neste trabalho encontra-se descrito o sistema de conversão texto-fala para o português do Brasil desenvolvido no Laboratório de Processamento de Sinais da Faculdade de Engenharia Elétrica e de Computação da Unicamp. O sistema recebe como entrada um texto genérico em português e produz em sua saída o sinal de fala correspondente. O processo de conversão texto-fala divide-se em três etapas básicas, cada uma englobando vários módulos:

- **Processamento de texto:** pré-processamento, classificação gramatical, divisão silábica e transcrição ortográfico-fonética.
- **Processamento prosódico:** determinação de fronteiras prosódicas, geração de contornos de entonação e geração de durações de segmentos.
- **Processamento de sinal:** síntese do sinal de fala utilizando o sintetizador de formantes de Klatt.

Os módulos de processamento prosódico empregam dados de duração e entonação extraídos de elocuções naturais, ajustando-os às particularidades do texto de entrada com base em informações provenientes do classificador gramatical. A abordagem de síntese por regras é utilizada para a geração dos parâmetros de controle do sintetizador. Uma linguagem e um compilador específicos foram criados para a descrição das regras de síntese. Embora não tenham sido realizadas avaliações formais da qualidade do sistema, testes informais indicaram um bom desempenho geral em termos de inteligibilidade e naturalidade.

Abstract

This work contains a description of the text-to-speech conversion system for the Portuguese of Brazil developed at the Signal Processing Laboratory of the Electrical and Computer Engineering School of Unicamp. The system receives as input a generic text in Portuguese and produces as output the corresponding speech signal. The text-to-speech conversion process is divided into three basic steps, each one including several modules:

- **Text processing:** preprocessing, grammatical classification, syllabic division and orthographic-phonetic transcription.
- **Prosodic processing:** determination of prosodic boundaries, generation of intonation patterns and generation of segmental durations.
- **Signal processing:** synthesis of the speech signal using the Klatt formant synthesizer.

The prosodic processing modules use duration and intonation data extracted from natural utterances, adjusting them to the particularities of the input text on the basis of information provided by the grammatical classifier. The synthesis-by-rule approach is used for generating the synthesizer control parameters. A specific language and a compiler have been created for the description of the synthesis rules. Although formal evaluations of the system quality have not been made, informal tests have indicated a good general performance in terms of intelligibility and naturalness.

Índice

Introdução 1

Capítulo 1 Sistemas de conversão texto-fala 4

1.1 Estrutura básica de um sistema de conversão texto-fala 4
1.2 Evolução dos sistemas de conversão texto-fala 5
1.3 Visão geral do sistema implementado 7
1.4 Operação do sistema 8
1.5 Módulo de controle do sistema 13

Capítulo 2 Processamento de texto 15

2.1 Pré-processamento de texto 15
2.2 Classificação gramatical 21
2.3 Divisão silábica e determinação da sílaba tônica lexical 28
2.4 Transcrição ortográfico-fonética 29

Capítulo 3 Processamento prosódico 36

3.1 Determinação de fronteiras prosódicas 36
3.2 Geração de contornos de entonação 38
3.3 Geração de durações de segmentos 45
3.4 Formato de saída do processador prosódico 50

Capítulo 4 Síntese do sinal de fala 52

4.1 Técnicas para síntese de fala 52
4.2 O sintetizador de formantes de Klatt 53
4.3 Modelos para as classes de fones 56
4.4 Linguagem para descrição de regras 61
4.5 O módulo conversor 79

Capítulo 5 Considerações finais 84

5.1 Principais contribuições 84
5.2 Avaliação do desempenho geral do sistema 86
5.3 Possibilidades de estudos futuros 87

Apêndice A Regras utilizadas na etapa de processamento de texto 91

Apêndice B Regras de síntese e arquivo de dados para síntese 100

Referências bibliográficas 106

Introdução

Este trabalho descreve o sistema de conversão texto-fala para a língua portuguesa desenvolvido no Laboratório de Processamento de Sinais (LPS) da Faculdade de Engenharia Elétrica e de Computação da Unicamp. O sistema, que roda em microcomputadores PC, divide-se em três componentes básicos: (1) processamento de texto, incluindo módulos para pré-processamento, classificação gramatical, divisão silábica e transcrição ortográfico-fonética; (2) processamento prosódico, incluindo módulos para determinação de fronteiras prosódicas, geração de padrões de entonação e obtenção de durações de segmentos; e (3) processamento de sinal, utilizando o sintetizador de formantes de Klatt e a abordagem de síntese por regras para a geração dos parâmetros de controle do sintetizador. A partir de um texto genérico em português no formato ASCII estendido, o sistema gera um sinal de fala que pode ser armazenado em arquivo ou reproduzido imediatamente após a síntese. Nesta introdução, são discutidas a importância dos sistemas de conversão texto-fala e a sua rápida disseminação em ambientes computacionais. É também apresentada a estrutura geral da dissertação, comentando-se brevemente cada um de seus capítulos.

Difusão dos sistemas de síntese e reconhecimento de fala

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 ao longo da vida. A própria escrita nada mais é do que um registro simbólico da fala. Embora a escrita constitua um meio eficiente para o armazenamento e a transmissão de informações, a comunicação através da linguagem falada é com freqüência mais conveniente, permitindo a rápida assimilação do conteúdo de uma mensagem sem a necessidade de que os olhos se voltem para um determinado ponto.

Os computadores atuais apresentam como uma de suas principais características a capacidade para manipulação de grande quantidade de símbolos; é natural, portanto, que a linguagem escrita tenha predominado nas interfaces computador-usuário. O rápido aumento na capacidade de processamento dos computadores e os avanços nos campos da síntese e do reconhecimento de fala, no entanto, vêm tornando as interfaces baseadas na linguagem falada progressivamente mais acessíveis. Sistemas de conversão texto-fala de alto desempenho para a língua inglesa encontram-se disponíveis comercialmente há vários anos, enquanto sistemas de reconhecimento de fala comerciais, também para a língua inglesa,

apresentam eficiência crescente aliada a baixo custo. Previsões indicam que no início do próximo século mais da metade das principais companhias americanas fará uso de tecnologias de síntese ou reconhecimento de fala [Comerford et al. 1997]. Embora sistemas de síntese e reconhecimento para o português ainda não apresentem a variedade e a qualidade dos produtos disponíveis para o inglês, intensa atividade de pesquisa vem sendo conduzida nesse sentido, impulsionada pelo grande potencial do mercado brasileiro.

Algumas aplicações

A seguir são comentadas algumas aplicações típicas dos sistemas de conversão texto-fala.

- **Serviços de atendimento automático por telefone:** A comunicação entre uma pessoa e um computador através de uma linha telefônica é talvez a aplicação de maior potencial imediato dos sistemas de conversão texto-fala, permitindo o fornecimento automático de informações como saldo bancário, previsão do tempo, noticiários etc. A maioria dos sistemas de atendimento automático atualmente em operação utiliza técnicas de síntese por concatenação de palavras ou frases previamente gravadas, obtendo um resultado em geral muito pobre em termos de naturalidade. O uso de conversores texto-fala de alta qualidade produziria resultados sensivelmente superiores, além de evitar a realização de novas gravações a cada alteração no vocabulário do sistema.
- **Leitura de mensagens enviadas por correio eletrônico:** A rápida expansão no uso do correio eletrônico deu origem a outro mercado para os sistemas de conversão texto-fala. Um profissional que se encontre em trânsito e não disponha de um computador portátil poderia telefonar para seu provedor de acesso à Internet e ouvir suas mensagens, lidas por um computador utilizando um sistema de conversão texto-fala, sem a intermediação de um operador humano. Conforme as técnicas de reconhecimento de fala se tornem mais confiáveis, passará a ser possível também o envio de mensagens utilizando este método.
- **Auxílio a deficientes visuais:** Uma aplicação clássica dos sistemas de conversão texto-fala é o seu emprego como máquinas de leitura para deficientes visuais. Sistemas deste tipo em geral operam acoplados a um *scanner* e a um *software* de OCR para reconhecimento de texto impresso; estes acessórios, no entanto, vêm se tornando menos necessários conforme aumenta a quantidade de textos diretamente acessíveis aos computadores, sobretudo com a expansão da Internet.

- **Auxílio a indivíduos incapacitados para a fala:** Sistemas de conversão texto-fala podem ser utilizados como meio de expressão por indivíduos incapacitados para a fala. Para esta aplicação, é necessário que o sistema opere em um microcomputador portátil e que a interface com o usuário permita a construção de sentenças de maneira simples e eficiente.

Estrutura da dissertação

A dissertação está organizada em cinco capítulos e dois apêndices.

No capítulo 1 é apresentada uma descrição genérica dos sistemas de conversão texto-fala, incluindo um breve histórico, seguida de comentários sobre a estrutura e a operação do sistema implementado.

O capítulo 2 trata dos módulos de processamento de texto do sistema: o pré-processador, o classificador gramatical, o divisor silábico e o transcritor ortográfico-fonético.

No capítulo 3 são apresentados os módulos de processamento prosódico: o determinador de fronteiras prosódicas, o gerador de contornos de entonação e o gerador de durações de segmentos.

O capítulo 4 apresenta os módulos de processamento de sinal: o sintetizador de formantes de Klatt, responsável pela síntese do sinal de fala, e o módulo conversor, responsável pela obtenção dos parâmetros de controle do sintetizador. São também apresentadas as regras de síntese, bem como a linguagem e o compilador que permitem descrevê-las.

O capítulo 5 destaca as principais contribuições deste trabalho e discute o desempenho geral do sistema. São ainda relacionadas algumas possibilidades de estudos futuros.

Nos apêndices A e B estão reunidos algoritmos e tabelas empregados no sistema. Ao final da dissertação, encontram-se listadas as referências bibliográficas completas.

CAPÍTULO 1

Sistemas de conversão texto-fala

Este capítulo apresenta uma descrição genérica dos sistemas de conversão texto-fala, seguida de um breve histórico da sua evolução. É também apresentada a estrutura do sistema implementado, descrevendo-se resumidamente cada um de seus módulos. Por fim, é comentada a operação do módulo de controle do sistema e é detalhada a sua interface com o usuário.

1.1 ESTRUTURA BÁSICA DE UM SISTEMA DE CONVERSÃO TEXTO-FALA

Sistemas de conversão texto-fala têm por função gerar um sinal de fala a partir de um texto genérico. A meta principal destes sistemas é a maximização da inteligibilidade da fala sintetizada. É também desejável que a fala produzida soe natural ao ouvinte, isto é, que o seu caráter sintético não seja fortemente perceptível.

A estrutura mais comum para um sistema de conversão texto-fala está esquematizada na figura 1.1, mostrando as três etapas básicas presentes nestes sistemas: processamento de texto, processamento prosódico e processamento de sinal.

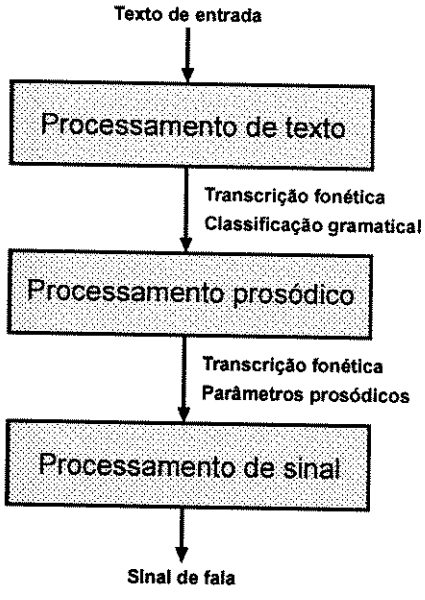

Figura 1.1 Estrutura de um sistema de conversão texto-fala genérico, mostrando as etapas de processamento de texto, processamento prosódico e processamento de sinal; são indicadas também a entrada geral do sistema e as saídas de cada etapa.

A etapa de processamento de texto é responsável pela conversão do texto de entrada em uma transcrição fonética que represente de forma suficientemente acurada os sons a sintetizar. A obtenção dessa transcrição envolve vários passos. Em primeiro lugar, é necessário efetuar um pré-tratamento sobre o texto de entrada, convertendo siglas, abreviações, símbolos e algarismos em suas correspondentes formas extensas. Em seguida, deve-se submeter o texto a uma análise gramatical, associando-se uma classificação a cada palavra. De posse dessas informações, pode-se efetuar a transcrição fonética propriamente dita.

Na etapa de processamento prosódico, são atribuídos padrões de entonação às frases sintetizadas e são determinadas durações para cada fone (denominadas *durações de segmentos*); eventualmente, é tratada também a intensidade (energia) dos sons. Esta etapa é de importância fundamental para conferir naturalidade à fala sintetizada. Informações provenientes da análise gramatical geralmente são utilizadas para a seleção de características prosódicas adequadas às frases do texto de entrada.

A etapa de processamento de sinal conclui o processo de conversão texto-fala, realizando a síntese do sinal de fala a partir da transcrição fonética acrescida de informações prosódicas.

Cada uma das etapas descritas acima pode ser implementada através de técnicas com diferentes graus de complexidade. Sistemas de conversão texto-fala mais sofisticados podem possuir módulos adicionais, assim como sistemas mais simples podem não possuir alguns dos módulos citados.

1.2 EVOLUÇÃO DOS SISTEMAS DE CONVERSÃO TEXTO-FALA [1]

Os primeiros sistemas para produção artificial de fala surgiram já nas décadas iniciais deste século. Data de 1922 o mais antigo sintetizador de formantes elétrico de que se tem registro. Em 1939, foi desenvolvido nos laboratórios Bell um sintetizador de fala eletromecânico, denominado *Voder,* controlado através de pedais e de um teclado semelhante ao de um piano. Embora apresentasse baixa inteligibilidade, o *Voder* deixou claro o potencial das pesquisas no ramo da síntese de fala.

A larga disseminação dos computadores eletrônicos digitais nas instituições de pesquisa a partir dos anos 60 teve importância fundamental para os trabalhos relacionados à conversão texto-fala. Tornou-se possível a implementação em *software* de sistemas que, tendo como entrada um texto genérico, controlavam um sintetizador analógico externo ao computador. O enorme avanço na capacidade de armazenamento e de processamento dos computadores nas décadas de 70 e 80 tornou possível a implementação dos próprios sinte-

[1] Uma revisão histórica detalhada dos sistemas de conversão texto-fala para a língua inglesa encontra-se em [Klatt 1987].

tizadores em *software,* restringindo os equipamentos requeridos pelos sistemas de conversão texto-fala, além do próprio computador, a um conversor digital-analógico, um amplificador e um alto-falante.

A teoria acústica da produção da fala proposta por Fant [Fant 1960] constituiu um grande avanço no campo da síntese de fala. Esta teoria representa matematicamente o processo de produção da fala e forneceu a base para a elaboração de diversos modelos empregados na implementação de sintetizadores. Em particular, o sintetizador de formantes de Klatt, utilizado no sistema de conversão texto-fala descrito neste trabalho, modela o processo de produção da fala através de um sistema linear com três componentes principais: fontes de vozeamento e ruído, característica de filtragem do trato vocal e característica de irradiação para o meio externo. A seção 4.2 apresenta uma breve descrição deste sintetizador.

Outra abordagem para o processo de síntese consiste na concatenação de trechos de fala natural para a produção de elocuções mais complexas. Técnicas desse tipo são utilizadas há várias décadas; sistemas baseados em concatenação de palavras ou frases para o fornecimento automático de informações por telefone, por exemplo, encontravam-se disponíveis já na década de 30. Sistemas mais versáteis podem ser obtidos utilizando unidades de concatenação menores, como semi-sílabas (trecho de fala que vai do meio de uma sílaba ao meio da sílaba seguinte) e difones (trecho de fala que vai do meio de um fone ao meio do fone seguinte). O emprego de difones como unidade de concatenação em sistemas de síntese foi inicialmente proposto em 1958. Mais recentemente, o advento de técnicas que permitem variar a duração e a freqüência fundamental de trechos curtos de fala pré-gravada tornou viável a implementação de sintetizadores de alta qualidade baseados na concatenação de difones.

Outros aspectos do processo de conversão texto-fala, como o processamento de texto e o processamento prosódico, vêm também sendo estudados há vários anos. Trabalhos iniciais relativos ao comportamento da freqüência fundamental para a língua inglesa datam da década de 40. Nas décadas de 60 e 70, foram desenvolvidos vários conjuntos de regras para a geração automática de durações de segmentos, também com foco na língua inglesa.

O campo da conversão texto-fala para a língua portuguesa permanece ainda relativamente pouco explorado, embora vários trabalhos abrangendo as diversos etapas do processo tenham sido apresentados nos últimos anos. A demanda do mercado por sistemas de conversão texto-fala de alta qualidade, aliada ao interesse demonstrado por empresas nacionais e estrangeiras, permite prever um rápido desenvolvimento desse campo no futuro próximo.

1.3 VISÃO GERAL DO SISTEMA IMPLEMENTADO

A estrutura do sistema de conversão texto-fala apresentado neste trabalho é muito similar à estrutura genérica descrita na seção 1.1[2]. A figura 1.2 ilustra os módulos que o compõem.

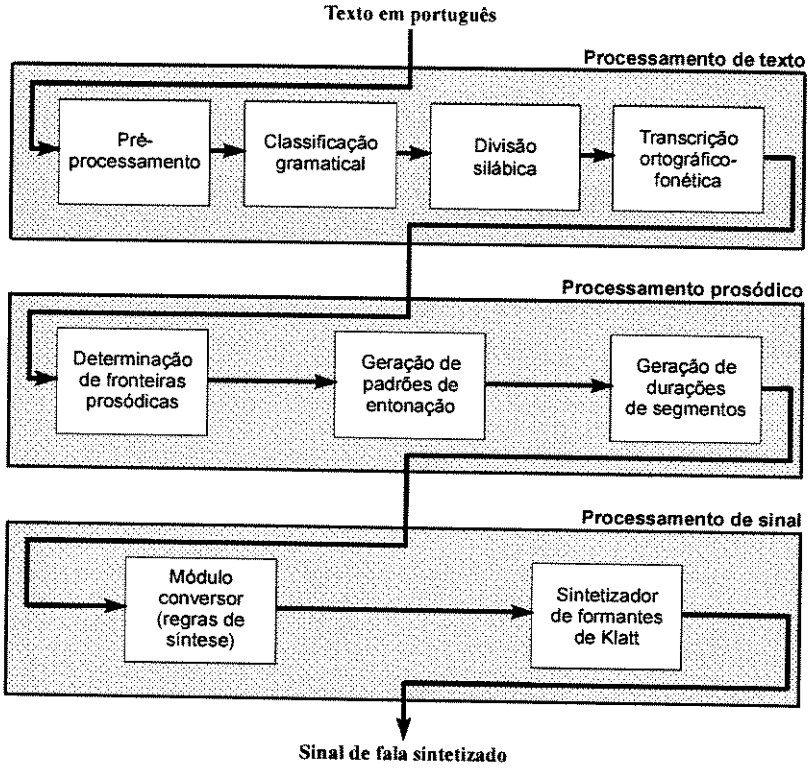

Figura 1.2 Estrutura do sistema de conversão texto-fala implementado, ilustrando as três etapas principais do processo e os módulos que as compõem.

A etapa de processamento de texto subdivide-se em quatro módulos:

- **Módulo de pré-processamento de texto (arquivo "PREPROC.CPP"):** responsável pelo tratamento de siglas, abreviações, símbolos e algarismos.
- **Módulo de classificação gramatical (arquivo "CLAGRAM.CPP"):** responsável pela obtenção da classe gramatical associada a cada palavra do texto de entrada.
- **Módulo para divisão silábica e determinação de sílabas tônicas (arquivo "DIVSIL.CPP"):** responsável pela localização de fronteiras entre sílabas e pela determinação de sílabas tônicas lexicais.
- **Módulo de transcrição ortográfico-fonética (arquivo "TRANSFON.CPP"):** responsável pela conversão do texto de entrada em uma seqüência de símbolos fonéticos.

[2] Uma descrição resumida do sistema apresentado neste trabalho encontra-se em [Gomes et al. 1998].

A etapa de processamento prosódico é constituída por três módulos:

- **Módulo para determinação de fronteiras prosódicas (arquivo "PROSOD.CPP"):** responsável pela quebra de frases em unidades menores, denominadas *grupos prosódicos,* tratadas separadamente pelos demais módulos desta etapa.
- **Módulo para geração de contornos de entonação (arquivo "PROSOD.CPP"):** responsável pela geração de contornos de freqüência fundamental (entonação) para os grupos prosódicos.
- **Módulo para geração de durações para os sons (arquivo "PROSOD.CPP"):** responsável pela geração de durações para cada fone presente nos grupos prosódicos.

Finalmente, a etapa de processamento de sinal subdivide-se em dois módulos:

- **Módulo conversor (arquivo "CONV.CPP"):** responsável pela obtenção de valores para os parâmetros de controle do sintetizador de Klatt a partir da transcrição fonética e da saída dos módulos de processamento prosódico.
- **Sintetizador de Klatt (arquivo "CONV.CPP"):** responsável pela síntese da forma de onda do sinal de fala.

Um módulo adicional (arquivo "SISCON.CPP") é responsável pelo controle do sistema, realizando a interface com o usuário e acionando os demais módulos na seqüência apropriada. Rotinas de uso comum de todos os módulos localizam-se no arquivo "GERAL.CPP".

Além dos módulos listados acima, foi implementado um programa, denominado *compilador de regras* (arquivo "CPREG.CPP"), cuja função se encontra descrita em detalhe na seção 4.4.

1.4 OPERAÇÃO DO SISTEMA

O sistema foi implementado inteiramente em linguagem C e roda em microcomputadores PC em ambiente DOS. Para a reprodução da fala sintetizada, utiliza-se uma placa de som padrão Sound Blaster ou o módulo CSL (*Computerized Speech Lab.*)[3] . Este último é um equipamento da Kay Elemetrics que opera ligado a um microcomputador e é utilizado também para captura e análise de sinais de fala.

[3] Para a reprodução do sinal de fala no CSL, foi utilizada uma biblioteca de rotinas fornecida pela Kay Elemetrics [STR 1994].

Os seguintes arquivos devem estar presentes em um mesmo diretório para que o sistema opere de forma adequada:

- **"SISCON.EXE":** arquivo executável que contém todos os módulos do sistema de conversão texto-fala.
- **"SISCON.CFG":** arquivo de configuração do sistema, descrito em detalhe mais adiante nesta seção; se não for encontrado, é utilizada a configuração *default*.
- **"PREPROC.SUB"**: arquivo-texto que contém dados utilizados pelo módulo de pré-processamento de texto (descrito no capítulo 2) e necessário para a sua operação.
- **"CLAGRAM.CLA":** arquivo-texto que contém dados utilizados pelo módulo de classificação gramatical (descrito no capítulo 2) e necessário para a sua operação.
- **"TRANSFON.EXC":** arquivo-texto que contém o dicionário de exceções do transcritor ortográfico-fonético (descrito no capítulo 2); se não for encontrado, o módulo opera sem um dicionário de exceções.
- **"PROSOD.DIC":** arquivo-texto que contém os dicionários de duração e de contornos de entonação (descritos no capítulo 3), necessário para a etapa de processamento prosódico.
- **"DADOS.TAB":** arquivo de dados para síntese (descrito no capítulo 4), necessário para a operação do módulo conversor.
- **"REGRAS.BIN":** versão compilada das regras de síntese (descritas no capítulo 4), necessária para a operação do módulo conversor.
- **"PLAY.EXE":** utilitário responsável pela reprodução de um sinal de fala em uma placa de som Sound Blaster; se não presente, somente a reprodução no CSL estará operacional.

Inserindo-se o diretório que contém estes arquivos na lista de trajetórias consultadas automaticamente pelo DOS para a busca de programas (através do comando PATH do DOS), o sistema de conversão texto-fala pode ser executado a partir de qualquer diretório. Não é necessária para a operação do sistema a presença do arquivo executável do compilador de regras ("CPREG.EXE") no diretório que contém os demais arquivos citados; este programa somente é requerido para a geração da versão compilada das regras, conforme descrito no capítulo 4. Os nomes dos arquivos "PREPROC.SUB", "CLAGRAM.CLA", "TRANSFON.EXC", "PROSOD.DIC", "DADOS.TAB" e "REGRAS.BIN" podem ser alterados através do arquivo de configuração "SISCON.CFG", descrito mais adiante.

O acionamento do sistema é feito através da linha de comando do DOS, obedecendo ao seguinte formato (colchetes indicam que um elemento não é obrigatório):

```
SISCON [opções] [nome de arquivo]
```

As opções do programa são especificadas no formato Unix, isto é, através de uma letra precedida por um hífen. Múltiplas opções podem ser fornecidas após um único hífen; por exemplo, as três formas de acionamento mostradas a seguir são equivalentes:

```
SISCON -k -f -p
SISCON -kf -p
SISCON -kfp
```

Nos três casos, as opções "k", "f" e "p" são ativadas. É indiferente o uso de letras maiúsculas ou minúsculas nas opções.

As opções disponíveis no sistema são[4]:

- **a:** mostra uma tela de ajuda sobre a operação do sistema.
- **c:** aciona o modo de reprodução no CSL (*default*) e define o formato NSP (comentado no capítulo 4) para o arquivo de forma de onda de saída.
- **d:** ativa a geração do arquivo de depuração (descrito no capítulo 4) e especifica quais parâmetros de controle do sintetizador devem ser rastreados; opcionalmente, pode ser fornecido um nome para este arquivo (extensão *default* ".DEP").
- **e:** indica em qual módulo deve ser encerrada a execução do sistema (*default*: módulo conversor e sintetizador de Klatt).
- **f:** indica que o texto a sintetizar deve ser obtido a partir de um arquivo, cujo nome pode ser opcionalmente especificado (extensão *default* ".TXT"); quando a execução do sistema tem início em um módulo que não o pré-processador, a opção "f" se aplica a esse módulo.
- **g:** indica que a saída de um determinado módulo deve ser registrada em arquivo.
- **h:** reproduz no CSL ou em uma placa de som padrão Sound Blaster um arquivo de forma de onda armazenado em disco, cujo nome pode ser opcionalmente especificado (extensão *default* ".NSP" para CSL e ".WAV" para Sound Blaster).
- **i:** indica a partir de qual módulo deve ser iniciada a execução do sistema (*default*: pré-processador de texto).

[4] Diversos termos e siglas aqui citados serão definidos ao longo dos capítulos seguintes.

- **k:** gera um arquivo de parâmetros de controle do sintetizador no formato original definido por Klatt (descrito no capítulo 4); o nome deste arquivo pode ser opcionalmente fornecido (extensão *default* ".KLA").
- **o:** desabilita a geração do arquivo de forma de onda (habilitada por *default*).
- **p:** gera um arquivo de parâmetros no formato compacto (descrito no capítulo 4); o nome deste arquivo pode ser opcionalmente fornecido (extensão *default* ".PAR").
- **r:** reproduz o sinal de fala no CSL ou em uma placa de som padrão Sound Blaster.
- **s:** aciona o modo de reprodução em placa de som padrão Sound Blaster e define o formato Wave (comentado no capítulo 4) para o arquivo de forma de onda de saída.
- **t:** realiza a síntese a partir de um arquivo de parâmetros no formato compacto (descrito no capítulo 4); o nome deste arquivo pode ser opcionalmente fornecido (extensão *default* ".PAR").
- **u:** determina um único módulo do sistema que deve ser executado, desabilitando a execução dos demais.
- **v:** realiza a síntese a partir de um arquivo de parâmetros no formato original definido por Klatt (descrito no capítulo 4); o nome deste arquivo pode ser opcionalmente fornecido (extensão *default* ".KLA").

Para as opções que aceitam a especificação opcional de um nome de arquivo ("d", "f", "h", "k", "p", "t" e "v"), este nome deve ser fornecido entre colchetes imediatamente após a opção; por exemplo:

```
SISCON -f[TEXTO]p[PARAM]
```

As opções mostradas acima especificam que o texto de entrada do sistema será obtido do arquivo "TEXTO.TXT" (aplicando-se a extensão *default* ".TXT") e será gerado o arquivo de parâmetros no formato compacto "PARAM.PAR" (aplicando-se a extensão *default* ".PAR"). Se os nomes fornecidos contivessem extensões, elas prevaleceriam sobre as extensões *default.*

A opção "d" permite a especificação de uma lista de parâmetros a rastrear, conforme descrito em detalhe no capítulo 4; essa lista é fornecida entre chaves, como no exemplo a seguir:

```
SISCON -d[DEPUR]{AV,F1,B2}
```

Nesse exemplo, é especificado o nome "DEPUR.DEP" para o arquivo de depuração e são selecionados os parâmetros AV, F1 e B2 para rastreamento. Os nomes dos parâmetros po-

dem ser separados por qualquer um dos seguintes caracteres: ",", ".", "–", "_" ou "/". Se essa lista é omitida, são rastreados por *default* os parâmetros AV, AF e F1.

O nome de arquivo especificado opcionalmente após as opções de linha de comando é utilizado para o arquivo de forma de onda de saída (inserindo-se as extensões ".NSP" ou ".WAV" caso nenhuma extensão seja fornecida). Além disso, ele é utilizado para os arquivos cujos nomes não foram fornecidos juntamente com as opções de linha de comando, inserindo-se a extensão *default* para cada tipo de arquivo. Por exemplo:

```
SISCON -fps TEXTO
```

Neste exemplo, o sistema obtém o texto de entrada do arquivo "TEXTO.TXT" e gera o arquivo de parâmetros no formato compacto "TEXTO.PAR" e o arquivo de forma de onda "TEXTO.WAV" (com extensão ".WAV" devido à seleção do modo de reprodução em placa de som padrão Sound Blaster através da opção "s"). Caso o nome de um arquivo não seja fornecido juntamente com a opção correspondente e nem após as opções, o sistema solicita ao usuário que o forneça através do teclado.

As opções "u", "i", "e" e "g" requerem a especificação de um módulo do sistema, indicado através de um algarismo posicionado imediatamente após a opção. A correspondência estabelecida entre módulos e algarismos segue a ordem de execução do sistema: 1 – pré-processador de texto; 2 – classificador gramatical; 3 – divisor silábico; 4 – transcritor ortográfico-fonético; 5 – determinador de fronteiras prosódicas; 6 – geração de durações de segmentos e de contornos de freqüência fundamental; 7 – módulo conversor e sintetizador de Klatt. As opções "i" e "e" utilizadas em conjunto permitem que seja definido um grupo de módulos a executar, enquanto a opção "u" determina a execução de um único módulo. Quando a execução do sistema é encerrada em um módulo que não o conversor e o sintetizador de Klatt, a saída do último módulo executado é registrada em arquivo. Pode-se também registrar em arquivo a saída de um módulo qualquer através da opção "g". Na versão atual do sistema, estes arquivos têm nomes fixos para cada módulo, exceto para o módulo conversor:

- **Pré-processador de texto:** "PREPROC.SPT"
- **Classificador gramatical:** "CLAGRAM.SCG"
- **Divisor silábico:** "DIVSIL.SDS"
- **Transcritor ortográfico-fonético:** "TRANSFON.STF"
- **Determinador de fronteiras prosódicas:** "FRONPRO.SFP"
- **Gerador de durações segmentais e contornos de entonação:** "PROSOD.SPR"

Se a opção "g" é aplicada ao módulo conversor, ela determina que seja gerado o arquivo de forma de onda mesmo que a sua geração tenha sido previamente inibida através da opção "o".

No momento, apenas o pré-processador e o módulo conversor estão preparados para receber uma entrada fornecida diretamente pelo usuário; os demais módulos devem necessariamente receber informações provenientes dos módulos que os antecedem. As opções "i" e "u", portanto, somente podem ser aplicadas aos módulos 1 (pré-processador) e 7 (conversor). Quando o primeiro módulo a executar é o conversor, a sua entrada pode ser obtida a partir do teclado (*default*) ou através de um arquivo (caso seja especificada a opção "f") com extensão *default* ".SPR".

Através do arquivo de configuração ("SISCON.CFG"), é possível utilizar nomes diferentes do *default* para os arquivos de dados lidos pelos módulos do sistema. Esses nomes devem ser especificados no arquivo, um por linha, na seguinte ordem: arquivo de dados do pré-processador de texto (nome *default* "PREPROC.SUB"), arquivo de dados do classificador gramatical (nome *default* "CLAGRAM.CLA"), arquivo de dados do transcritor ortográfico-fonético (nome *default* "TRANSFON.EXC"), arquivo que contém os dicionários de duração e de contornos de freqüência fundamental (nome *default* "PROSOD.DIC"), arquivo de dados para síntese (nome *default* "DADOS.TAB") e arquivo de regras compilado (nome *default* "REGRAS.BIN"). Para manter-se o nome *default* de um arquivo, especifica-se um hífen ("-") em seu lugar. Outra função do arquivo de configuração é a definição de opções acionadas de forma automática pelo sistema; essas opções seguem o mesmo padrão descrito para as opções fornecidas diretamente pela linha de comando do DOS e são listadas em uma única linha após os nomes dos arquivos de dados. Este recurso é útil, por exemplo, quando o sistema é operado em um microcomputador que não possui o módulo CSL mas possui uma placa de som padrão Sound Blaster; especificando-se a opção "s" (que seleciona a placa de som como dispositivo de reprodução) no arquivo de configuração, não é necessário fornecê-la sempre que o sistema for acionado. Podem ser inseridos comentários no arquivo de configuração, devendo-se precedê-los pelo símbolo "%".

1.5 MÓDULO DE CONTROLE DO SISTEMA

O módulo de controle do sistema (arquivo "SISCON.CPP") coordena o processo de conversão texto-fala, acionando cada um dos demais módulos na seqüência adequada. Ele é também responsável pela interface com o usuário, tratando as opções e os parâmetros de linha de comando (fornecidos ao acionar o sistema a partir do DOS) e solicitando nomes

de arquivos. O módulo de controle é ainda responsável pela leitura e tratamento do arquivo de configuração ("SISCON.CFG").

Uma função adicional desempenhada pelo módulo de controle é a quebra do texto de entrada em frases. O final de uma frase é definido pela presença de um dos seguintes caracteres: ponto final ("."), ponto e vírgula (";"), ponto de exclamação ("!"), ponto de interrogação ("?"), reticências ("..."), dois pontos (":"), travessão (representado por um hífen dobrado: "--"), retorno de carro (CR) e mudança de linha (LF). As informações do caractere terminador e da posição da frase no parágrafo (início, meio ou fim) são fornecidas aos demais módulos do sistema, pois podem ter influência na atribuição de características prosódicas adequadas à frase. O final de um parágrafo é definido pelos caracteres de retorno de carro (CR) ou de mudança de linha (LF).

Como o caractere de ponto final (".") é utilizado também após abreviações e em meio a siglas, foram adotados os seguintes critérios para a definição de seu significado:

- Se o ponto está inserido em meio a uma palavra que contém um ponto após cada letra, ele é considerado um indicador de sigla. Esta regra não se aplica ao último ponto de uma sigla, já que este pode desempenhar também a função de finalizador de frase.
- Se o ponto é seguido por uma palavra cuja primeira letra é maiúscula, ele é considerado um finalizador de frase. Exceções a esta regra são as abreviações normalmente seguidas por nomes próprios (em geral iniciados com letra maiúscula), como "Sr.", "Dr.", "Prof." etc.; nestes casos, o ponto é considerado um simples indicador de abreviação.
- Se o ponto é seguido por uma palavra cuja primeira letra é minúscula, ele é considerado um indicador de abreviação.
- Se o ponto está localizado após a última palavra do parágrafo, ele é considerado um terminador de frase.

Embora não cubram todos os casos possíveis, estes critérios na maioria das vezes levam à atribuição da função correta ao caractere de ponto final.

Cada frase do texto de entrada é tratada seqüencialmente pelos módulos de pré-processamento de texto, classificação gramatical, transcrição ortográfico-fonética, determinação de fronteiras prosódicas, geração de contornos de freqüência fundamental e geração de durações de segmentos. Os resultados do processamento de cada frase são registrados cumulativamente em um arquivo temporário utilizado como entrada do módulo conversor, que gera a forma de onda completa do sinal de fala. Após a execução do módulo conversor, o sinal de fala pode ser reproduzido no CSL ou em uma placa de som padrão Sound Blaster.

CAPÍTULO 2

Processamento de texto

Este capítulo descreve os módulos que constituem a etapa de processamento de texto do sistema de conversão texto-fala implementado: o pré-processador de texto, o classificador gramatical, o divisor silábico e o transcritor ortográfico-fonético. Todo o processamento descrito assume como entrada uma única frase em português; o módulo de controle do sistema, conforme discutido no capítulo anterior, é responsável pela quebra do texto de entrada em frases que são tratadas isoladamente pelos módulos de processamento de texto e de processamento prosódico.

2.1 PRÉ-PROCESSAMENTO DE TEXTO

A necessidade de um pré-processador de texto

Além dos símbolos ortográficos do alfabeto latino (eventualmente acompanhados de sinais diacríticos ou cedilha), um texto genérico em português pode apresentar uma grande variedade de caracteres, tais como sinais de pontuação, aspas, algarismos e diversos símbolos especiais ("%" e "&", por exemplo). Mesmo os símbolos do alfabeto latino podem ser usados em configurações particulares que impossibilitam a sua leitura direta; este é o caso, por exemplo, de abreviações, cuja leitura exige a sua expansão em um ou mais termos que devem ser de conhecimento prévio do leitor. Como um sistema de conversão texto-fala de finalidade geral deve, em princípio, tratar satisfatoriamente qualquer texto em português, é necessário criar mecanismos que lhe permitam interpretar corretamente tais ocorrências.

Esses mecanismos são providos pelo *pré-processador de texto* (contido no arquivo "PREPROC.CPP"), que é o módulo inicial do sistema de conversão texto-fala implementado e é responsável pela conversão do texto de entrada em um formato apropriado aos módulos seguintes. As principais tarefas desempenhadas pelo pré-processador são:

- **Isolamento e tratamento inicial das palavras:** forma-se uma lista contendo as palavras da frase de entrada submetidas a um tratamento inicial.
- **Tratamento de siglas, abreviações e símbolos especiais:** siglas, abreviações e símbolos especiais são substituídos por suas respectivas formas extensas, isto é, pelas palavras que representam a sua pronúncia.

- **Tratamento de algarismos:** algarismos presentes na frase de entrada são substituídos por suas formas extensas.

Cada uma dessas tarefas encontra-se descrita em detalhe nos itens seguintes.

Especificando-se a opção "g1" ao acionar o sistema de conversão texto-fala, o pré-processador registra a sua saída em um arquivo-texto ("PREPROC.SPT"). Essa opção encontra-se descrita na seção 1.4.

Isolamento e tratamento inicial de palavras

A primeira etapa do pré-processamento consiste no isolamento das palavras presentes na frase de entrada, dando origem a uma *lista de palavras* que exclui símbolos tais como vírgulas, aspas e outros. Estes símbolos têm sua presença registrada em outra lista, denominada *lista de atributos,* associada à lista de palavras. A lista de atributos registra também informações referentes ao posicionamento de cada palavra na frase. Os atributos são armazenados em *flags* contidos em dois bytes (dos quais treze bits são utilizados), conforme a tabela a seguir:

Bit	Significado para bit em 1
0	Palavra seguida de vírgula
1	Palavra seguida de ponto (abreviação)
2	Palavra com um ponto após cada letra (sigla)
3	Palavra seguida de barra
4	Palavra precedida de abertura de aspas duplas
5	Palavra seguida de fechamento de aspas duplas
6	Palavra precedida de abertura de aspas simples
7	Palavra seguida de fechamento de aspas simples
8	Palavra precedida de abertura de parênteses, colchetes ou chaves
9	Palavra seguida de fechamento de parênteses, colchetes ou chaves
10	Palavra composta associada à seguinte
11	Primeira palavra da frase
12	Última palavra da frase

Para exemplificar o processo de elaboração das listas de palavras e de atributos, tome-se a frase seguinte:

No dia 4 de julho, comemora-se a "Independência dos EUA"

A frase não é finalizada por um caractere de pontuação, pois esse caractere foi eliminado previamente pelo módulo de controle do sistema, conforme descrito no capítulo anterior.

A tabela a seguir apresenta as listas de palavras e de atributos correspondentes:

Lista de palavras	Lista de atributos (bin / hex)	Significado dos atributos ativos
No	0100000000000B / 0800H	Primeira palavra da frase
dia	0000000000000B / 0000H	—
4	0000000000000B / 0000H	—
de	0000000000000B / 0000H	—
julho	0000000000001B / 0001H	Palavra seguida de vírgula
comemora	0010000000000B / 0400H	Palavra composta associada à próxima
se	0000000000000B / 0000H	—
a	0000000000000B / 0000H	—
Independência	0000000010000B / 0010H	Palavra precedida de abertura de aspas duplas
dos	0000000000000B / 0000H	—
EUA	1000000100000B / 1020H	Palavra seguida de fechamento de aspas duplas; última palavra da frase

Uma vez isoladas, as palavras são submetidas a um tratamento inicial que consiste na eliminação de sinais diacríticos e cedilhas e na conversão de todas as letras para o caso minúsculo, bem como na eliminação de letras dobradas (como a letra "l" em "Salles")[1] e de algumas letras normalmente não utilizadas na língua portuguesa ("w" e "y"). Símbolos não alfabéticos são mantidos inalterados nesta etapa do pré-processamento. Para cada palavra, é gerada uma lista de atributos associados aos caracteres. Os atributos de caracteres são armazenados em *flags* contidos em dois bytes (dos quais onze bits são utilizados), conforme a tabela abaixo:

Bit	Significado para bit em 1
0	Acento agudo
1	Acento grave
2	Acento circunflexo
3	Til
4	Trema
5	Cedilha
6	Letra "w"
7	Letra "y"
8	Letra maiúscula
9	Letra minúscula
10	Letra dobrada

[1] Na língua portuguesa, consoantes dobradas em geral apresentam pronúncia idêntica à de consoantes simples, com exceção das letras "c", "s" e "r", tratadas de forma diferenciada nas regras de transcrição ortográfico-fonética.

Aplicando-se este processamento sobre a palavra "Independência", obtém-se a palavra processada "independencia" e a seguinte lista de atributos associados às letras:

Letra	Atributos (bin / hex)	Significado dos atributos ativos
i	00100000000B / 0100H	Letra maiúscula
n	00000000000B / 0000H	—
d	00000000000B / 0000H	—
e	00000000000B / 0000H	—
p	00000000000B / 0000H	—
e	00000000000B / 0000H	—
n	00000000000B / 0000H	—
d	00000000000B / 0000H	—
e	00000000100B / 0040H	Acento circunflexo
n	00000000000B / 0000H	—
c	00000000000B / 0000H	—
i	00000000000B / 0000H	—
a	00000000000B / 0000H	—

Para a palavra "EUA", obtém-se a palavra processada "eua" e a seguinte lista de atributos:

Letra	Atributos (bin / hex)	Significado dos atributos ativos
e	00100000000B / 0100H	Letra maiúscula
u	00100000000B / 0100H	Letra maiúscula
a	00100000000B / 0100H	Letra maiúscula

Submetendo-se a palavra "Rússia" a este processamento, obtém-se a palavra processada "rusia" (eliminando-se o acento agudo e a letra dobrada) e a seguinte lista de atributos:

Letra	Atributos (bin / hex)	Significado dos atributos ativos
r	00100000000B / 0100H	Letra maiúscula
u	00000000001B / 0001H	Acento agudo
s	10000000000B / 0400H	Letra dobrada
i	00000000000B / 0000H	—
a	00000000000B / 0000H	—

Após o isolamento e o tratamento inicial das palavras, são efetuadas comparações procurando-se identificar siglas, abreviações, símbolos especiais e algarismos. Essas etapas do pré-processamento são descritas nos itens a seguir.

Tratamento de siglas, abreviações e símbolos especiais

A busca por siglas, abreviações e símbolos especiais faz uso de uma tabela armazenada em arquivo, denominada *tabela de símbolos* (arquivo "PREPROC.SUB"), que lista esses elementos e suas respectivas formas extensas. A seguir encontram-se alguns dos itens presentes nessa tabela.

Símbolo especial, abreviação ou sigla	Tipo de ocorrência	Forma extensa
Sr	Abreviação	senhor
Sra	Abreviação	senhora
Srta	Abreviação	senhorita
Dr	Abreviação	doutor
Dra	Abreviação	doutora
Prof	Abreviação	professor
Profa	Abreviação	professora
Jr	Abreviação	júnior
pág	Abreviação	página
tel	Abreviação	telefone
V	Abreviação	você
etc	Abreviação	et cétera
VS	Sigla	vossa senhoria
RJ	Sigla	Rio de Janeiro
EUA	Sigla	Estados Unidos
&	Símbolo especial	e
%	Símbolo especial	por cento
@	Símbolo especial	arroba

Uma palavra somente é considerada uma abreviação caso seja seguida por um ponto e esteja presente na tabela de símbolos. Por outro lado, há três possibilidades para que uma palavra seja reconhecida como uma sigla: caso cada uma de suas letras seja seguida por um ponto (como em "A.B.N.T."), caso ela esteja presente na tabela de símbolos (como "EUA") ou caso ela contenha seqüências de letras impronunciáveis em português (como em "SKF"). Uma sigla que não conste da tabela de símbolos é simplesmente soletrada; por exemplo, a sigla "A.B.N.T." é expandida para "a bê ene tê". Termos particulares (normalmente de origem estrangeira) que contenham seqüências de letras impronunciáveis no português e não devam ser soletrados são listados em uma seção específica da tabela de símbolos, não sendo alterados pelo pré-processador. As pronúncias apropriadas para esses termos devem ser fornecidas ao sistema através do dicionário de exceções do transcritor ortográfico-fonético, conforme descrito na seção 2.3.

Antes da busca por abreviações e siglas, as palavras listadas na tabela de símbolos são submetidas ao mesmo tratamento inicial descrito para as palavras da frase de entrada. Des-

se modo, um termo é localizado na tabela de símbolos mesmo que haja diferenças em termos de sinais diacríticos, cedilhas, letras dobradas, letras maiúsculas e minúsculas, letras "w" e "v" e letras "y" e "i". Por exemplo, as abreviações "pag", "pág" e "Pág" são todas expandidas para "página". Essa característica torna o sistema robusto a erros de acentuação e mesmo à falta total de acentos (comum em mensagens enviadas através de correio eletrônico). Se houver necessidade de distinção entre termos de grafia similar (com diferenças de acentuação, por exemplo), todos devem ser especificados na tabela de símbolos em posições adjacentes. Quando dois ou mais termos presentes na tabela de símbolos satisfazem a comparação com uma palavra da frase de entrada, para cada um é calculado um escore que indica o grau de similaridade em termos de atributos dos caracteres. Esse escore, inicialmente nulo, é incrementado de 1 para cada atributo não coincidente entre a palavra da frase de entrada e o termo da tabela de símbolos, escolhendo-se ao final o termo com menor escore. Por exemplo, se a palavra da frase de entrada fosse "maçã" e constassem da tabela de símbolos as palavras "maca" e "maçã", ambas satisfariam o critério de comparação (que não leva em conta o cedilha e o til), mas a segunda palavra seria escolhida por apresentar escore nulo (nenhuma diferença em relação à palavra da frase de entrada), enquanto a primeira teria escore igual a 2 (devido à ausência do cedilha e do til). Em caso de empate nos escores, o termo especificado anteriormente na tabela de símbolos é escolhido; assim, termos de grafia similar devem ser especificados em ordem decrescente de freqüência de uso.

Tratamento de algarismos

O pré-processador é também responsável pela conversão de algarismos para a sua forma extensa. Essa conversão pode ser feita de dois modos: leitura do número completo (convertendo, por exemplo, "193" para "cento e noventa e três") ou leitura algarismo a algarismo (convertendo "193" para "um nove três"). Para números isolados, é utilizada a forma de leitura completa, com a vírgula decimal (se presente) pronunciada literalmente; por exemplo, o número "23,12" é convertido para "vinte e três vírgula doze". Determinadas configurações de algarismos, no entanto, recebem tratamentos diferenciados: datas, no formato "*dia / mês*" (convertendo-se, por exemplo, "12/10" para "doze do dez") ou "*dia / mês / ano*" (convertendo-se "12/10/97" para "doze do dez de noventa e sete"); quantidades monetárias, no formato "R$ *valor*" (convertendo-se "R$ 12" para "doze reais") ou "R$ *valor,valor*" (convertendo-se "R$ 12,20" para "doze reais e vinte centavos"); e números de telefone, no formato "*##-####*", "*###-####*" ou "*####-####*", onde o símbolo "#" corresponde a um algarismo (convertendo-se "289-3134" para "dois oito nove três um três

quatro"). Números ordinais e números de telefone com códigos de localidades não são reconhecidos pela versão atual do sistema.

Embora seja parte integrante do módulo de pré-processamento de texto, a rotina para tratamento de números somente é executada após a conclusão da classificação gramatical (descrita na próxima seção). Isso se deve à necessidade da informação de gênero para definir-se a conversão apropriada para os algarismos "1" ("um" ou "uma") e "2" ("dois" ou "duas"), assim como para as centenas ("duzentos" ou "duzentas", "trezentos" ou "trezentas" etc.).

2.2 CLASSIFICAÇÃO GRAMATICAL

A finalidade da classificação gramatical

Um classificador gramatical é parte essencial de um sistema de conversão texto-fala. O transcritor ortográfico-fonético (descrito na seção 2.3) toma por base a classificação gramatical na escolha da pronúncia adequada para determinadas palavras, e o módulo de processamento prosódico utiliza informações gramaticais no processo de geração de durações de segmentos e contornos de freqüência fundamental para a frase de entrada. A qualidade da transcrição fonética e do processamento prosódico estão, portanto, diretamente relacionados à precisão e à abrangência da classificação gramatical. O classificador gramatical implementado (arquivo "CLAGRAM.CPP") procura identificar as classes das palavras presentes na frase de entrada, não entrando no mérito das relações sintáticas entre elas.

Classes e subclasses gramaticais[2]

O classificador pode atribuir as seguintes *classes gramaticais* a uma palavra da frase de entrada: substantivo, adjetivo, artigo, pronome, advérbio, preposição, conjunção, verbo, interjeição, numeral e palavra denotativa. Palavras para as quais o classificador não é capaz de atribuir uma classe gramatical única podem receber múltiplas classificações.

Além das classes listadas acima, são atribuídas também *subclasses gramaticais* às palavras da frase de entrada, refinando-se a classificação obtida. Para cada classe gramatical associada a uma palavra, são alocados cinco bytes para armazenar as respectivas subclasses, com dois bits por subclasse; totaliza-se assim um máximo de vinte subclasses por classe gramatical. Cada subclasse recebe um coeficiente que indica o grau de confiabilidade da

[2] Foram empregadas neste trabalho as classes e subclasses gramaticais definidas em [Lima 1991].

classificação; esse coeficiente, que varia entre 0 e 3, é armazenado nos dois bits associados à subclasse. Um coeficiente nulo indica que a subclasse muito provavelmente não corresponde à classificação correta da palavra; coeficientes variando entre 1 e 3 indicam graus crescentes de probabilidade de correspondência com a classificação correta (1 = pouco provável; 2 = provável; 3 = muito provável).

O módulo de classificação gramatical pode opcionalmente registrar a sua saída em um arquivo-texto ("CLAGRAM.SCG"). Nesse arquivo, as classes e subclasses gramaticais são representadas por abreviações de três e dois caracteres, respectivamente, sendo os coeficientes de confiabilidade especificados imediatamente após as abreviações das subclasses correspondentes. As tabelas a seguir listam as subclasses que podem ser associadas a cada classe gramatical, juntamente com as respectivas abreviações.

SUBSTANTIVO – SUB

Abreviação	Descrição
MA	Masculino
FE	Feminino
SI	Singular
PL	Plural
CP	Composto
DI	Diminutivo
AU	Aumentativo
PR	Próprio
CM	Comum
CL	Coletivo

ADJETIVO – ADJ

Abreviação	Descrição
MA	Masculino
FE	Feminino
SI	Singular
PL	Plural
CP	Composto
DI	Diminutivo
AU	Aumentativo
SU	Superlativo
CR	Comparativo

PRONOME – PRO

Abreviação	Descrição
RE	Pessoal reto
OA	Pessoal objetivo átono
OT	Pessoal objetivo tônico
TR	Pessoal de tratamento
PO	Possessivo
DM	Demonstrativo
IN	Indefinido
RL	Relativo
EU	Primeira pessoa
TU	Segunda pessoa
EL	Terceira pessoa
SI	Singular com relação à pessoa
PL	Plural com relação à pessoa
MA	Masculino com relação à pessoa
FE	Feminino com relação à pessoa
SJ	Singular com relação ao objeto
PJ	Plural com relação ao objeto
MJ	Masculino com relação ao objeto
FJ	Feminino com relação ao objeto
DA	Adverbial demonstrativo

ARTIGO – ART

Abreviação	Descrição
DE	Definido
IN	Indefinido
SI	Singular
PL	Plural
MA	Masculino
FE	Feminino

ADVÉRBIO – ADV

Abreviação	Descrição
DV	Dúvida
IS	Intensidade
LU	Lugar
MO	Modo
TE	Tempo
LO	Locução
RL	Relativo
IT	Interrogativo
DI	Diminutivo
AU	Aumentativo
SU	Superlativo
CR	Comparativo

CONJUNÇÃO – CON

Abreviação	Descrição
AD	Coordenativa aditiva
AV	Coordenativa adversativa
AL	Coordenativa alternativa
CO	Coordenativa conclusiva
EX	Coordenativa explicativa
CS	Subordinativa causal
CC	Subordinativa concessiva
CD	Subordinativa condicional
CF	Subordinativa conformativa
CR	Subordinativa comparativa
CN	Subordinativa consecutiva
FI	Subordinativa final
PC	Subordinativa proporcional
TM	Subordinativa temporal
IG	Subordinativa integrante

PALAVRA DENOTATIVA–DEN

Abreviação	Descrição
AF	Afirmação
NG	Negação
EC	Exclusão
IL	Inclusão
AA	Avaliação
DG	Designação
EP	Explicação
RT	Retificação
AP	Apreciação

PREPOSIÇÃO – PRE

Abreviação	Descrição
ES	Essencial
AC	Acidental
LO	Locução

VERBO – VER

Abreviação	Descrição
ID	Indicativo
SB	Subjuntivo
IF	Infinitivo
PA	Particípio
GE	Gerúndio
IM	Imperativo afirmativo
IE	Imperativo negativo
PS	Presente
PP	Pretérito perfeito
PI	Pretérito imperfeito
PM	Pretérito mais-que-perfeito
FP	Futuro do presente
FT	Futuro do pretérito
EU	Primeira pessoa
TU	Segunda pessoa
EL	Terceira pessoa
SI	Singular
PL	Plural
MA	Masculino
FE	Feminino

INTERJEIÇÃO – INT

Abreviação	Descrição
AG	Alegria
DS	Desejo
DO	Dor
CH	Chamamento
SL	Silêncio
AE	Advertência
IC	Incredulidade

NUMERAL – NUM

Abreviação	Descrição
CA	Cardinal
OR	Ordinal
FR	Fracionário
MU	Multiplicativo
DU	Dual

Os itens seguintes descrevem as técnicas utilizadas para a atribuição de classes e sub-classes gramaticais às palavras da frase de entrada.

Classificação por comparação

A primeira etapa da classificação gramatical consiste na comparação de cada palavra da frase de entrada com termos listados em uma tabela armazenada em arquivo, denominada *tabela de classificação gramatical* (arquivo "CLAGRAM.CLA"). Essa tabela contém os principais pronomes, preposições, conjunções, advérbios, artigos, interjeições e palavras denotativas da língua portuguesa, bem como as conjugações completas de alguns verbos irregulares muito comuns (como "ser", "estar", "ter", "haver", "pôr" etc.). As classes e subclasses gramaticais de cada termo presente na tabela são especificadas através das abreviações listadas no item anterior. A palavra "eu", por exemplo, é classificada da seguinte forma:

```
eu PRO RE3 EU3 SI3
```

indicando que se trata de um pronome (PRO) pessoal reto (RE) de primeira pessoa (EU) no singular (SI). Os números que seguem as especificações de subclasses são os coeficientes de confiabilidade da classificação; nesse caso, todas as subclasses receberam coeficiente 3, indicando que a classificação muito provavelmente está correta.

Diversas palavras presentes na tabela de classificação gramatical podem desempenhar duas ou mais funções gramaticais, conforme o contexto em que são utilizadas; nesse caso, a classificação não é única, sendo atribuídas várias classes gramaticais a uma mesma palavra. Pode ocorrer também a atribuição de múltiplas subclasses gramaticais conflitantes dentro de uma mesma classe gramatical. A palavra "a", por exemplo, é classificada da seguinte forma:

```
a ART DE3 FE3 SI3  PRE ES2  PRO OA1 DM1 EL1 FE1 SI1
```

indicando que pode se tratar de um artigo (ART) definido (DE) feminino (FE) singular (SI) com probabilidade elevada (coeficiente igual a 3 para todas as subclasses), ou de uma preposição (PRE) essencial (ES) com probabilidade média (coeficiente igual a 2), ou ainda de um pronome (PRO) pessoal objetivo átono (OA) de terceira pessoa (EL) feminino (FE) singular (SI) com probabilidade baixa (coeficiente igual a 1 para todas as subclasses). Ao invés de um pronome pessoal objetivo átono, a palavra pode também corresponder a um pronome demonstrativo (DM) com probabilidade baixa (coeficiente igual a 1). Os coeficientes atribuídos a cada subclasse são baseados na freqüência com que a palavra é utilizada desempenhando a função gramatical correspondente; embora não tenha sido realizado um estudo formal dessas freqüências, a simples busca pelas palavras em textos genéricos permitiu na maioria dos casos a definição razoavelmente precisa dos valores dos coeficientes para cada subclasse.

Além do caso descrito no parágrafo anterior, existe outra situação na qual se torna necessária a atribuição de classes gramaticais múltiplas: quando uma palavra é formada pela aglutinação de duas outras, como muitas vezes ocorre com preposições, artigos e conjunções. Por exemplo, a palavra "da" é obtida pela aglutinação da preposição "de" com o artigo "a". Ela é classificada da seguinte forma:

```
da PRE ES3  ART DE3 FE3 SI3
```

indicando que se trata de uma preposição (PRE) essencial (ES) e de um artigo (ART) definido (DE) feminino (FE) singular (SI). Como a palavra desempenha as duas funções gramaticais simultaneamente, valores idênticos são atribuídos aos coeficientes das subclasses associadas a cada uma das classes.

A etapa seguinte do processo de classificação gramatical consiste na verificação das terminações das palavras presentes na frase de entrada. Diversas terminações características da língua portuguesa são listadas na tabela de classificação gramatical, incluindo verbos regulares, advérbios, aumentativos (geralmente substantivos, adjetivos ou advérbios) e outros. A classificação por terminação somente é efetuada para as palavras que não tenham sido classificadas na etapa anterior do processo. A cada terminação é atribuída uma ou mais classes gramaticais, de forma idêntica à descrita anteriormente para palavras completas. Por exemplo, a terminação "mente", característica de advérbios, é classificada como:

```
mente ADV MO3 DV1 IS1 LU1 TE1
```

Como não se conhece o tipo de advérbio em questão, todas as possibilidades são levadas em conta (modo, dúvida, intensidade, lugar e tempo), mas a subclasse correspondente aos advérbios de modo recebe um coeficiente maior do que as demais, pois a terminação "mente" é mais comum em advérbios desta subclasse.

Classificação por regras posicionais

Após a classificação por comparação, passa-se à etapa de classificação por regras posicionais. Essas regras atribuem classes gramaticais às palavras com base nas classificações de palavras vizinhas. O conjunto completo de regras posicionais encontra-se no apêndice A.

O primeiro grupo de regras posicionais procura definir com maior precisão as classificações de palavras às quais foram atribuídas duas ou mais classes gramaticais na etapa de classificação por comparação (excluindo palavras que de fato desempenham funções gramaticais múltiplas, como "do" e "pelo"). Como exemplo, tem-se a palavra "se", que pode

desempenhar a função de pronome e de conjunção. Uma das regras posicionais verifica se essa palavra (ou qualquer outro pronome objetivo átono) está ligada à anterior através de um hífen (como em "faz-se"); em caso afirmativo, a palavra é definitivamente classificada como pronome e as demais classificações são descartadas. Outra ocorrência comum é a classificação incorreta de um substantivo como um verbo em decorrência da sua terminação; caso haja um artigo precedendo a palavra, no entanto, torna-se evidente que se trata de um substantivo (ou, com menor probabilidade, de um adjetivo).

O segundo grupo de regras posicionais atribui classes gramaticais às palavras que não receberam classificação alguma na etapa de comparação. Essas palavras correspondem em sua maioria a substantivos e adjetivos, para os quais não há um conjunto restrito de terminações características. As regras procuram determinar quais dessas palavras são provavelmente substantivos e quais são provavelmente adjetivos. Na língua portuguesa, os adjetivos posicionam-se tipicamente após os substantivos, embora em certos casos a inversão seja aceitável ou mesmo aconselhável. Como a diferenciação precisa entre substantivos e adjetivos somente é possível através da consulta a um léxico amplo previamente classificado (não disponível para o sistema), optou-se por utilizar uma regra simples: caso seja encontrada uma seqüência de palavras não classificadas, a primeira delas é considerada um substantivo com probabilidade elevada e um adjetivo com probabilidade baixa, enquanto as demais são consideradas adjetivos com probabilidade elevada e substantivos com probabilidade baixa. Uma regra alternativa é aplicada no caso em que uma das palavras não classificadas se inicia por letra maiúscula (com as demais minúsculas) e não está em início de frase; nessa situação, a palavra é considerada um substantivo próprio e as demais são classificadas como adjetivos. Regras adicionais atribuem subclasses de gênero e número às palavras classificadas como substantivos ou adjetivos, baseando-se nas suas terminações (substantivos e adjetivos terminados em "o", por exemplo, são provavelmente masculinos) e na classificação da palavra precedente (artigos e pronomes possessivos, por exemplo, permitem definir o gênero e o número da palavra que os segue).

Desempenho do classificador gramatical

Apesar de sua estrutura relativamente simples, o classificador gramatical implementado apresenta bom desempenho para frases típicas em português. Para frases muito complexas ou de estrutura não usual, no entanto, o classificador pode apresentar uma taxa de erros elevada. A tabela a seguir mostra as classificações obtidas pelo sistema, juntamente com os coeficientes atribuídos às subclasses gramaticais, para cada palavra da seguinte frase: "teriam sido palavras certeiras se a Gazeta soubesse a verdadeira natureza dos fatos". São mostradas também as classificações exatas de cada palavra.

Palavra	Classificação determinada pelo sistema	Classificação exata
teriam	VER ID3 FT3 EL3 PL3	VER ID FT EL PL
sido	VER PA3 SI3 MA3	VER PA SI MA
palavras	ADJ FE1 PL1 SUB FE1 PL1 CM1 VER ID1 SB1 IE1 PS1 TU1 SI1	SUB FE PL CM
certeiras	SUB FE3 PL3 CM3 ADJ FE1 PL1	ADJ FE PL
se	CON CD3 IG2 CC1 CR1 PRO OA2 EL2	CON CD
a	ART DE3 FE3 SI3 PRE ES3 PRO OA2 DM1 EL2 FE2 SI2	ART DE FE SI
Gazeta	SUB FE3 SI3 CM2 PR3	SUB FE SI PR
soubesse	VER SB3 PI3 EU2 EL3 SI3	VER SB PI SI EL
a	ART DE3 FE3 SI3 PRE ES2 PRO OA1 EL1 FE1 SI1	ART DE FE SI
verdadeira	SUB FE3 SI3 CM3 ADJ FE2 SI2 VER ID1 SB1 IM1 PS1 PM1 EU1 TU1 EL1 SI1	ADJ FE SI
natureza	SUB FE3 SI3 CM3 ADJ FE1 SI1 VER ID1 PS1 TU1 SI1 EL1 IM1 SB1 EU1 IE1	SUB FE SI CM
dos	PRE ES3 ART DE3 MA3 PL3	PRE ES ART DE MA PL
fatos	SUB MA3 PL3 CM3 ADJ MA2 PL2	SUB MA PL CM

Como se vê, neste exemplo são obtidas classificações próximas das corretas para a maioria das palavras. As principais falhas ocorreram para os termos "certeiras" e "verdadeira", classificados mais provavelmente como substantivos do que como adjetivos. Como já dito, a diferenciação precisa entre substantivos e adjetivos somente é possível utilizando-se um léxico extenso previamente classificado. Além disso, o termo "palavras" foi classificado de forma igualmente provável como adjetivo, substantivo e verbo; o mesmo ocorreu para o primeiro "a", classificado como artigo e preposição com coeficientes de subclasses idênticos. Para todas as palavras da frase acima, a classificação exata corresponde ao menos à segunda classificação mais provável determinada pelo sistema.

A classificação gramatical tem influência decisiva nos módulos de transcrição ortográfico-fonética e de processamento prosódico, conforme discutido mais adiante. A precisão obtida foi suficiente para garantir um bom desempenho do sistema como um todo, embora futuras sofisticações no processamento prosódico e na transcrição ortográfico-fonética possam requerer uma classificação mais exata, eventualmente implicando na adoção de técnicas mais complexas (como métodos baseados em análise morfológica, por exemplo).

2.3 DIVISÃO SILÁBICA E DETERMINAÇÃO DA SÍLABA TÔNICA LEXICAL

Divisão silábica

Uma sílaba corresponde a um fone ou grupo de fones pronunciado num único impulso de expiração [Lima 1991]. Informações referentes à divisão silábica apresentam implicações importantes a nível prosódico, conforme discutido no capítulo 3. Um módulo específico (arquivo "DIVSIL.CPP") foi implementado com o objetivo de efetuar a divisão silábica das palavras presentes na frase de entrada, bem como determinar as suas sílabas tônicas (discutido no próximo item).

Na língua portuguesa, a determinação das fronteiras entre sílabas é relativamente simples. Toda sílaba apresenta uma vogal como núcleo, podendo estar cercada por consoantes ou por outras vogais (desempenhando a função de semivogais). A base de um algoritmo de divisão silábica consiste em localizar as vogais que formam os núcleos das sílabas e isolar as consoantes e semivogais a elas associadas. O algoritmo desenvolvido, detalhado no apêndice A, mostrou-se eficiente na quase totalidade dos casos. Foram detectadas falhas somente para seqüências de vogais que podem constituir tanto ditongos como hiatos, sendo necessária informação sobre a composição morfológica do vocábulo para que se possa determinar a divisão silábica correta. A palavra "traidor", por exemplo, é separada pelo sistema como "trai-dor", considerando-se a seqüência "ai" como um ditongo; a separação correta, no entanto, é "tra-i-dor", seguindo o verbo "trair", no qual o "i" tônico (evidenciado pelo "r" final) força o hiato. Erros na divisão silábica podem ser corrigidos utilizando o dicionário de exceções do transcritor fonético, descrito na seção 2.4.

Determinação da sílaba tônica lexical

A determinação de sílabas tônicas lexicais é de importância fundamental para o processamento prosódico, conforme discutido no capítulo 3. Em muitos casos a inteligibilidade de uma palavra depende do correto posicionamento da sílaba tônica; como exemplos, têm-se as palavras "ira" e "irá", ou "sábia", "sabia" e "sabiá".

A língua portuguesa apresenta três possibilidades para o posicionamento da sílaba tônica lexical: ao final da palavra (palavra oxítona), imediatamente antes da última sílaba (palavra paroxítona) e imediatamente antes da penúltima sílaba (palavra proparoxítona). Palavras que contenham acento agudo ou circunflexo têm a posição da sílaba tônica automaticamente determinada. Como as palavras proparoxítonas são sempre acentuadas, a determinação da sua sílaba tônica é trivial. Para uma palavra não acentuada, a sílaba tônica é determinada através de um conjunto de regras levando em conta a terminação da palavra,

tomada a partir da primeira vogal da última sílaba (desconsiderando-se o "u" quando este forma dígrafo com as letras "q" ou "g"). Estas regras são:

- **Para palavras polissilábicas:** Se a palavra termina em "a", "as", "e", "es", "o", "os", "am", "em" ou "ens", a penúltima sílaba é tônica (palavra paroxítona); senão, a última sílaba é tônica (palavra oxítona).
- **Para monossílabos:** Se a palavra termina em "a", "as", "e", "es", "o" ou "os", a sílaba é átona; em caso contrário, a sílaba é tônica.

As regras falham na determinação da sílaba tônica de alguns prefixos não acentuados, como "super" (em "super-homem") e "semi" (em "semi-analfabeto"). Erros como esses podem ser corrigidos por meio do dicionário de exceções do transcritor fonético, descrito na seção 2.4.

Em vocábulos longos, sobretudo os derivados, é comum a existência de uma sílaba tônica secundária localizada antes da antepenúltima sílaba. Na palavra "razoavelmente", por exemplo, o segundo "a" corresponde a uma sílaba tônica secundária (devido à sua tonicidade na palavra "razoável"). Embora tenham implicações prosódicas perceptíveis, as sílabas tônicas secundárias não afetam significativamente a inteligibilidade da fala e não são tratadas na versão atual do sistema.

2.4 TRANSCRIÇÃO ORTOGRÁFICO-FONÉTICA

A necessidade de uma representação fonética

A linguagem escrita constitui um mecanismo alternativo à linguagem falada, permitindo a representação através de símbolos ortográficos de sentenças formuladas em uma determinada língua. Para as línguas que adotam o alfabeto latino, são usadas 26 letras, além de diversos símbolos auxiliares (acentos, sinais de pontuação, algarismos etc.). Não há, no entanto, uma correspondência unívoca entre símbolos ortográficos e fones. Dois sím-bolos ortográficos distintos podem representar um mesmo fone; isso ocorre, por exemplo, para as letras "s" e "c" nas palavras portuguesas "selva" e "cela", ambas representando o fone /s/. Além disso, um mesmo símbolo ortográfico pode representar fones diferentes conforme o contexto em que se insere; por exemplo, nas palavras "cedo" e "casa", a letra "c" corresponde aos fones /s/ e /k/, respectivamente. Certas letras podem apresentar pronúncias variadas mesmo quando inseridas em contextos similares; a letra "x" nas palavras "fixo" e "lixo", por exemplo, corresponde respectivamente aos fones /ks/ e /x/. Outro caso comum é a combinação de duas letras para formar um único fone (dígrafos), como em "lh", "nh" e "rr".

Como um sistema de conversão texto-fala necessita de uma representação precisa dos fones a sintetizar, é requerido um módulo que a obtenha com base nos símbolos ortográficos presentes no texto de entrada. Esse módulo, denominado *transcritor ortográfico-fonético,* produz a partir de uma frase em português uma seqüência de símbolos univocamente associados aos fones passíveis de síntese pelo sistema.

A transcrição ortográfico-fonética pode ser efetuada com diferentes graus de precisão. Em uma *transcrição larga,* não são levadas em conta diferenças pequenas, embora perceptíveis, entre fones. Realizações acústicas de um dado fonema podem apresentar características espectrais diversas em função do seu contexto fonético, mas são representadas por um símbolo único em uma transcrição larga. Já uma *transcrição estreita* leva em consideração detalhes que diferenciam as várias realizações possíveis de um fonema, atribuindo-lhes símbolos distintos.

Técnicas para transcrição ortográfico-fonética

Diversas técnicas podem ser empregadas no processo de transcrição ortográfico-fonética. A técnica de implementação mais simples é a transcrição por regras, que procura deduzir os fones a serem pronunciados com base no contexto em que se insere cada letra do texto de entrada. Para línguas como o português, na qual a correspondência entre letras e fones é razoavelmente estável, esta técnica permite obter bons resultados; para o inglês, ao contrário, a sua eficiência é menor, pois a pronúncia associada a letras ou grupos de letras pode sofrer variações fortes e muitas vezes não sistemáticas (comparem-se, por exemplo, as pronúncias das palavras inglesas "rough" e "through"). Mesmo no português há casos em que uma transcrição totalmente baseada em regras se torna inviável; a letra "x", por exemplo, é associada aos fones /x/, /s/, /ks/ e /z/ de forma pouco sistemática. Outro exemplo é a determinação da pronúncia apropriada para as letras "e" e "o" não acentuadas, que podem corresponder a sons abertos (como em "peste" e "bola") ou fechados (como em "neste" e "tola").

Uma abordagem alternativa que conduz potencialmente a uma transcrição ortográfico-fonética de alta qualidade é a utilização de técnicas de análise morfológica. Uma base de dados deve, em princípio, conter todos os morfemas da língua, juntamente com as respectivas transcrições fonéticas. Cada palavra do texto de entrada é decomposta em morfemas e as transcrições correspondentes são concatenadas. Esta técnica requer um conjunto de regras que atue nas junções entre morfemas, pois é freqüente a ocorrência de mutações sonoras nessas regiões. Como a maioria dos neologismos é criada a partir de morfemas já existentes, o sistema automaticamente os reconhece e transcreve corretamente. Caso seja en-

contrado um morfema desconhecido, um algoritmo de transcrição por regras como o descrito no parágrafo anterior é usado. Embora em geral a pronúncia de um morfema seja idêntica em todas as palavras que o contêm, há casos em que ocorre alteração; por exemplo, a letra "x" é pronunciada como /s/ na palavra "máximo" e como /ks/ em "maximizar". Nessa situação, a palavra que foge à regra geral deve ser incluída como um morfema único na base de dados, juntamente com a sua transcrição correta. A técnica de transcrição por análise morfológica apresenta duas desvantagens principais: (1) é comum que haja várias decomposições morfológicas plausíveis para uma palavra, podendo ocorrer erros de transcrição caso seja escolhida uma decomposição incorreta; e (2) uma grande quantidade de memória de massa é necessária para armazenar a lista de morfemas e suas transcrições. Além disso, a elaboração da lista dos morfemas de uma língua é uma tarefa complexa e muito extensa.

Uma técnica intermediária entre as apresentadas acima efetua a transcrição ortográfico-fonética por meio de um conjunto de regras, mas utiliza uma base de dados, denominada *dicionário de exceções*, que lista as palavras transcritas incorretamente pelas regras, juntamente com as transcrições apropriadas. Inicialmente busca-se a palavra a transcrever no dicionário de exceções; caso ela não seja encontrada, são aplicadas as regras de transcrição. Para regras com índice de acerto elevado, o dicionário de exceções pode ser significativamente menor do que a base de dados utilizada na análise morfológica. O transcritor ortográfico-fonético implementado baseia-se em uma variação desta técnica, conforme descrito nos itens seguintes.

Regras de transcrição

O módulo de transcrição ortográfico-fonética (arquivo "TRANSFON.CPP") utiliza um conjunto de regras de síntese e um dicionário de exceções (descrito no próximo item). A transcrição efetuada é relativamente larga, pois os fonemas são representados em sua maioria por símbolos únicos, com exceção de alguns sons vocálicos e de consoantes nasais. A tabela a seguir lista os símbolos fonéticos utilizados no sistema, juntamente com palavras da língua portuguesa exemplificando a sua utilização[3]. Em cada palavra, as letras correspondentes ao fone exemplificado foram grafadas em itálico.

[3] Embora a rigor as pausas não constituam fones, um símbolo específico (o caractere de cifrão, "$") foi a elas atribuído, podendo ser utilizado em transcrições fornecidas pelo usuário diretamente ao módulo conversor (para o qual não há distinção entre pausas e fones propriamente ditos).

Classe	Símbolo adotado	Exemplo
Pausa	$	—
Vogais orais tônicas	i	ap*i*to
	e	par*e*de
	é	bon*e*co
	a	at*a*lho
	ó	b*o*la
	o	b*o*lo
	u	*u*va
Vogais orais átonas	y	salár*i*o, pared*e*
	A	pat*a*
	w	lou*c*o, apit*o*
Vogais nasais tônicas	ï	cap*i*m
	ë	*e*ntrada
	ä	d*a*nça
	ö	t*o*mbo
	ü	b*u*mbo
Vogais nasais átonas	ÿ	mã*e*
	Ü	emoç*ão*
Fricativas sonoras	z	ca*s*a
	j	ágil
	v	*v*azio
	R	cor*r*eio

Classe	Símbolo adotado	Exemplo
Fricativas surdas	x	fe*ch*ado
	f	so*f*á
	s	ba*c*ia
	S[4]	li*s*ta
Africadas[5]	J	*d*isco
	X	*t*ipo
Oclusivas surdas	p	*p*ato
	k	di*c*a
	t	li*s*ta
Oclusivas sonoras	d	aju*d*a
	b	*b*oné
	g	*gu*eto
Líquidas	r	pa*r*ada
	P[6]	pa*r*te
	l	fi*l*a
	L	pa*lh*a
Consoantes nasais	m	cha*m*a
	n	fi*n*o
	M	lo*m*bo
	N	ma*n*to, ma*n*do
	Ñ	ma*n*ga, ma*n*co
	ñ	li*nh*a

Para cada letra ou grupo de letras da frase de entrada, as regras de transcrição determinam os símbolos fonéticos correspondentes através de uma análise da sua vizinhança. As regras são em geral bastante simples, na maioria das vezes consistindo na substituição de uma ou duas letras por um símbolo fonético. A letra "x" é uma exceção, apresentando um conjunto de regras relativamente elaborado e que nem sempre produz resultados corretos. Sinais diacríticos são levados em conta através da análise dos atributos das letras (descritos na seção 2.1), que indicam também a presença de letras dobradas (relevante para a pronúncia das letras "c", "s" e "r"). A letra "h" é geralmente ignorada, exceto quando sucede as letras "c", "l", "p" e "s", com as quais forma dígrafos[7]. As letras "w" e "y", substituídas por "v" e "i" pelo pré-processador, não recebem tratamento diferenciado.

[4] O fone /S/ somente é utilizado nos finais de sílaba (posição de coda); nas demais posições, utiliza-se o fone /s/.

[5] A africada /J/ somente é utilizada em conjunto com o fone /d/ para a formação do som /dJ/, como em "disco" (/dJiSkw/); já a africada /X/ é empregada em conjunto com o fone /t / para a formação do som /tX/, como em "tiro" (/tXirw/).

[6] O fone /P/ somente é utilizado nos finais de sílaba (posição de coda); nas demais posições, utiliza-se o fone /r/.

[7] Embora os dígrafos "ph" e "sh" a rigor não constem da língua portuguesa, eles aparecem em várias marcas comerciais e em diversas palavras estrangeiras utilizadas na linguagem corrente; por isso, foram levados em conta nas regras de transcrição fonética.

As letras "a", "i" e "u" são transcritas diretamente para os fones correspondentes, utilizando as variedades tônicas, átonas ou nasais[8]. Uma exceção é a letra "u" precedida por "q" ou "g" e seguida de "e" ou "i"; nesses casos, os pares "qu" ou "gu" correspondem aos fones /k/ e /g/, respectivamente. As letras "e" e "o" não acentuadas apresentam um complicador adicional, pois podem ser transcritas para sons abertos ou fechados. Como os sons fechados são mais comuns, as letras "e" e "o" são a princípio associadas aos fones /e/ e /o/ (fechados), mas podem ser convertidas nos fones /é/ e /ó/ (abertos) caso determinadas condições sejam satisfeitas. Por exemplo, em palavras paroxítonas femininas que contenham as letras "e" ou "o" na sílaba tônica, os sons correspondentes tendem a ser abertos, como no pronome "ela" (em contraposição ao pronome "ele"). Mesmo palavras homógrafas podem apresentar pronúncias variadas: no substantivo "peso" (como em "eu levanto o peso"), por exemplo, a letra "e" corresponde a um som fechado, enquanto no verbo "peso" (como em "eu peso 70 quilos") a mesma letra corresponde a um som aberto[9]. Tanto neste caso como no anterior são utilizadas informações providas pelo classificador gramatical. Persistindo uma transcrição incorreta das letras "e" ou "o" após a aplicação dessas regras, inclui-se a palavra no dicionário de exceções.

A transcrição gerada pelas regras apresenta uma taxa de acerto próxima de 100% para textos genéricos. Conforme esperado, os erros concentram-se sobretudo nas letras "x", "e" e "o". Como o transcritor fonético faz uso de informações provenientes do classificador gramatical, o desempenho deste último constitui um fator limitante para a qualidade da transcrição produzida. O dicionário de exceções, descrito no próximo item, permite corrigir grande parte dos erros de transcrição.

As regras de transcrição ortográfico-fonética completas encontram-se no apêndice A.

Dicionário de exceções

O dicionário de exceções (arquivo "TRANSFON.EXC") contém palavras da língua portuguesa transcritas incorretamente pelas regras de transcrição ortográfico-fonética, além de palavras estrangeiras de uso corrente no português, como "champagne", "shampoo" e "pizza". Não foi realizado um estudo sistemático para a construção do dicionário de exceções (por exemplo, aplicando-se as regras de transcrição a todas as palavras de um dicio-

[8] Conforme [Aquino 1997], as vogais pré-tônicas são muito semelhantes às tônicas em termos de propriedades espectrais; por isso, os mesmos símbolos foram utilizados para transcrevê-las. As variedades átonas (/A/, /y/, /w/, /ÿ/ e /Ü/) somente são usadas para as vogais pós-tônicas e na formação de certos ditongos.

[9] Há casos em que palavras homógrafas de mesma classe gramatical apresentam pronúncias diferentes para a letra "e" ou a letra "o". Tomem-se, por exemplo, as frases "ele tem fome e sede" e "a sede da empresa fica nesta cidade"; a palavra "sede" é pronunciada "sedJy" na primeira frase e "sédJy" na segunda. Em situações como essa, a classificação gramatical não é suficiente para a determinação da pronúncia correta, sendo necessárias informações de ordem semântica.

nário da língua portuguesa), tendo sido tomados como base os erros detectados na transcrição de textos genéricos extraídos de jornais.

Como exemplo de utilização do dicionário de exceções, tem-se a palavra "lixa", transcrita incorretamente como /liksa/ pelas regras. O item correspondente do dicionário é especificado da seguinte forma:

```
lixa li-xa
```

O primeiro termo presente na linha corresponde à forma ortográfica da palavra ("lixa"), enquanto o segundo corresponde à sua transcrição (/lixa/); neste caso particular, a forma ortográfica e a transcrição são idênticas. O hífen é um indicador de separação silábica.

O fato de existirem diversas variações para uma palavra (plural, superlativo, conjugações, inclusão de prefixos etc.) tenderia a aumentar exageradamente o tamanho do dicionário de exceções. Esse problema é contornado utilizando-se uma versão adaptada das expressões regulares definidas no sistema operacional Unix. Somente são especificadas no dicionário as porções principais das palavras (correspondendo geralmente aos seus radicais), juntamente com suas transcrições, permitindo-se a anexação de prefixos e sufixos que são transcritos pelas regras de transcrição. São quatro os caracteres utilizados na formação de expressões regulares:

- **Ponto final (".")**: Substitui uma letra qualquer. Por exemplo, dada a expressão regular "vive.", são reconhecidas as palavras "viver", "vivem", "viveu" etc. Para a expressão regular ".iv..", além das palavras já citadas, são também reconhecidas "vivia", "livre", "divas" etc.

- **Ponto de interrogação ("?")**: Substitui uma letra qualquer ou nenhuma letra. Por exemplo, dada a expressão regular "vive?", são reconhecidas as palavras "vive", "viver", "vivem" etc. Para a expressão regular "vive???", além das palavras já citadas, são também reconhecidas "viveu", "vivera", "viveria" etc., mas não "viveríamos".

- **Asterisco ("*")**: Substitui zero ou mais letras quaisquer. Por exemplo, dada a expressão regular "cal*", são reconhecidas as palavras "cal", "calota", "calombo" etc. Para a expressão regular "*cal*", além das palavras já citadas, são também reconhecidas "escala", "recalcado", "fiscal" etc.

- **Sinal de adição ("+")**: Substitui uma ou mais letras quaisquer. Por exemplo, dada a expressão regular "cal+", são reconhecidas as palavras "calota", "calombo" etc., mas não a palavra "cal". Para a expressão regular "+cal+", são reconhecidas as palavras "escala", "recalcado" etc., mas não "fiscal" e "calor".

Esses caracteres somente podem ser empregados no início ou no final de uma expressão regular, não sendo permitidas expressões como "cal*a" ou "ca.a". Os caracteres "." e "?" podem ser combinados, sempre com "." precedendo "?"; a expressão regular "cal..??", por exemplo, permite o reconhecimento de palavras compostas por "cal" seguido de duas, três ou quatro letras quaisquer. Outras combinações de caracteres não são permitidas, sendo inválidas expressões regulares como "cal.*" e "cal?+".

O exemplo apresentado anteriormente para a palavra "lixa" pode ser reformulado fazendo uso de uma expressão regular. O item do dicionário de exceções passa a ser especificado da seguinte forma:

```
*lix+ li-x
```

Por meio desta expressão regular, são reconhecidas todas as palavra da família de "lixa", como "lixas", "lixar", "lixando", "lixado", "lixei", "relixar", "relixaríamos" etc. Palavras como "lixo", "lixeiro" e "lixão", embora não pertençam à família de "lixa", também têm suas transcrições corrigidas por este item do dicionário. Além disso, palavras terminadas em "lix" (como "félix"), nas quais a letra "x" deve ser transcrita como /ks/, não são afetadas pela expressão regular devido ao uso do caractere "+".

Como a informação de divisão silábica é fornecida juntamente com a transcrição fonética no dicionário de exceções, falhas do divisor silábico podem ser corrigidas por meio deste dicionário. O hífen é utilizado para indicar fronteiras entre sílabas; além disso, o caractere "=" (sinal de igual) é empregado em lugar do hífen antes de sílabas tônicas. Por exemplo, as palavras "traidor", "traidora", "traidores" etc., para as quais as regras de divisão silábica falham, podem ser introduzidas no dicionário de exceções da seguinte forma:

```
traidor  tra-i=doP
traidor+ tra-i=do-r
```

A primeira linha é utilizada na transcrição da palavra "traidor"; para as demais variantes dessa palavra, é empregada a transcrição fornecida na segunda linha.

Conforme explicado neste item, o dicionário de exceções associa uma seqüência de caracteres ortográficos a uma seqüência de símbolos fonéticos. Palavras homógrafas com pronúncias diferentes não podem, portanto, ter suas transcrições corrigidas por meio do dicionário de exceções, pois as seqüências de caracteres ortográficos a elas associadas são idênticas, requerendo o uso de informação gramatical (ou mesmo semântica) para a obtenção da pronúncia correta.

CAPÍTULO 3
Processamento prosódico

Neste capítulo são descritos os três módulos que compõem a etapa de processamento prosódico do sistema de conversão texto-fala implementado. O primeiro destes módulos tem por função quebrar a frase de entrada em grupos prosódicos, tomando por base a pontuação do texto e informações sobre a classificação gramatical das palavras. Cada grupo prosódico é tratado isoladamente pelos módulos seguintes, responsáveis pela geração de durações de segmentos e contornos de entonação.

3.1 DETERMINAÇÃO DE FRONTEIRAS PROSÓDICAS

Justificativa para o emprego de grupos prosódicos

Ao formular uma sentença, um falante lhe confere uma *estrutura prosódica* que permite ao ouvinte dividi-la em blocos lógicos, facilitando o seu entendimento. Esses blocos, denominados *grupos prosódicos,* delimitam regiões de uma frase nas quais os parâmetros prosódicos apresentam um comportamento dentro de certa medida independente do verificado nas demais regiões, embora na prática ocorra um certo acoplamento nas fronteiras entre grupos prosódicos adjacentes. Os principais parâmetros prosódicos são a freqüência fundamental (entonação), a duração de segmento e a intensidade sonora; esta última geralmente apresenta pouca influência sobre a inteligibilidade e a naturalidade da fala sintética, não tendo sido considerada neste trabalho.

A divisão de uma sentença em grupos prosódicos permite o processamento em separado de cada grupo; como resultado, os módulos para geração de durações de segmentos e contornos de entonação podem ser elaborados partindo-se do pressuposto de que as suas entradas contêm um único grupo prosódico, simplificando-os significativamente. A escolha de uma definição adequada para o conceito de grupo prosódico depende da técnica empregada na geração dos parâmetros prosódicos. Neste trabalho, esta definição coincide na maioria das vezes com o conceito de oração, ou seja, um membro de uma sentença que pode ser dividido em sujeito e predicado. Trabalhos prévios na área de processamento prosódico para a língua portuguesa [Silva 1995, Silva e Violaro 1996] consideram o sujeito e o predicado de uma oração como grupos prosódicos independentes; a metodologia aqui utilizada para a geração de parâmetros prosódicos, no entanto, dispensa a localização da fronteira entre esses elementos.

Em certos casos, a definição de grupo prosódico como equivalente ao conceito de oração torna-se inadequada. Frases que não contêm verbos (como "Socorro!" ou "Que belo quadro!") não são orações, mas constituem grupos prosódicos. Além disso, orações podem apresentar termos acessórios que quebram a sua continuidade prosódica, como na frase "O homem, assustado, correu."; neste exemplo, são definidos três grupos prosódicos independentes, embora haja somente uma oração. Por fim, orações muito longas e que não apresentem termos acessórios intercalados exigem a inserção de pausas para respiração, quebrando a sua continuidade prosódica e dando origem a múltiplos grupos prosódicos.

Algoritmo para determinação de fronteiras prosódicas

O algoritmo elaborado para a determinação de fronteiras prosódicas (inserido no arquivo "PROSOD.CPP") é bastante simples e mostrou-se eficiente na maioria dos casos. São tomadas por base informações provenientes do classificador gramatical, descrito na seção 2.2, e a pontuação do texto. O primeiro passo do algoritmo consiste em inserir fronteiras prosódicas, acompanhadas por pausas de 300 ms, nos pontos da sentença onde haja vírgulas ou parênteses. Em seguida, são localizados os verbos da sentença (possíveis indicadores de orações) e são inseridas fronteiras prosódicas antes de conjunções, pronomes relativos, pronomes pessoais retos, pronomes possessivos ou artigos, nesta ordem de prioridade, posicionados entre dois verbos. Não foi dado um tratamento específico a pausas para respiração, presentes em orações longas; embora contribuam para aumentar a naturalidade da fala sintetizada, essas pausas não afetam a sua inteligibilidade. Números cardinais e números de telefone são considerados grupos prosódicos independentes. Outras classes de números, como os ordinais, não foram tratadas neste trabalho.

Classificação de grupos prosódicos

Após a determinação das fronteiras prosódicas, cada grupo prosódico recebe uma classificação conforme o tipo de frase em que se encontra e a relação que apresenta com os grupos vizinhos. Essas informações são relevantes para os módulos de geração de durações de segmentos e contornos de entonação, descritos nas próximas seções deste capítulo. São definidas seis categorias para a classificação geral de grupos prosódicos, cada uma representada por abreviações de três letras:

- **Declarativo (DEC):** Grupos prosódicos pertencentes a frases não terminadas por ponto de interrogação e cujos verbos não estejam no modo imperativo.

- **Interrogativo (INT):** Grupos prosódicos pertencentes a frases terminadas por ponto de interrogação.
- **Imperativo (IMP):** Grupos prosódicos pertences a frases não terminadas por ponto de interrogação e que contêm ao menos um verbo no modo imperativo.
- **Indicativo (IND):** Grupos prosódicos pertencentes a frases não terminadas por ponto de interrogação e que não contêm verbos.
- **Número cardinal (NUC):** Grupo prosódico correspondente a um número cardinal.
- **Número de telefone (NTL):** Grupo prosódico correspondente a um número de telefone.

Além disso, são definidas oito subcategorias, representadas por abreviações de duas letras, que podem aplicar-se a qualquer uma das categorias acima:

- **Exclamativa (EX):** Grupos prosódicos pertencentes a frases terminadas por ponto de exclamação (eventualmente acompanhado por outro símbolo finalizador de sentença).
- **Afirmativa (AF):** Grupos prosódicos que não contêm palavras negativas (não, nunca, jamais etc.).
- **Negativa (NG):** Grupos prosódicos que contêm palavras negativas.
- **Início de frase (IF):** Grupos prosódicos localizados no início de uma frase.
- **Meio de frase (MF):** Grupos prosódicos localizados no meio de uma frase.
- **Final de frase (FF):** Grupos prosódicos localizados no final de uma frase.
- **Precedido por pausa (PP):** Grupos prosódicos que são precedidos por uma pausa.
- **Sucedido por pausa (SP):** Grupos prosódicos que são sucedidos por uma pausa.

As abreviações correspondentes a cada categoria e subcategoria são utilizadas nos dicionários de durações e de contornos de entonação, descritos nas próximas seções deste capítulo.

3.2 GERAÇÃO DE CONTORNOS DE ENTONAÇÃO

Técnicas para geração de contornos de entonação

Várias técnicas vêm sendo utilizadas em sistemas de conversão texto-fala para a geração de contornos de entonação. Uma abordagem possível consiste na utilização de um

contorno-base padrão, aplicado sobre todas as frases sintetizadas, sobre o qual são efetuadas alterações com base em um conjunto de regras que se vale de informações sintáticas e do conhecimento de fronteiras entre palavras e sílabas, além da informação sobre a posição da sílaba tônica em cada palavra. Descrições de implementações dessa abordagem podem ser encontradas em [Allen et al. 1987, Silva 1995].

Outra possibilidade consiste no emprego de contornos de entonação extraídos de elocuções naturais, que são ajustados às particularidades do texto a sintetizar. Esta técnica apresenta como vantagem o fato de os contornos naturais já trazerem consigo uma grande quantidade de detalhes que, na abordagem anterior, deveriam ser gerados através de regras. No entanto, é justamente a ausência de um controle preciso sobre a adaptabilidade desses detalhes às frases sintetizadas que constitui a principal dificuldade de implementação desta técnica. Descrições de sistemas de processamento prosódico baseados nessa abordagem podem ser encontradas em [Emerard et al. 1992, Traber 1992].

O módulo para geração de contornos de entonação implementado neste trabalho (contido no arquivo "PROSOD.CPP") utiliza a técnica de adaptação de contornos naturais, conforme descrito nos itens seguintes.

O dicionário de contornos de entonação

A partir de curvas de freqüência fundamental extraídas de um conjunto de trinta elocuções proferidas por um locutor, foi criado o *dicionário de contornos de entonação* (arquivo "PROSOD.DIC"). Esse dicionário encontra-se em um arquivo-texto lido pelo processador prosódico. Cada item do dicionário, correspondente a um grupo prosódico, contém as seguintes informações: (1) classificação geral do grupo prosódico, utilizando as categorias descritas na seção 3.1; (2) classe gramatical de cada palavra da elocução, obtida utilizando-se o classificador gramatical descrito na seção 2.2; (3) contorno de freqüência fundamental, em hertz, ao longo da elocução; (4) número de fones em cada palavra; e (5) localização de sons vocálicos pertencentes a sílabas tônicas lexicais e à sílaba tônica frasal (se presente). A seguir encontra-se um exemplo de declaração de um item do dicionário:

```
DEC AF
[20][0,113,26,113,57,112]a ART DE3 MA3 SI3
[18]a[17][14,118,94,92]t[10]a[2][46,99,73,98]a SUB MA3 SI3 CM3
[-8][44,102]a[0][72,121]a[7][3,121,66,114]t VER ID3 PS3 EL3 SI3
[15]a[17][16,127,52,110]T[18][55,103]a[19][56,102]a ADJ MA3 SI3
FIM
```

A primeira linha do exemplo acima indica a classificação geral do grupo prosódico (no caso, declarativo afirmativo, conforme as abreviações definidas na seção 3.1). Cada uma

das quatro linhas seguintes corresponde a uma palavra pertencente ao grupo prosódico. Como a palavra em si é irrelevante, cada um de seus fones é representado pela letra "a" (minúscula), exceto a vogal principal de uma sílaba tônica lexical, representada pela letra "t" (minúscula), e a vogal principal de uma sílaba tônica frasal, representada pela letra "T" (maiúscula). Cada letra é precedida por duas seqüências de números entre colchetes. A primeira seqüência, que sempre possui apenas um número, é obrigatória e corresponde aos dados do dicionário de durações, descrito na próxima seção. A segunda seqüência, opcional, corresponde a pontos do contorno de entonação localizados dentro das fronteiras do fone que a segue. Após o último fone, é especificada a classificação gramatical da palavra, obtida através do classificador gramatical descrito na seção 2.2. O identificador "FIM" indica que a definição do item do dicionário chegou ao seu final.

A seqüência de números que define um trecho do contorno de entonação corresponde a pares *(tempo, valor de freqüência fundamental),* com o tempo dado em relação ao início do fone que a segue e em termos de porcentagem da duração desse fone. Através de seqüências desse tipo, são especificados diversos pontos do contorno de entonação ao longo da elocução, obtendo-se valores intermediários por interpolação linear. Para o exemplo anterior, o contorno de entonação ao longo de toda a elocução tem a forma mostrada na figura 3.1.

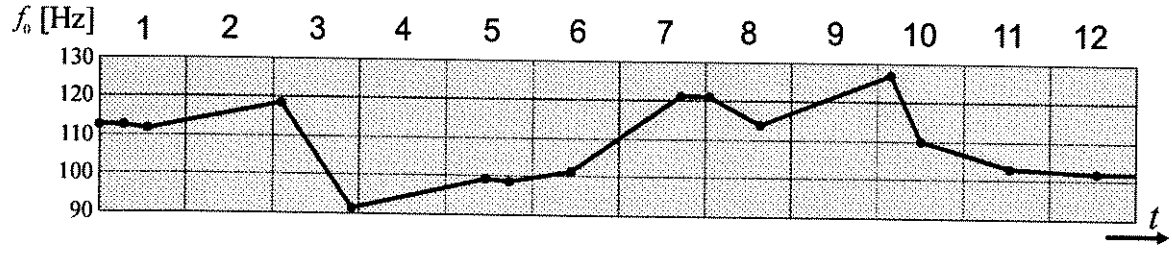

Figura 3.1 Variação da freqüência fundamental (f_0) em função do tempo para o exemplo de declaração de um item do dicionário de contornos de entonação. Cada divisão numerada corresponde a um fone. Como as durações absolutas dos fones não são armazenadas no dicionário, as doze divisões foram traçadas com larguras idênticas.

Seleção de um contorno do dicionário

O primeiro passo efetuado pelo gerador de contornos de entonação consiste em selecionar um item do dicionário cujo contorno se adapte bem às características do grupo prosódico de entrada. Essa seleção é feita com base na classificação gramatical prévia, definindo-se um índice que mede a similaridade entre características gramaticais do grupo prosódico de entrada e de cada item do dicionário; ao final, escolhe-se o item com índice mais elevado. Esse índice é calculado através das seguintes regras:

1. Se uma determinada classe gramatical presente no grupo prosódico de entrada estiver presente também no item do dicionário, soma-se o seu peso (característico de cada classe gramatical) ao índice de similaridade.
2. Se uma determinada classe gramatical presente no grupo prosódico de entrada não estiver presente no item do dicionário, subtrai-se o seu peso do índice de similaridade.
3. Se uma determinada classe gramatical presente no item do dicionário não estiver presente no grupo prosódico de entrada, subtrai-se o seu peso do índice de similaridade.
4. Se duas classes gramaticais estiverem presentes na mesma ordem no grupo prosódico de entrada e no item do dicionário, seus pesos são somados novamente ao índice de similaridade.

A primeira regra favorece os itens do dicionário que apresentem as mesmas classes gramaticais observadas no grupo prosódico de entrada, enquanto a segunda e a terceira regras desfavorecem os itens que apresentem classes gramaticais diferentes das observadas no grupo prosódico de entrada. A quarta regra favorece itens que apresentem seqüências de classes gramaticais em ordem similar à verificada no grupo prosódico de entrada. Na aplicação da primeira regra, caso sejam encontradas em um item do dicionário duas ou mais palavras com classificação gramatical idêntica à de uma palavra do grupo prosódico de entrada, é utilizada aquela cuja posição no item é mais semelhante à posição da palavra no grupo prosódico de entrada.

Os pesos associados a cada classe gramatical foram escolhidos da seguinte forma: palavras classificadas como de conteúdo[1] (verbos, substantivos, adjetivos, advérbios, interjeições, palavras denotativas) recebem um peso arbitrariamente escolhido como 30, enquanto palavras classificadas como funcionais (conjunções, preposições, artigos, numerais, pronomes) recebem um peso igual a 15. Além disso, esse peso pode ser reduzido com base na análise das subclasses gramaticais:

- **Quando a classe gramatical estiver presente tanto no grupo prosódico de entrada como no item do dicionário:** se não houver coincidência entre subclasses, o peso será reduzido por um fator de 5; se houver coincidência entre subclasses, será tomado o me-

[1] Palavras essenciais à compreensão de uma sentença são definidas como *palavras de conteúdo,* enquanto *palavras funcionais* são aquelas cuja eliminação não inviabiliza o entendimento do sentido geral de uma sentença [Allen et al. 1987]. É claro que, em dados contextos, uma palavra normalmente classificada como funcional pode assumir importância fundamental na compreensão de uma sentença, caracterizando-se como palavra de conteúdo, o que é relativamente comum sobretudo com pronomes e numerais. Como, no entanto, o sistema não dispõe de recursos para análise semântica, optou-se por utilizar uma classificação rígida e independente do contexto.

nor dentre os coeficientes a elas associados, reduzindo-se o peso por um fator de 3 para coeficiente igual a 1 ou por um fator de 2 para coeficiente igual a 2, não havendo redução alguma caso ambos os coeficientes sejam iguais a 3.

- **Quando a classe gramatical estiver presente somente no grupo prosódico de entrada ou somente no item do dicionário:** será tomada a subclasse que possui maior coeficiente, reduzindo-se o peso por um fator de 3 para coeficiente igual a 1 ou por um fator de 2 para coeficiente igual a 2, não havendo redução alguma caso o coeficiente seja igual a 3.

O processo de seleção de contornos leva também em conta as categorias e subcategorias do grupo prosódico de entrada e dos itens do dicionário. Para que um item do dicionário possa ser escolhido, sua categoria deve necessariamente coincidir com a do grupo prosódico de entrada. Além disso, para cada subcategoria não coincidente, o índice de similaridade é reduzido por um fator de 0,75.

Se houver empate entre os índices de similaridade obtidos para dois ou mais itens do dicionário, o critério de escolha passa a ser a quantidade de sílabas tônicas presentes em cada item. É escolhido o item que apresentar o número de sílabas tônicas mais próximo do encontrado no grupo prosódico de entrada (e preferencialmente maior do que este último). Esse critério se baseia no fato de as sílabas tônicas desempenharem um papel importante no posterior ajuste do contorno selecionado, como descrito mais à frente nesta seção. Caso ainda haja empate após a aplicação deste critério, opta-se pelo contorno que estiver definido anteriormente no dicionário; deste modo, contornos considerados de uso mais freqüente devem ser definidos antes de contornos de utilização mais rara.

Ajuste do contorno selecionado

Uma vez selecionado um item do dicionário, passa-se ao ajuste do contorno às particularidades do grupo prosódico de entrada. Esse ajuste consiste em uma compressão ou expansão temporal de segmentos do contorno limitados pelas vogais principais das sílabas tônicas e pelo início e final das sentenças. Os trechos do contorno correspondentes às vogais principais de sílabas tônicas não são alterados, de modo a preservar as variações notáveis de freqüência fundamental geralmente observadas nesses fones. A figura 3.2 ilustra o processo de ajuste de contornos de entonação para o caso mais simples, no qual o grupo prosódico de entrada e o item selecionado do dicionário apresentam um número idêntico de sílabas tônicas. Os fatores de escala são dados pela relação entre a quantidade de fones contidos em segmentos correspondentes do contorno no item do dicionário e no grupo prosódico de entrada.

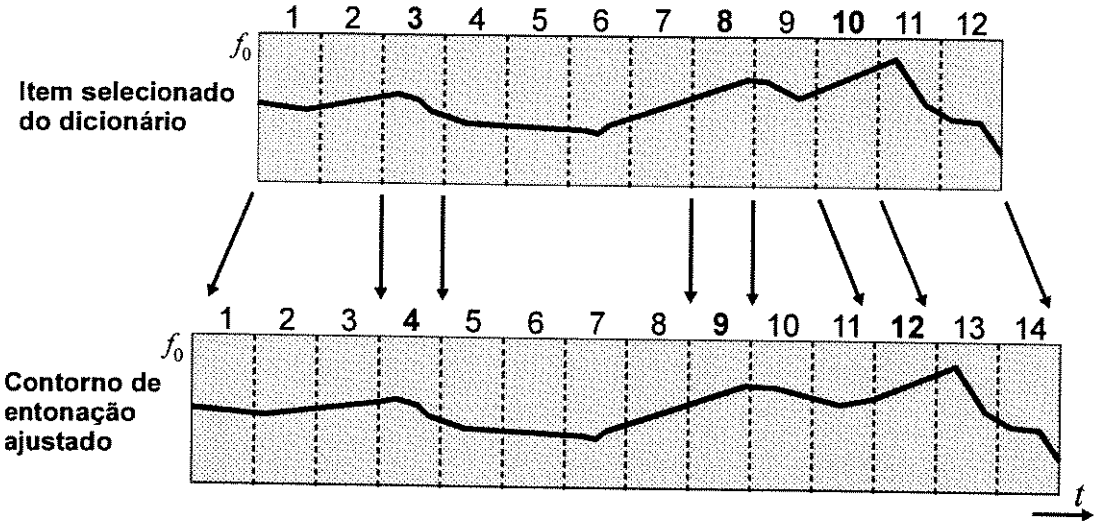

Figura 3.2 Ilustração do processo de ajuste de contornos de freqüência fundamental (f_0) para o caso em que há coincidência entre o número de sílabas tônicas no grupo prosódico de entrada e no item selecionado do dicionário. Cada divisão numerada corresponde a um fone. Os números em negrito indicam as vogais principais das sílabas tônicas.

Caso não haja coincidência entre a quantidade de sílabas tônicas no grupo prosódico de entrada e no item selecionado do dicionário, é necessário desprezar sílabas tônicas até atingir-se a igualdade. As sílabas tônicas a desprezar são escolhidas por meio de comparações entre classes gramaticais no grupo prosódico de entrada e no item do dicionário. As comparações têm início nas palavras iniciais de cada sentença. Havendo coincidência entre suas classes gramaticais, avança-se às palavras seguintes e segue-se com as comparações; em caso contrário, despreza-se a sílaba tônica da palavra pertencente à sentença que as possui em excesso e avança-se à palavra seguinte, prosseguindo-se com as comparações. O processo continua até desprezar-se o número necessário de sílabas tônicas. Caso se chegue ao final da sentença que possui sílabas tônicas em excesso antes de atingir-se o equilíbrio, retorna-se à primeira palavra desta sentença e desprezam-se seqüencialmente tantas sílabas tônicas quantas forem necessárias. Se, por outro lado, for atingido o final da sentença que possui menos sílabas tônicas, despreza-se seqüencialmente a partir da última palavra comparada o número necessário de sílabas tônicas na sentença que as possui em excesso, retornando-se à primeira palavra caso o final dessa sentença seja atingido.

Tratamento de interações entre grupos prosódicos

Como os grupos prosódicos são tratados independentemente pelo módulo de geração de contornos de entonação, ocorre um problema de continuidade do contorno na fronteira

entre grupos prosódicos adjacentes. Quando há uma pausa entre os grupos prosódicos, esse problema não é crítico, pois há uma quebra natural na continuidade do contorno; se não houver pausa, no entanto, uma mudança brusca no valor da freqüência fundamental prejudicaria a naturalidade da fala sintetizada.

Para evitar problemas de descontinuidade da freqüência fundamental quando não há pausa entre grupos prosódicos adjacentes, efetua-se um deslocamento dos trechos de contorno de modo que o valor final do primeiro trecho coincida com o valor inicial do segundo, ambos iguais à média geométrica dos valores originais. Esse deslocamento é efetuado em escala logarítmica (correspondendo em escala linear à multiplicação dos trechos de contorno por constantes), de modo a manter as proporções do contorno inalteradas em termos da resposta do ouvido humano. A figura 3.3 ilustra esse processo.

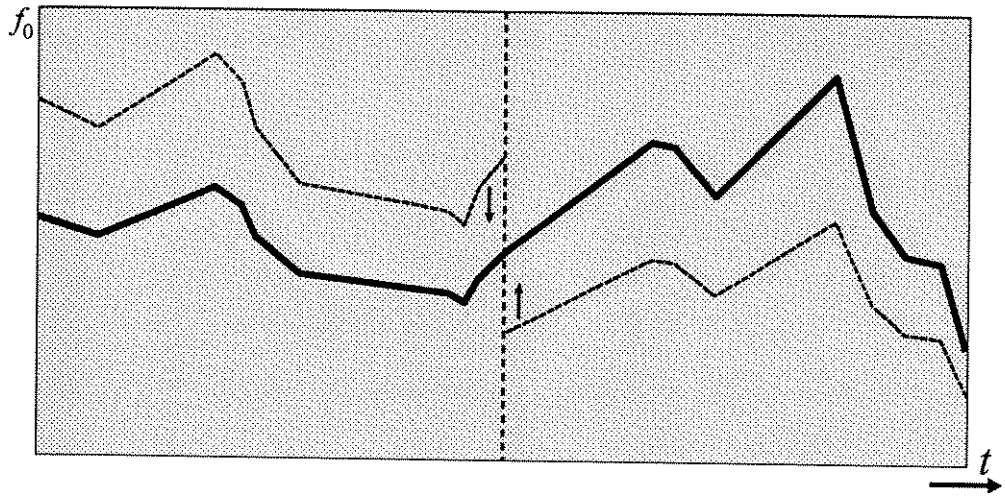

Figura 3.3 Ilustração em escala linear do processo de deslocamento de contornos de freqüência fundamental (f_0) na fronteira entre grupos prosódicos adjacentes. As curvas tracejadas correspondem aos contornos originais e a curva contínua corresponde ao contorno único resultante ao final do processo.

Caso haja três ou mais grupos prosódicos adjacentes e não separados por pausas, o procedimento descrito acima é inicialmente efetuado entre os dois primeiros trechos de contorno, que passam a ser considerados como um trecho único. O procedimento é então repetido entre este trecho e o seguinte, formando um novo trecho único. O processo se repete até atingir-se uma pausa ou o final da frase.

Geração de contornos de entonação para números

A pronúncia de números apresenta características prosódicas particulares e bastante regulares. Números cardinais e números de telefone são considerados grupos prosódicos

independentes, recebendo um tratamento particular por parte do módulo de geração de contornos de entonação. Seções específicas do dicionário de contornos são destinadas a esses números. Números ordinais não foram abordados neste trabalho, mas o seu tratamento seria muito similar ao empregado para os números cardinais.

3.3 GERAÇÃO DE DURAÇÕES DE SEGMENTOS

Técnicas para geração de durações de segmentos

A atribuição de durações a cada fone (ou *segmento*) da fala é parte fundamental do processamento prosódico. O emprego de durações fixas, independentes do contexto em que um fone se insere, torna a fala artificial e muitas vezes prejudica a sua compreensão. As vogais principais de sílabas tônicas, por exemplo, apresentam durações maiores do que as verificadas em outros contextos, e são por vezes fundamentais para a compreensão de uma palavra.

Uma técnica bastante divulgada para a geração de durações de segmentos para a língua inglesa é a desenvolvida por Klatt [Allen et al. 1987]. Essa técnica faz uso de um conjunto de regras que agem sucessivamente sobre as durações, inicialmente estabelecidas em valores intrínsecos característicos de cada fone. Conforme o contexto em que se insere um fone, as regras podem aumentar ou diminuir a sua duração, respeitando limites mínimos previamente fixados. Os sistemas de geração automática de durações de segmentos para a língua portuguesa descritos em [Egashira 1992, Silva 1995] utilizam adaptações desta técnica.

Outra possibilidade para a geração de durações de segmentos é a utilização direta de dados extraídos de elocuções naturais. Esta foi a abordagem empregada no presente trabalho, tendo sido desenvolvido um algoritmo próprio baseado na hipótese de que efeitos globais e locais sobre as durações podem ser isolados. Por efeitos globais entende-se a influência do grupo prosódico como um todo sobre a duração de um fone, levando em conta, por exemplo, a posição do fone no grupo prosódico e a classificação desse grupo (afirmativo, interrogativo etc.). Os efeitos locais, por outro lado, respondem pela influência da vizinhança imediata de um fone sobre a sua duração, considerando, por exemplo, os fones vizinhos e os acentos lexicais. Tomando-se dois grupos prosódicos de estrutura gramatical similar, os efeitos globais sobre as durações observados em cada um deles muito provavelmente são também similares. Entretanto, o mesmo não se aplica aos efeitos locais, pois eles têm pouca relação com a estrutura gramatical do grupo prosódico. Os próximos itens descrevem em detalhe o algoritmo utilizado no sistema.

O dicionário de durações

O algoritmo desenvolvido para a geração de durações de segmentos utiliza um *dicionário de durações* (arquivo "PROSOD.DIC") que contém, para cada fone de uma elocução natural, o desvio percentual da duração em relação ao seu valor médio, normalizado pelo desvio padrão dessa duração. A seqüência de desvios percentuais normalizados pode ser vista como a representação amostral de uma curva contínua que fornece, para cada instante de tempo, a quantidade pela qual a duração deve diferir de seu valor médio. Essa curva, denominada *curva de variação de duração,* é ajustada ao grupo prosódico de entrada de forma similar à descrita na seção anterior para os contornos de entonação. Os desvios obtidos através da curva ajustada correspondem aos efeitos globais sobre as durações para o grupo prosódico de entrada. É claro que as variações decorrentes de efeitos locais também estão presentes ao extrair-se a seqüência de desvios percentuais normalizados de uma elocução natural; para evitar as distorções daí decorrentes, a curva de variação de duração é suavizada antes que suas amostras sejam introduzidas no dicionário de durações.

O dicionário de durações é armazenado juntamente com o dicionário de contornos de entonação, ambos utilizando o mesmo conjunto de sentenças naturais. O exemplo da seção 3.2, no qual são declarados um item do dicionário de contornos de entonação e um item do dicionário de durações, encontra-se reproduzido a seguir.

```
DEC AF
[20][0,113,26,113,57,112]a ART DE3 MA3 SI3
[18]a[17][14,118,94,92]t[10]a[2][46,99,73,98]a SUB MA3 SI3 CM3
[-8][44,102]a[0][72,121]a[7][3,121,66,114]t VER ID3 PS3 EL3 SI3
[15]a[17][16,127,52,110]T[18][55,103]a[19][56,102]a ADJ MA3 SI3
FIM
```

A primeira e a última linhas dessa declaração já foram comentadas na seção 3.2, bem como as informações referentes ao dicionário de contornos de entonação. Os desvios percentuais normalizados das durações são especificados entre colchetes antes dos símbolos "a", "t" e "T" (que representam os fones da elocução natural), eventualmente seguidos por dados do dicionário de contornos de entonação, também entre colchetes. Como os desvios percentuais normalizados normalmente se encontram na faixa entre –1 e 1, eles são multiplicados por 100 antes de serem introduzidos no dicionário, permitindo a utilização de números inteiros.

Seleção e ajuste de curvas de variação de duração

O critério para seleção de itens do dicionário de durações é idêntico ao utilizado para o dicionário de contornos de entonação, baseando-se na avaliação da similaridade gramatical entre o grupo prosódico de entrada e cada item do dicionário. Como o dicionário de durações é elaborado a partir das mesmas elocuções naturais utilizadas para o dicionário de contornos de entonação, o resultado da seleção efetuada pelo gerador de contornos de entonação automaticamente determina o item do dicionário de durações que melhor se adapta às características do grupo prosódico de entrada.

O processo de ajuste de uma curva de variação de duração ao grupo prosódico de entrada é similar ao descrito na seção anterior para os contornos de entonação, mas apresenta algumas particularidades. As curvas são divididas em segmentos limitados pelas vogais principais das sílabas tônicas e pelo início e final da elocução. Como no caso do ajuste de contornos de entonação, é necessário que haja igualdade entre o número de sílabas tônicas no grupo prosódico de entrada e no item selecionado do dicionário; caso essa igualdade não se verifique, as mesmas sílabas tônicas desprezadas no ajuste do contorno de entonação também o são aqui. O ajuste da curva de duração consiste em um processo de reamostragem de seus segmentos. O número de amostras presentes em um segmento da curva original é igual ao número de fones no trecho correspondente do item do dicionário, pois um valor de desvio normalizado de duração é fornecido para cada fone; já para um segmento da curva ajustada, o número de amostras passa a ser igual ao número de fones no trecho correspondente do grupo prosódico de entrada. As amostras da curva ajustada fornecem diretamente os desvios percentuais normalizados de duração para cada fone do grupo prosódico de entrada. A figura 3.4 ilustra o processo de reamostragem de uma curva de variação de duração.

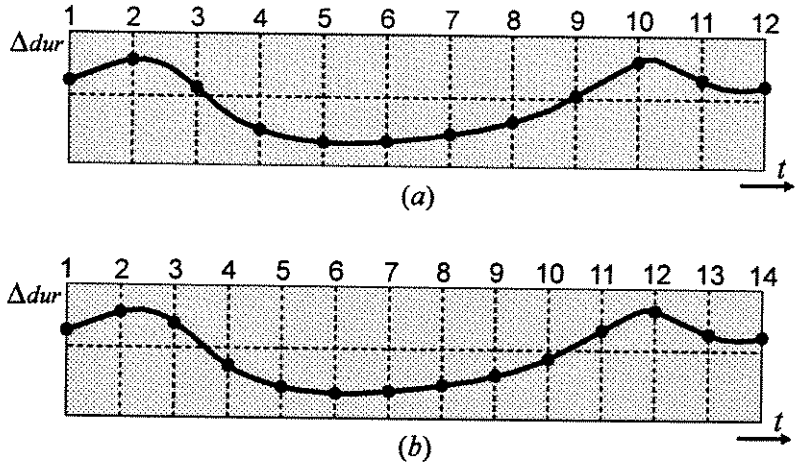

Figura 3.4 Ilustração do processo de reamostragem de curvas de variação de duração (*Δdur*): (a) amostras originais obtidas do dicionário de durações; (b) amostras resultantes após o processo de reamostragem. Cada linha vertical numerada corresponde a um fone, indicando um ponto em que é tomada uma amostra da curva.

A reamostragem é realizada através de cálculo direto das novas amostras a partir das antigas. Supondo que o segmento da curva original possua N_1 amostras (representadas por $x_1[n_1]$, com n_1 variando entre 0 e N_1-1) e o segmento ajustado deva possuir N_2 amostras (representadas por $x_2[n_2]$, com n_2 variando entre 0 e N_2-1), cada amostra do segmento ajustado é obtida através da fórmula de interpolação:

$$x_2[n_2] = \sum_{n_1=0}^{N_1-1} x_1[n_1] \operatorname{sinc}\left(n_2 \frac{N_1-1}{N_2-1} - n_1 \right)$$

onde sinc x é definido como:

$$\operatorname{sinc} x = \begin{cases} \dfrac{\operatorname{sen} \pi x}{\pi x} & \text{para } x \neq 0 \\ 1 \text{ para } x = 0 \end{cases}$$

Não é necessário usar a fórmula de interpolação para a primeira e a última amostras da curva ajustada, pois elas são sempre iguais, respectivamente, à primeira e à última amostras da curva original.

Ao contrário do que ocorre com os contornos de entonação, não há para as durações de segmentos problemas relacionados à interação entre grupos prosódicos adjacentes, mesmo que não separados por pausa. Uma eventual descontinuidade na curva de variação de duração na fronteira entre grupos prosódicos em princípio não prejudica a naturalidade da fala sintetizada, desde que o seu ritmo geral seja mantido (isto é, não se passe bruscamente de uma fala lenta para uma fala rápida). Como todas as elocuções naturais utilizadas na elaboração do dicionário de durações foram pronunciadas mantendo-se um ritmo aproximadamente constante, as chances de descontinuidade perceptível na fronteira entre grupos prosódicos são mínimas.

Uma vez que as estruturas gramaticais do grupo prosódico de entrada e do item selecionado do dicionário de durações são similares, as posições das sílabas tônicas frasais de cada uma dessas sentenças provavelmente coincidem. Desse modo, a sílaba tônica frasal do grupo prosódico de entrada pode ser determinada com base na posição da palavra que contém a sílaba tônica frasal do item do dicionário (cuja vogal principal é indicada pela letra "T"). Esta informação é utilizada nas regras para efeitos locais, discutidas no próximo item.

Regras para efeitos locais nas durações

Após a introdução dos efeitos globais sobre as durações, passa-se à determinação dos efeitos locais. É usado um conjunto de regras que se baseia na análise da vizinhança de um fone, levando em conta os fones a ele adjacentes, a sua localização no interior da palavra e a sua posição em relação às sílabas tônicas. Essas regras foram inspiradas no modelos apresentados em [Allen et al. 1987] e [Silva 1995].

O conjunto de regras para efeitos locais encontra-se listado a seguir:

1. Segmentos de palavras com mais do que três sílabas têm suas durações reduzidas por um fator de 0,92 para vogais e de 0,95 para consoantes, exceto /r/ e /P/.
2. Segmentos da sílaba final de uma palavra têm suas durações aumentadas por um fator de 1,08 para vogais e de 1,05 para consoantes, exceto /P/.
3. Consoantes em início de palavra têm suas durações aumentadas por um fator de 1,05.
4. Segmentos em sílabas pós-tônicas ou monossílabos átonos têm suas durações reduzidas por um fator de 0,92 para vogais, exceto /y/, /A/, /w/, /ÿ/ e /Ü/, e de 0,95 para consoantes, exceto /r/ e /P/.
5. Segmentos em sílabas tônicas lexicais têm suas durações aumentadas por um fator de 1,3 para vogais e de 1,1 para consoantes, exceto /r/ e /P/.
6. Segmentos em sílabas tônicas frasais têm suas durações aumentadas por um fator de 1,5 para vogais e de 1,15 para consoantes, exceto /r/ e /P/.
7. Vogais têm suas durações multiplicadas por um fator de 1,08 quando sucedidas por fricativa sonora ou precedidas por oclusiva surda, de 1,05 quando sucedidas por oclusiva sonora, de 0,95 quando sucedidas por consoante nasal e de 0,92 quando sucedidas ou precedidas por vogal ou sucedidas por oclusiva surda.
8. Consoantes, com exceção de /r/ e /P/, têm suas durações reduzidas por um fator de 0,95 quando precedidas ou sucedidas por consoante.

As variações impostas às durações por estas regras são em geral menores do que as empregadas no modelo descrito em [Silva 1995]. Isto se deve em parte ao fato de os efeitos globais aplicados anteriormente sobre as durações já imporem um certo nível de variação ao qual são sobrepostos os efeitos locais. Além disso, testes realizados com o modelo citado levaram a variações de duração perceptualmente exageradas, possivelmente em decorrência das particularidades da voz tomada como base para a elaboração das regras de síntese.

Geração de durações de segmentos para números

Como no caso dos contornos de entonação, são definidas curvas de variação de duração específicas para números cardinais e números de telefone no dicionário de durações. Para os efeitos locais, são utilizadas as mesmas regras aplicadas aos demais grupos prosódicos.

3.4 FORMATO DE SAÍDA DO PROCESSADOR PROSÓDICO

O processador prosódico produz como saída uma transcrição fonética à qual são acrescidos parâmetros prosódicos (durações e freqüência fundamental). O formato dessa saída é compatível com a entrada do módulo conversor, descrito na seção 4.5.

A duração de um fone e o valor de freqüência fundamental (em hertz) ao final deste fone são especificados entre colchetes, separados por vírgula, imediatamente antes do símbolo fonético correspondente. Não é obrigatório o fornecimento de um valor de freqüência fundamental. A transcrição "[70,120]i[80,125]e[80]é[100,115]a", por exemplo, corresponde ao contorno de entonação mostrado na figura 3.5.

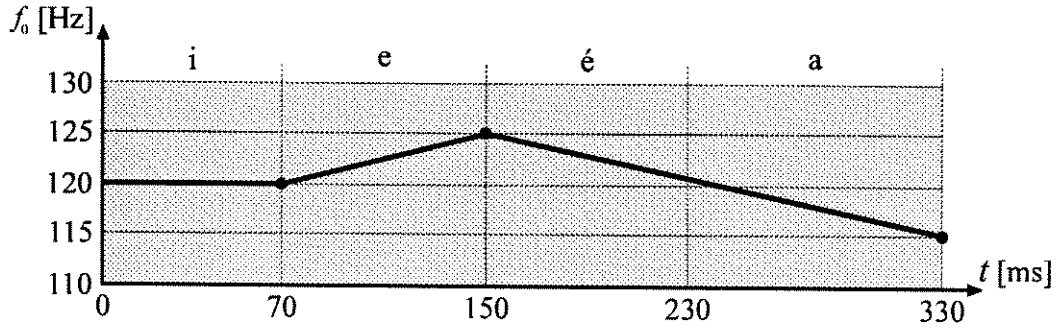

Figura 3.5 Contorno de freqüência fundamental (f_0) correspondente à transcrição "[70,120]i[80,125]e[80]é[100,115]a".

Caso deva haver variação na inclinação da curva de freqüência fundamental em meio a um fone, são especificadas diversas durações parciais (tantas quantas forem as regiões de inclinação diferenciada) que se alternam com valores de freqüência fundamental, separados por vírgulas, e a duração total do fone é obtida pela soma das durações parciais. A transcrição "[70,120]i[60,125,20,125]e[60,120,20]é[30,120,30,115,40,115]a", por exemplo, corresponde ao contorno de entonação e ao padrão de durações mostrados na figura 3.6.

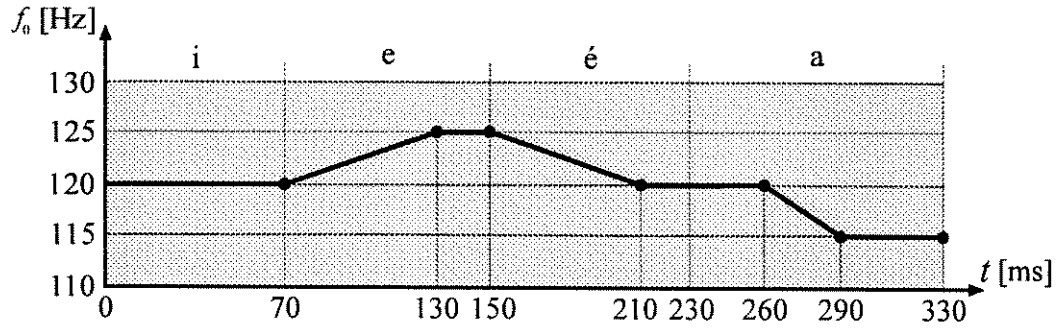

Figura 3.6 Contorno de freqüência fundamental (f_0) correspondente à transcrição "[70,120]i[60,125,20,125]e[60,120,20]é[30,120,30,115,40,115]a".

Uma transcrição com parâmetros prosódicos no formato descrito acima pode também ser fornecida diretamente pelo usuário ao módulo conversor (especificando-se a opção "i7" ou "u7" ao acionar o sistema, conforme descrito na seção 1.4), permitindo a síntese com durações de segmentos e contornos de entonação arbitrários. Caso não seja fornecido um valor de duração para um fone, é usada a duração média especificada no arquivo de dados para síntese ("DADOS.TAB", descrito no próximo capítulo).

CAPÍTULO 4

Síntese do sinal de fala

Este capítulo descreve o *módulo conversor* do sistema de conversão texto-fala, responsável pela geração de um sinal de fala a partir de uma transcrição fonética acrescida de parâmetros prosódicos. É descrito ainda o *sintetizador de Klatt,* responsável pela síntese do sinal de fala, bem como os modelos empregados na obtenção dos valores dos parâmetros de controle do sintetizador em função do tempo para as várias classes de fones. Por fim, são discutidos a *Linguagem para Descrição de Regras* (LDR), através da qual são implementados os modelos citados, e o *compilador de regras,* responsável pela compilação de programas LDR.

4.1 TÉCNICAS PARA SÍNTESE DE FALA

A técnica mais elementar para a síntese de fala consiste na gravação de palavras (ou mesmo frases inteiras) e na sua concatenação para a formação de sentenças maiores. Esse método é muito utilizado em sistemas de atendimento automático por telefone, nos quais o usuário fornece informações (como um número de conta bancária) através do disco ou das teclas e recebe respostas (o saldo de sua conta, por exemplo) por meio de fala sintetizada. A principal desvantagem desta técnica é a artificialidade das fronteiras entre palavras, pois não há uma continuidade entre o final de uma palavra e o início da próxima. Além disso, geralmente é necessário gravar cada palavra em vários contextos, de modo que se possa escolher a versão que melhor se adapte, em termos de entonação, à frase a sintetizar. Dada a grande quantidade de memória requerida para o armazenamento das gravações de palavras, esta técnica somente é viável para sistemas de síntese com vocabulário limitado.

Métodos concatenativos também podem ser empregados na implementação de sistemas de síntese de fala sem restrição de vocabulário. Neste caso, devem ser usadas unidades de concatenação menores do que palavras. Uma unidade que se adapta satisfatoriamente ao processo de concatenação é o difone, definido como metade de um fone seguido pela metade do fone seguinte. Embora o número de difones existentes em uma língua não seja muito elevado, seria necessário, em princípio, gravá-los em diversos contextos, obtendo-se uma variedade de durações e entonações suficiente para que se tenha um controle adequado sobre a prosódia ao longo da síntese. É possível, no entanto, variar a duração e a freqüência fundamental de um difone utilizando técnicas de processamento de sinal. A técnica PSOLA (*Pitch Synchronous Overlap and Add*), por exemplo, possibilita variar os parâmetros prosódicos de um difone, dentro de certos limites, através do deslocamento, da replicação e

da eliminação de trechos da forma de onda [Silva 1995, Egashira e Violaro 1995, Violaro et al. 1996]. Essa técnica vem sendo empregada com sucesso no desenvolvimento de sistemas de conversão texto-fala sem restrição de contexto.

Outra abordagem comum na implementação de sintetizadores de fala consiste na utilização de técnicas para síntese de formantes. Esses sintetizadores recebem como entrada um conjunto de parâmetros de controle atualizado periodicamente e produzem como saída as amostras do sinal de fala. Os parâmetros incluem amplitudes de fontes, freqüência fundamental de fontes de vozeamento e freqüências e larguras de banda dos formantes, entre outros. Não há uma preocupação em reproduzir-se todas as características acústicas de um sinal de fala natural, mas somente aquelas que se mostram perceptualmente mais relevantes [Allen et al. 1987]. A principal dificuldade na utilização desta técnica consiste na obtenção dos valores dos parâmetros de controle do sintetizador, o que em geral é realizado por meio de um conjunto de regras, denominadas *regras de síntese,* que atua com base na transcrição fonética do texto a sintetizar. A etapa de elaboração dessas regras é bastante demorada e trabalhosa, o que torna a técnica concatenativa (utilizando difones, por exemplo) mais atrativa caso se deseje resultados a curto prazo. No entanto, a total liberdade na manipulação dos parâmetros de controle do sintetizador dá a um sistema baseado na síntese de formantes potencial para atingir um nível de qualidade superior ao de um sistema concatenativo. Além disso, sistemas concatenativos em geral requerem muito mais memória de massa do que sistemas baseados na síntese de formantes.

4.2 O SINTETIZADOR DE FORMANTES DE KLATT

Para a etapa de síntese do sinal de fala, o sistema de conversão texto-fala implementado utiliza o *sintetizador de formantes de Klatt* (incorporado ao módulo conversor no arquivo "CONV.CPP"). A seguir será feita uma breve descrição deste sintetizador. Descrições mais detalhadas podem ser encontradas em [Klatt 1980, Allen et al. 1987, Nagle e Chiquito 1993].

O sintetizador de formantes de Klatt baseia-se na teoria acústica da produção da fala elaborada por Fant [Fant 1960]. O processo físico da produção da fala é modelado matematicamente por meio de três componentes básicos: fontes, característica de filtragem do trato vocal e característica de radiação para o meio externo. O comportamento do sintetizador é controlado por meio de um conjunto de parâmetros que lhe é fornecido a intervalos periódicos. O modelo original de Klatt apresenta 39 parâmetros de controle; foram ainda acrescentados dois novos parâmetros com o objetivo de aumentar a naturalidade da fala sintetizada, conforme descrito mais adiante neste item, chegando-se a um total de 41 parâmetros.

As fontes dividem-se em duas classes: fontes de vozeamento e fontes de ruído. As fontes de vozeamento são geradas a partir da filtragem passa-baixas de um trem de im-

pulsos. A fonte de *vozeamento normal* é obtida por meio de uma única filtragem, enquanto a fonte de *vozeamento quase senoidal* requer uma filtragem adicional. A composição da excitação final de vozeamento é definida pelos controles de amplitude de cada ramo (parâmetros AV e AVS, respectivamente). As fontes de ruído, por sua vez, são obtidas a partir de um gerador de números pseudo-aleatórios. A fonte de *ruído de aspiração* (cuja amplitude é controlada pelo parâmetro AH) é somada às fontes de vozeamento e tratada pelo mesmo conjunto simulador do trato vocal; já a fonte de *ruído de fricação* (controlada pelo parâmetro AF) requer uma filtragem separada, uma vez que os sons fricativos são gerados em outro ponto do aparelho fonador.

A característica de filtragem do trato vocal é reproduzida por meio de um filtro digital cuja função de transferência é variável no tempo. As ressonâncias do trato vocal (formantes) são associadas a células de segunda ordem (ressoadores), sendo caracterizadas pelas suas freqüências centrais (parâmetros F1 a F6) e larguras de banda (parâmetros B1 a B6). O modelo de Klatt permite associar os ressoadores tanto em cascata como em paralelo; no segundo caso, as amplitudes dos ressoadores são controladas pelos parâmetros A2-A6 e AB.

A característica de irradiação das ondas sonoras a partir da boca do falante é aproximada pela diferenciação do sinal gerado, enfatizando as altas freqüências.

A figura 4.1 ilustra a estrutura geral do sintetizador de formantes de Klatt.

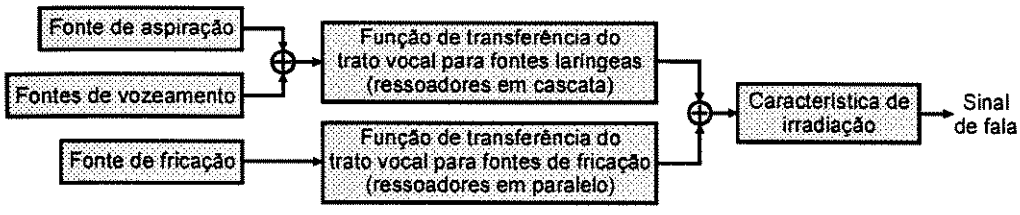

Figura 4.1 Estrutura geral do sintetizador de formantes de Klatt (configuração cascata/paralelo), mostrando as fontes, as características de filtragem do trato vocal e a característica de irradiação.

O sintetizador de Klatt recebe como entrada uma tabela na qual são especificados, a intervalos periódicos, os valores dos parâmetros de controle, e gera em sua saída as amostras do sinal de fala com taxa de amostragem determinada pelo parâmetro SR. O número de amostras gerado para cada conjunto de parâmetros de controle recebido pelo sintetizador é determinado pelo parâmetro NWS. O intervalo de tempo entre conjuntos sucessivos de parâmetros é dado, portanto, pela relação NWS / SR; valores típicos são SR = 16000 amostras/s e NWS = 96 amostras, levando a um intervalo de 6 ms entre conjuntos sucessivos de parâmetros de controle.

Além das características presentes no sintetizador original de Klatt, foram adicionados dois parâmetros denominados *jitter* (JI) e *shimmer* (SH). Esses parâmetros têm por função introduzir pequenas variações aleatórias na freqüência fundamental (*jitter*) e nas amplitu-

des das fontes de vozeamento (*shimmer*), aumentando a naturalidade da fala sintetizada. Foi ainda realizada uma modificação em relação ao algoritmo original do sintetizador: ao receber um conjunto de parâmetros de controle, o sintetizador gerava NWS amostras do sinal de fala mantendo constantes os coeficientes de seus filtros. Ao receber um novo conjunto de parâmetros, os coeficientes dos filtros eram alterados abruptamente e novamente mantidos constantes até concluir-se a geração de NWS amostras. A variação abrupta nesses coeficientes era responsável pela introdução de distorções no sinal de fala sintetizado. O problema foi contornado suavizando-se a variação dos coeficientes dos filtros através de interpolação linear, melhorando perceptivelmente a qualidade da fala produzida.

A tabela a seguir lista todos os parâmetros de controle do sintetizador de Klatt, incluindo os dois parâmetros acrescentados ao modelo original.

Parâmetro	Descrição
AV	Amplitude de vozeamento normal (dB)
AVS	Amplitude de vozeamento quase senoidal (dB)
AF	Amplitude de ruído de fricação (dB)
AH	Amplitude de ruído de aspiração (dB)
F0	Freqüência fundamental de vozeamento (Hz)
F1	Freqüência do primeiro formante (Hz)
F2	Freqüência do segundo formante (Hz)
F3	Freqüência do terceiro formante (Hz)
F4	Freqüência do quarto formante (Hz)
F5	Freqüência do quinto formante (Hz)
F6	Freqüência do sexto formante (Hz)
B1	Largura de banda do primeiro formante (Hz)
B2	Largura de banda do segundo formante (Hz)
B3	Largura de banda do terceiro formante (Hz)
B4	Largura de banda do quarto formante (Hz)
B5	Largura de banda do quinto formante (Hz)
B6	Largura de banda do sexto formante (Hz)
AN	Amplitude do formante nasal (dB)
A1	Amplitude do primeiro formante (dB)
A2	Amplitude do segundo formante (dB)
A3	Amplitude do terceiro formante (dB)
A4	Amplitude do quarto formante (dB)
A5	Amplitude do quinto formante (dB)
A6	Amplitude do sexto formante (dB)
AB	Amplitude do trajeto de *bypass* (dB)
FGP	Freqüência do ressoador glotal 1 (Hz)
BGP	Largura de banda do ressoador glotal 1 (Hz)
BGS	Largura de banda do ressoador glotal 2 (Hz)
FNZ	Freqüência do zero nasal (Hz)
BNZ	Largura de banda do zero nasal (Hz)
FNP	Freqüência do pólo nasal (Hz)
BNP	Largura de banda do pólo nasal (Hz)
FGZ	Freqüência do zero glotal (Hz)
BGZ	Largura de banda do zero glotal (Hz)
NFC	Número de formantes em cascata
G0	Controle geral de ganho (dB)
SR	Taxa de amostragem (Hz)
NWS	Número de amostras geradas por conjunto de parâmetros
SW	Seleção da configuração em cascata (0) ou paralela (1)
JI	Limite para *jitter*, em partes por milhar
SH	Limite para *shimmer*, em partes por milhar

Dos 41 parâmetros listados na tabela, os 20 parâmetros seguintes nunca são variados durante a síntese[1] : SR, NWS, NFC, AN, A1, F5, F6, B4, B5, B6, FGP, FGZ, FNP, BGP, BGZ, BNP, BNZ, BGS, G0 e SW.

4.3 MODELOS PARA AS CLASSES DE FONES

A primeira etapa no desenvolvimento de regras para a obtenção dos valores dos parâmetros de controle do sintetizador a partir de uma transcrição fonética consiste na definição de modelos que descrevam o comportamento desses parâmetros ao longo do tempo para cada classe de fones. Os modelos aqui apresentados foram criados por Edson José Nagle em sua pesquisa de doutorado [Nagle 1991]. Os contornos de parâmetros são traçados utilizando segmentos de reta. São empregadas três classes básicas de modelos, descritas nos próximos itens.

Modelo genérico

No modelo descrito neste item (denominado *genérico*), que se aplica à maioria dos fones sintetizados, os contornos dos parâmetros apresentam uma única região de valor constante, cercada de um lado pela transição proveniente do fone anterior e do outro pela transição que vai ao fone seguinte. A figura 4.2 ilustra esse modelo.

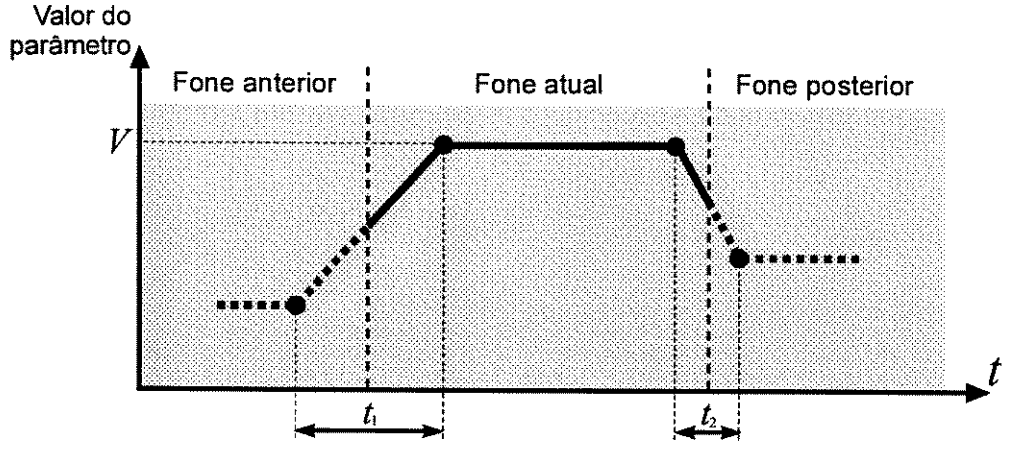

Figura 4.2 Modelo genérico para o comportamento dos parâmetros de controle do sintetizador. São mostradas as regiões de transição, com durações t_1 e t_2, e a região de valor constante V.

[1] Em princípio, alguns desses parâmetros considerados fixos (como F5, F6, B4, B5 e B6) poderiam ser variados durante a síntese para a reprodução na fala sintética de detalhes observados na fala natural; no entanto, dado o nível atual de sofisticação das regras de síntese, tais parâmetros são mantidos sempre constantes.

O valor constante V de cada parâmetro é uma característica inerente ao fone sintetizado, mas pode sofrer variações em função do contexto em que o fone se insere (isto é, em função de seus fones vizinhos). Além disso, os tempos de transição t_1 e t_2 são características de cada parâmetro e de cada par de fones (ou classes de fones). As transições são realizadas por meio de segmentos de reta que ligam duas regiões de valor constante, sempre de modo a centrar-se na fronteira entre os fones. A duração total de um fone é dada, portanto, pela soma da duração da região de valor constante e de metade das durações das transições para qualquer parâmetro que siga este modelo.

Quando a duração total de um fone é alterada, são mantidas as inclinações das retas de transição; desse modo, toda a alteração na duração de um fone se reflete nas regiões de valor constante dos parâmetros. Caso a duração total do fone seja reduzida a ponto de se tornar menor do que a soma da metade das durações das transições, a região de valor constante desaparece e ocorre um encontro entre as transições; neste caso, o valor V nunca é atingido. A figura 4.3 ilustra o comportamento da região de valor constante e das transições conforme a duração total de um fone é reduzida.

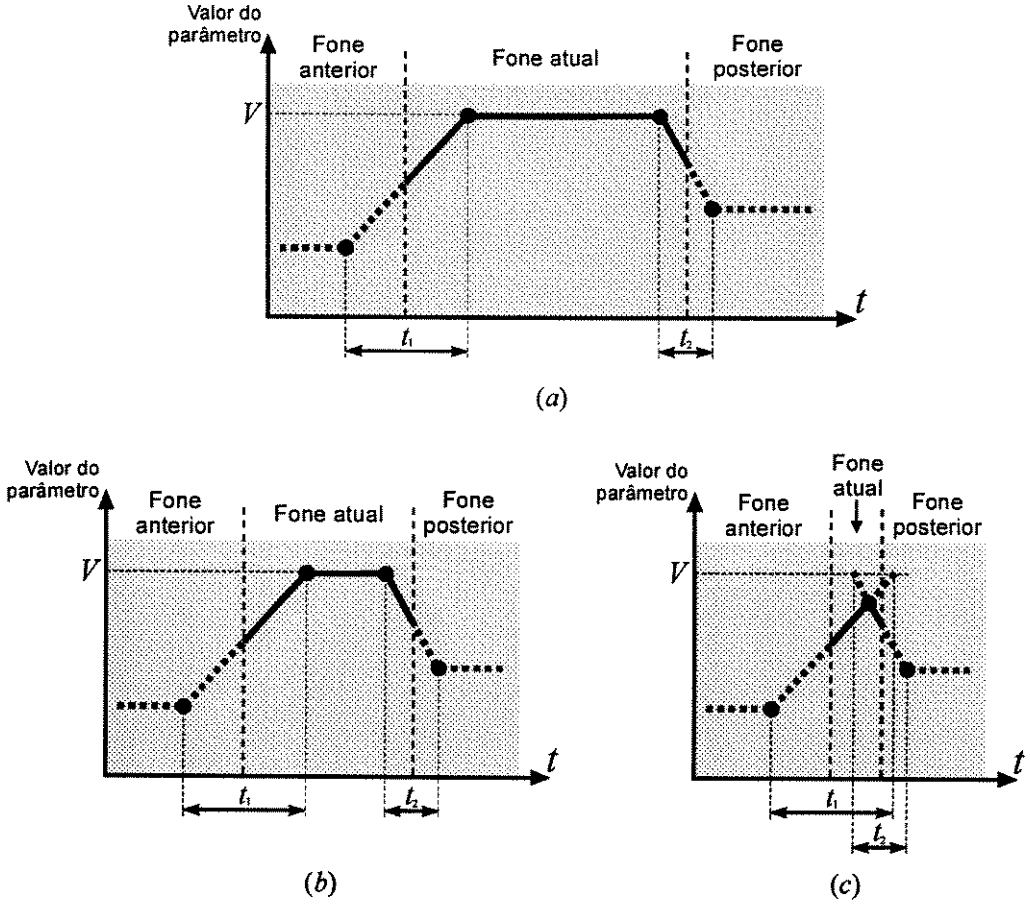

Figura 4.3 Efeito da redução na duração total de um fone sobre a região de valor constante e as transições para o modelo genérico: (a) duração original; (b) duração reduzida sem eliminação da região de valor constante; (c) duração reduzida com eliminação da região de valor constante.

O modelo genérico é utilizado na geração dos contornos dos parâmetros para a maioria dos fones sintetizados. A exceção fica por conta dos sons oclusivos, nos quais as amplitudes das fontes sonoras seguem modelos mais elaborados descritos no próximo item. Além disso, ao fone /f/ atribuiu-se um comportamento levemente alterado em relação às demais fricativas no que se refere ao parâmetro AF; para este parâmetro, a região entre transições apresenta uma inclinação positiva, como mostrado na figura 4.4.

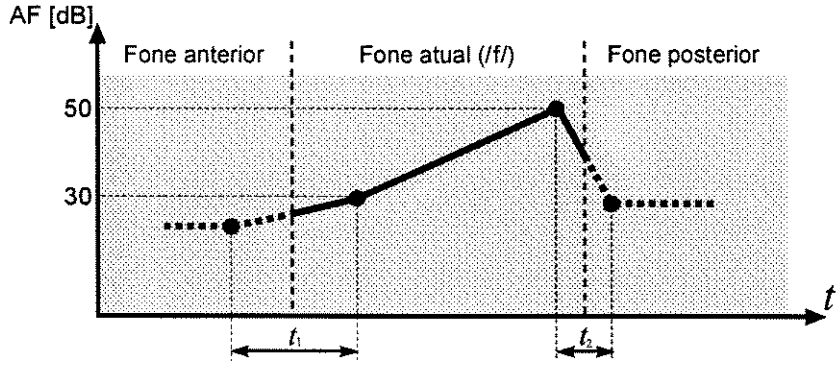

Figura 4.4 Contorno do parâmetro AF para o fone /f/.

Modelos para oclusivas

Os parâmetros correspondentes às amplitudes das fontes sonoras nos sons oclusivos seguem modelos diferentes do apresentado no item anterior, devido à presença de uma região de explosão na qual as fontes têm suas amplitudes bruscamente elevadas. Já para as freqüências e larguras de banda dos formantes, aplica-se o modelo genérico.

A figura 4.5 ilustra o modelo adotado para a amplitude das fontes de vozeamento (AV e AVS) e de ruído de aspiração (AH) nas oclusivas surdas (/p/, /t/ e /k/), para as quais essas fontes permanecem anuladas até o momento da explosão, quando sobem bruscamente. O modelo correspondente para as oclusivas sonoras (/b/, /d/ e /g/) encontra-se representado na figura 4.6; neste caso, as fontes já apresentam amplitude não nula antes da explosão, subindo gradativamente no momento em que esta ocorre.

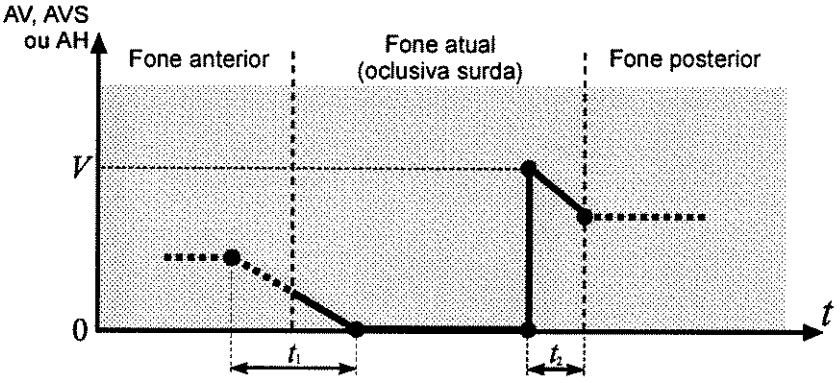

Figura 4.5 Modelo para o comportamento das fontes de vozeamento e de ruído de aspiração nas oclusivas surdas, mostrando as regiões de transição, de silêncio e de explosão.

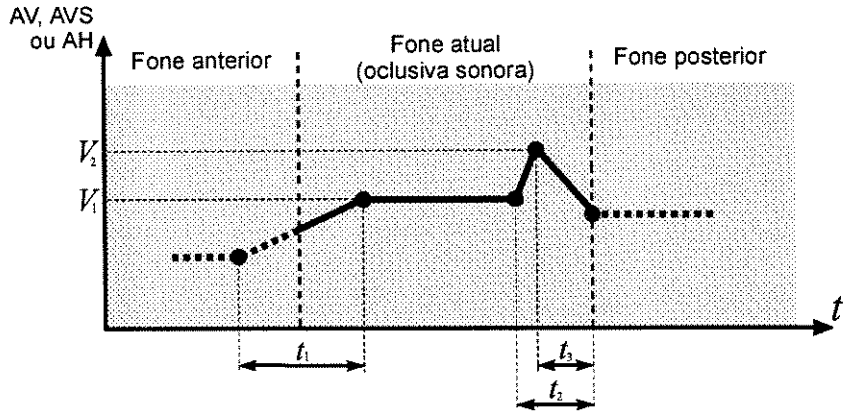

Figura 4.6 Modelo para o comportamento das fontes de vozeamento e de ruído de aspiração nas oclusivas sonoras, mostrando as regiões de transição, de estabilidade e de explosão.

Na figura 4.7 encontra-se representado o modelo para a amplitude da fonte de ruído de fricação (AF) nas oclusivas. Essa fonte permanece anulada até o momento da explosão, quando sua amplitude sobe gradativamente até um valor máximo, permanece nesse valor durante um certo tempo e volta gradativamente a zero.

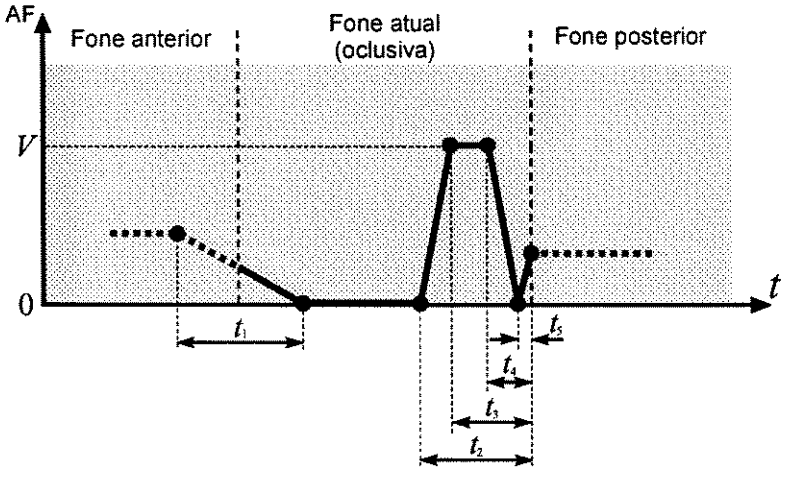

Figura 4.7 Modelo para o comportamento da fonte de ruído de fricação nas oclusivas, mostrando as regiões de transição, de silêncio e de explosão.

No que se refere às amplitudes das fontes, alterações na duração total de uma oclusiva refletem-se na região de valor constante que precede a explosão, mantendo-se fixas a transição inicial e a explosão propriamente dita. As durações não devem ser reduzidas a ponto de eliminar totalmente a região de valor constante, pois isso descaracterizaria os sons oclusivos.

Modelo de interligação

Uma possibilidade adicional de construção do contorno de um parâmetro corresponde à simples interligação, por meio de um segmento de reta, do último ponto do contorno no fone anterior até o primeiro ponto do contorno no fone seguinte, conforme ilustrado na figura 4.8a. Caso haja uma sucessão de fones que seguem este modelo, o segmento de reta se estende por todos eles, como mostrado na figura 4.8b.

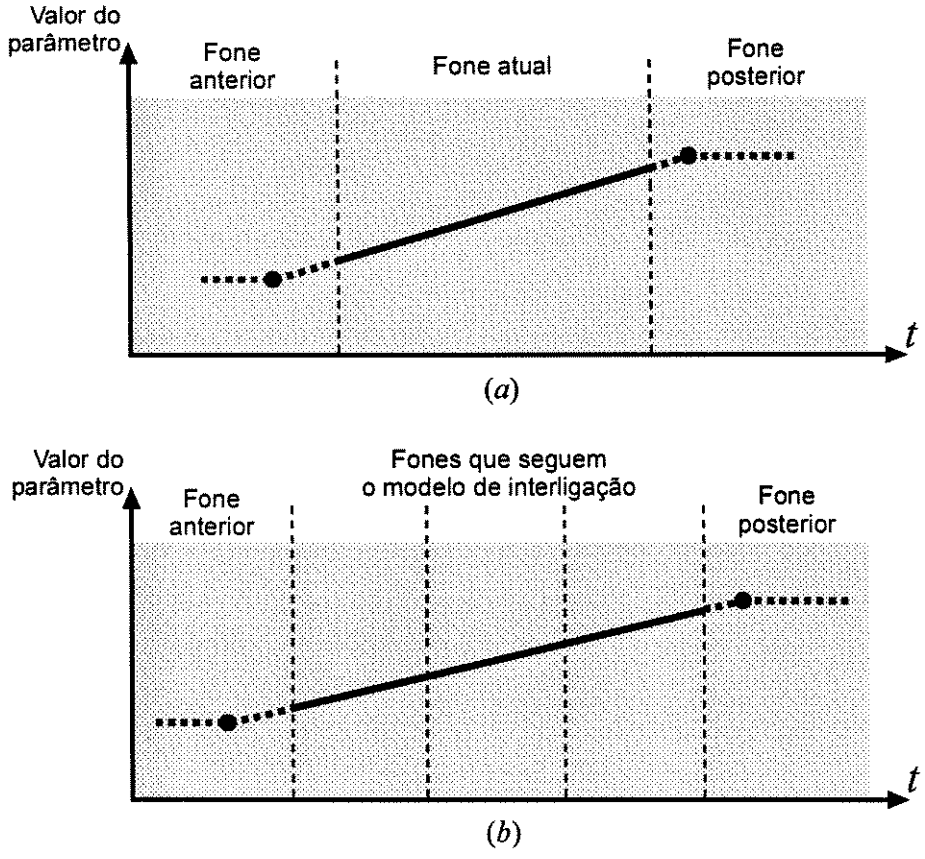

Figura 4.8 Modelo de interligação: (a) para um fone; (b) para vários fones consecutivos.

Como exemplo de aplicação deste modelo, podem-se citar as freqüências dos formantes no fone /R/, fortemente dependentes dos contextos anterior e posterior. O modelo se aplica também nas situações em que os valores assumidos por determinados parâmetros são irrelevantes; durante um período de silêncio, por exemplo, as freqüências dos formantes não afetam a fala sintetizada (pois as amplitudes das fontes estão zeradas), sendo utilizado o modelo de interligação.

Este modelo não pode ser aplicado diretamente quando o fone se localiza no início da frase, situação na qual não há um fone anterior, ou no final da frase, quando não há um fone posterior. A solução adotada nesses casos consiste em manter-se o valor do parâmetro

constante ao longo do fone, igualando-o ao valor do ponto mais próximo do fone adjacente. A figura 4.9 ilustra os modelos resultantes.

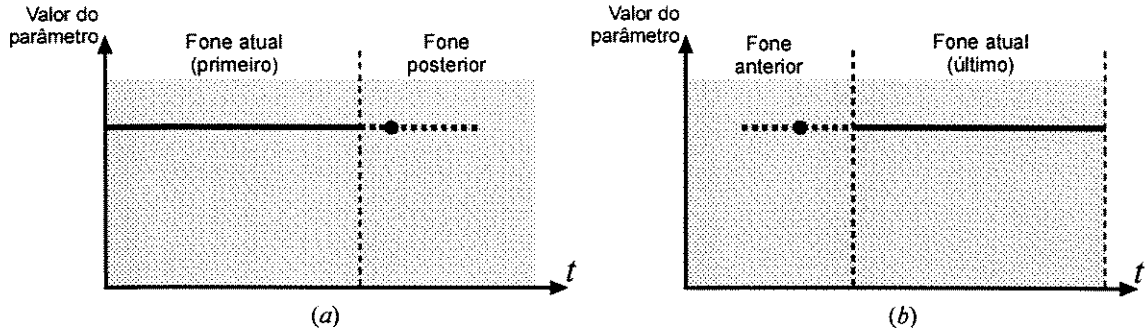

Figura 4.9 Casos especiais do modelo de interligação: (a) quando o fone atual é o primeiro da frase; (b) quando o fone atual é o último da frase.

4.4 LINGUAGEM PARA DESCRIÇÃO DE REGRAS

Para a implementação das regras de síntese utilizando os modelos apresentados na seção anterior, foi criada uma linguagem específica, denominada *Linguagem para Descrição de Regras* (LDR), e um compilador associado (descrito na próxima seção)[2]. A disponibilidade de uma linguagem dedicada torna o processo de implementação das regras muito mais simples do que o seria caso fosse usada uma linguagem de programação genérica (como C, por exemplo), já que os recursos disponíveis foram projetados para satisfazer às necessidades específicas do problema. Além disso, esta abordagem evita a recompilação do sistema a cada alteração nas regras.

Os próximos itens descrevem o processo através do qual são traçados os contornos dos parâmetros de controle do sintetizador para cada fone e explicam em detalhe os comandos da linguagem LDR.

O processo de construção dos contornos dos parâmetros

Os contornos dos parâmetros de controle do sintetizador são definidos através de um conjunto de pontos interligados por segmentos de reta, utilizando interpolação linear. Um programa LDR atua sobre esse conjunto, movendo e acrescentando pontos de modo a obter o contorno desejado. Antes da execução de um programa LDR, os contornos dos parâmetros são compostos por dois pontos localizados nas fronteiras do fone e com valores idênti-

[2] Uma descrição resumida do módulo conversor, da linguagem LDR e do compilador de regras encontra-se em [Gomes et al. 1996].

cos, conforme ilustrado na figura 4.10. Esse valor inicial é obtido de uma tabela armazenada no *arquivo de dados para síntese* ("DADOS.TAB"), descrito mais adiante.

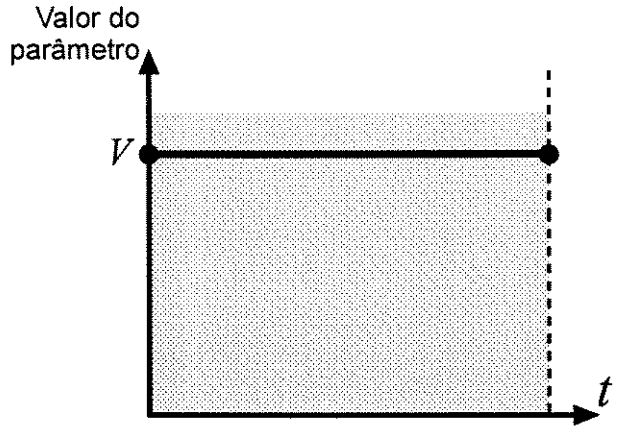

Figura 4.10 Contorno horizontal assumido antes da execução de um programa LDR.

A figura 4.11 ilustra algumas etapas possíveis para a construção de um contorno complexo partindo-se do contorno inicial mostrado na figura 4.10. Cada uma destas etapas corresponde a um único comando LDR.

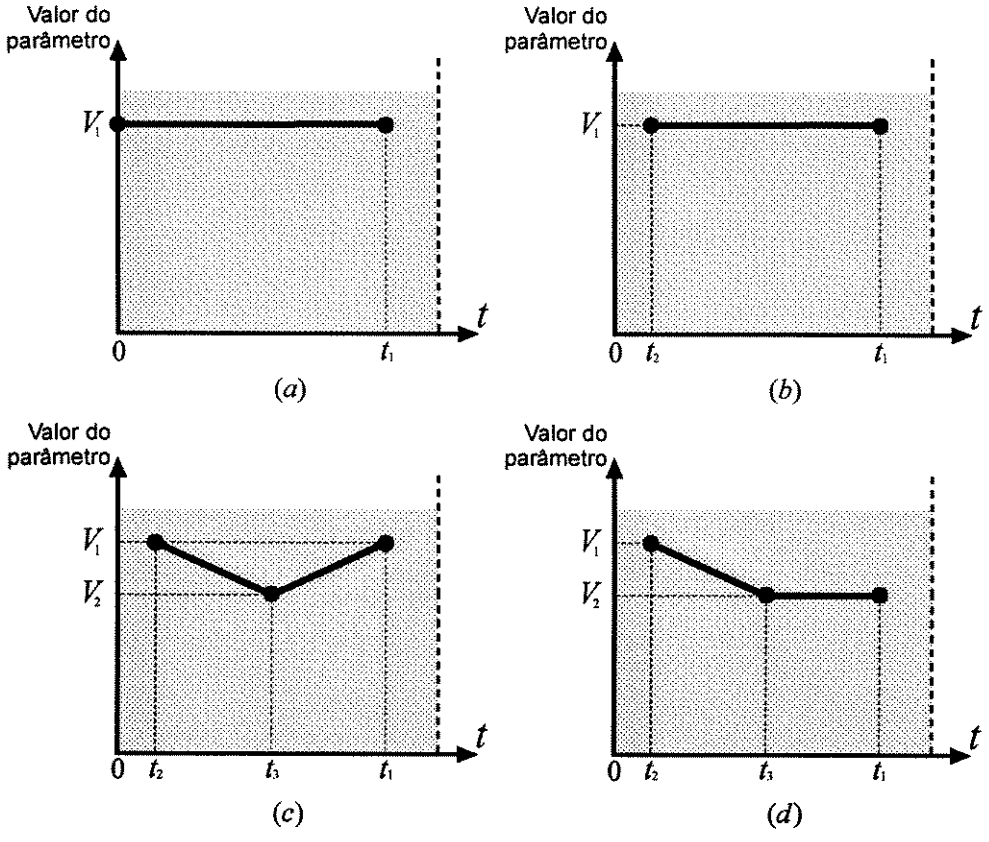

Figura 4.11 Algumas possibilidades de manipulação do contorno inicial mostrado na figura 4.10 utilizando comandos LDR: (a) alteração do tempo associado ao segundo ponto; (b) alteração do tempo associado ao primeiro ponto; (c) acréscimo de um ponto intermediário; (d) alteração do valor associado ao último ponto.

Na região de fronteira entre fones adjacentes, a transição no contorno de um parâmetro é efetuada por meio de um segmento de reta que liga o último ponto do primeiro fone ao ponto inicial do segundo fone. Transições geradas por este processo, em geral suaves, podem se tornar abruptas caso os pontos envolvidos estejam muito próximos da região de fronteira, ou mesmo sobre ela. A figura 4.12 ilustra o mecanismo de geração de transições entre fones.

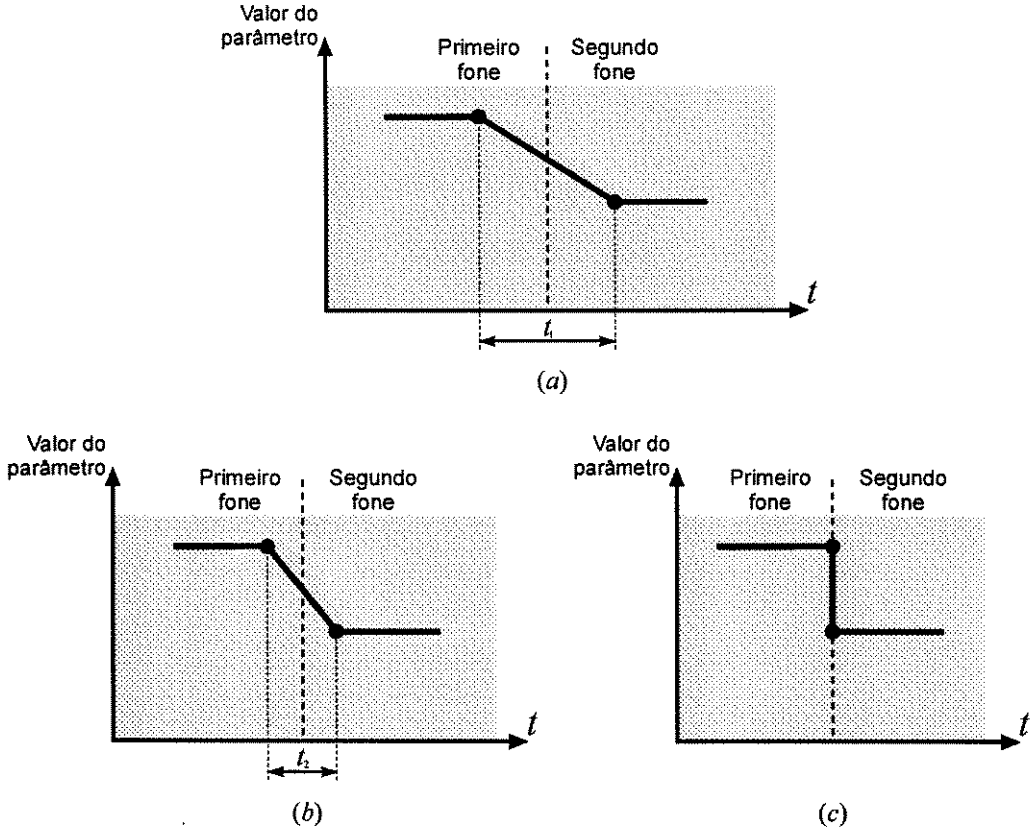

Figura 4.12 Possibilidades de transição entre fones: (a) transição suave (pontos afastados da fronteira entre os fones); (b) transição rápida (pontos próximos da fronteira); (c) transição abrupta (pontos sobre a fronteira).

Palavras reservadas da linguagem LDR

Neste item estão descritas as palavras reservadas da linguagem LDR. São definidas 11 palavras reservadas: ACP, CASO, DUIP, FIM, IGUALE, PARA, SE, SEMOD, SOME, SUB e VA. Cada descrição é acompanhada por um exemplo de utilização.

Os comandos LDR que agem diretamente sobre os pontos dos contornos permitem a especificação de um único parâmetro ou de grupos de parâmetros pré-definidos. Esses grupos são:

- **FONS:** AV, AH e AVS.
- **FONT:** AV, AF, AH e AVS.
- **Fn:** F1, F2, F3 e F4.
- **Bn:** B1, B2 e B3.
- **FBn:** F1, F2, F3, F4, B1, B2 e B3.
- **An:** A2, A3, A4, A5, A6 e AB.
- **FBAn:** F1, F2, F3, F4, B1, B2, B3, A2, A3, A4, A5, A6 e AB.
- **TODOS:** todos os parâmetros que podem ser variados.

Além disso, os comandos IGUALE, SOME e SUB aceitam a especificação de um parâmetro adicional, denominado DUR, que permite às regras de síntese alterar a duração do fone.

Os pontos que definem o contorno de um parâmetro no interior de um fone são numerados em ordem crescente a partir de zero; caso esse trecho do contorno contenha três pontos, por exemplo, a eles são associados os números 0, 1 e 2. Os comandos LDR que atuam diretamente sobre pontos de contornos de parâmetros utilizam esse número como identificador do ponto no interior de um fone. Além disso, esses comandos em geral permitem a atuação tanto sobre o tempo como sobre o valor associados a um ponto; nesse caso, precede-se o número correspondente ao ponto pela letra "T", caso se deseje atuar sobre o tempo, ou pela letra "V", caso se deseje atuar sobre o valor. Por exemplo, o tempo associado ao terceiro ponto de um contorno no interior de um fone é especificado por "T2" (pois a contagem dos pontos se inicia em zero), enquanto o valor associado a esse mesmo ponto é especificado por "V2".

A linguagem LDR permite a definição de grupos de fones; por exemplo, pode ser definido um grupo "VOGAL" que inclui todos os sons vocálicos. Os comandos da linguagem que requerem a especificação de um fone aceitam também a especificação de um grupo de fones. A sintaxe para a definição de grupos de fones está descrita no próximo item desta seção.

Outro recurso disponível na linguagem LDR são as *modificações*. Uma modificação corresponde a um *flag* que pode estar ativo ou inativo durante o processamento de um fone. O comando SEMOD, explicado em detalhe mais adiante, permite a execução condicional de um bloco de comandos com base no estado de uma modificação. A cada modificação é associado um nome que a identifica no programa LDR. Duas modificações são automaticamente definidas pelo sistema: "SOM-INICIAL", que está ativa durante o processamento do primeiro fone de uma frase e inativa nos demais, e "SOM-FINAL", que está ativa du-

rante o processamento do último fone de uma frase e inativa nos demais. Outras modificações podem ser definidas, devendo-se atribuir um nome e um *caractere modificador* a cada uma delas. Uma modificação é ativada durante o processamento de um fone quando este é precedido na transcrição fonética de entrada pelo caractere modificador correspondente. Na versão atual das regras de síntese, este recurso é utilizado na definição de símbolos para a identificação de sons tônicos, permitindo ao sistema tratá-los adequadamente. A sintaxe para a definição de modificações está descrita no próximo item desta seção.

Os comandos que compõem um programa LDR são normalmente executados de forma seqüencial. A ordem de execução pode ser alterada por meio do comando VA, que efetua um desvio para um ponto arbitrário do programa identificado por um rótulo. Os rótulos são constituídos por palavras seguidas de dois pontos; por exemplo, "AQUI:" é um rótulo válido. A sintaxe do comando VA encontra-se descrita em detalhe mais adiante.

Em geral, os comandos da linguagem LDR que requerem a especificação de um valor para um ponto do contorno de um parâmetro aceitam, além de valores numéricos, os símbolos "/", "<" e ">". Esses símbolos somente podem ser associados ao primeiro ponto do contorno no interior de um fone (ou seja, o ponto "0"). O símbolo "/" indica que o contorno do parâmetro neste fone segue o modelo de interligação descrito na seção 4.3 e ilustrado na figura 4.8. O símbolo "<" indica que o valor do parâmetro deve permanecer constante ao longo do fone atual, mantendo o último valor observado no fone anterior. O símbolo ">", por sua vez, indica que o valor do parâmetro deve permanecer constante ao longo do fone atual e igual ao valor observado no fone seguinte. Esses dois últimos símbolos permitem implementar o modelo de interligação modificado descrito na seção 4.3 e ilustrado na figura 4.9.

A seguir encontram-se as descrições das palavras reservadas da linguagem LDR, em ordem alfabética. Quando presentes, são descritos também os respectivos argumentos. Para referir-se ao fone cujos contornos de parâmetros estão sendo traçados, é utilizada a expressão "fone atual"; as expressões "fone anterior" e "fone posterior", por sua vez, referem-se aos fones que precedem e que sucedem o fone atual na transcrição fonética, respectivamente.

ACP *<parâmetro> <ponto do contorno> <tempo> <valor>*

Descrição: Este comando insere um novo ponto na posição indicada do contorno, utilizando o tempo e valor especificados.

Argumentos:

- *parâmetro:* parâmetro ou grupo de parâmetros sobre o qual deve atuar o comando.
- *ponto do contorno:* número correspondente ao ponto que deve ser inserido no contorno.
- *tempo:* tempo associado ao ponto que deve ser inserido no contorno, em milissegundos.
- *valor:* valor associado ao ponto que deve ser inserido no contorno, em hertz para freqüências e larguras de banda, em decibéis para amplitudes ou em partes por milhar para *jitter* e *shimmer*.

Exemplo: Dado o contorno mostrado na figura 4.13a para o parâmetro AVS, a execução do comando *ACP FONS 1 50 30* leva ao contorno mostrado na figura 4.13b.

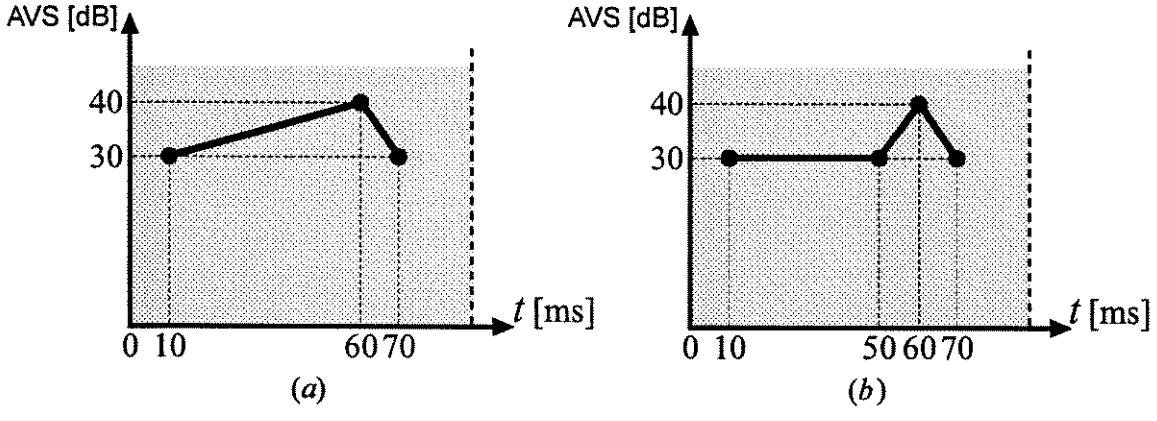

Figura 4.13 Exemplo de utilização do comando ACP: (a) contorno original para o parâmetro AVS; (b) contorno após a execução do comando *ACP FONS 1 50 30*.

CASO *<posição do fone>*

Descrição: Este comando permite a execução condicional de um dentre vários blocos de comandos com base no contexto em que se insere o fone atual. Para cada bloco é especificado, através da palavra reservada *PARA*, um fone ou grupo de fones que será comparado com o fone indicado pelo argumento *posição do fone,* executando-se o primeiro bloco que apresentar uma comparação bem sucedida.

(Para maiores detalhes, ver a descrição da palavra reservada PARA*)*

Argumento:

- *posição do fone:* número que especifica a posição do fone a comparar, relativa ao fone atual, na transcrição fonética de entrada; por exemplo, "0" corresponde ao fone atual, "1" corresponde ao fone posterior, "–1" corresponde ao fone anterior, "2" corresponde ao fone localizado duas posições à frente do atual, e assim por diante.

Exemplo: Dados a transcrição fonética de entrada /AmosAsaiwapresadA/, na qual o fone atual está em itálico, e o seguinte trecho de um programa LDR:

```
CASO -1
 PARA VOGAL:
  SOME AF V0 10
  FIM
 PARA FRICATIVA:
  SUB AF V0 10
  FIM
 PARA p:
  IGUALE AF V0 40
  FIM
```

o primeiro bloco de comandos, definido entre *PARA VOGAL:* e *FIM,* será executado caso o fone anterior pertença ao grupo de fones denominado "VOGAL"; o segundo bloco, definido entre *PARA FRICATIVA:* e *FIM,* será executado caso o fone anterior pertença ao grupo de fones denominado "FRICATIVA"; e o terceiro bloco, definido entre *PARA p:* e *FIM,* será executado caso o fone anterior seja /p/. Como o fone anterior é /a/, uma vogal, o primeiro bloco de comandos deve ser executado.

DUIP *<parâmetro> <tempo>*

Descrição: Este comando subtrai uma quantidade especificada do tempo associado ao último ponto do contorno no fone atual e soma essa mesma quantidade ao tempo associado ao primeiro ponto do contorno no fone posterior.

Argumentos:

- *parâmetro:* parâmetro ou grupo de parâmetros sobre o qual deve atuar o comando.
- *tempo:* quantidade que deve ser subtraída do tempo associado ao último ponto do contorno no fone atual e somada ao tempo associado ao primeiro ponto do contorno no fone posterior, expressa em milissegundos.

Exemplo: Dado o contorno mostrado na figura 4.14a para o parâmetro F2, correspondente a dois fones consecutivos, a execução do comando *DUIP Fn 20* leva ao contorno mostrado na figura 4.14b.

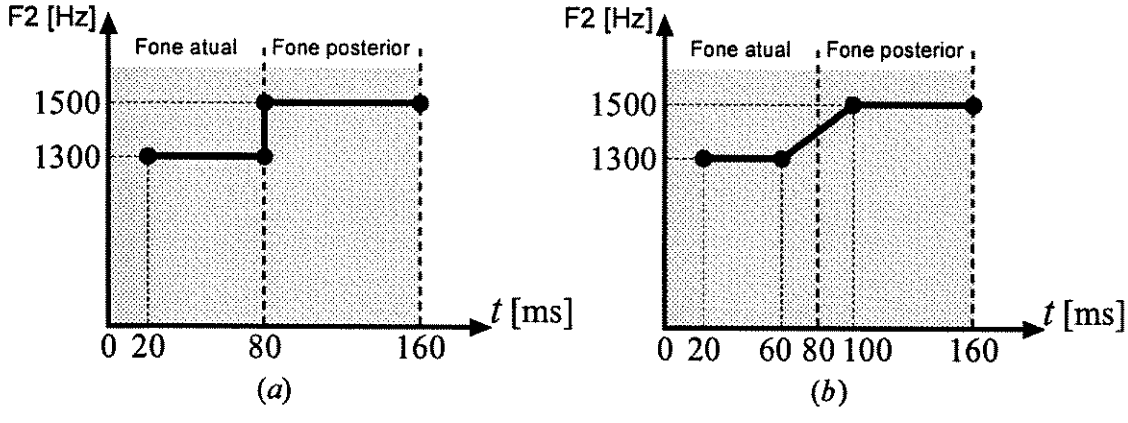

Figura 4.14 Exemplo de utilização do comando DUIP: (a) contornos originais para o parâmetro F2; (b) contornos após a execução do comando *DUIP Fn 20* para o fone atual.

FIM

Descrição: A palavra reservada *FIM* é utilizada para finalizar blocos de comandos, associando-se às palavras reservadas *SE*, *SEMOD* ou *PARA*. Além disso, ela é usada para indicar o final de seções do programa LDR.

Exemplo: No seguinte trecho de um programa LDR:

```
SE 2 a
 SOME DUR 10
 SUB AF T1 10
 FIM
```

a palavra reservada *FIM* indica o final do bloco de comandos cuja execução é condicionada pelo comando *SE 2 a.*

IGUALE *<parâmetro> <ponto do contorno> <tempo ou valor>*

Descrição: Este comando iguala a uma quantidade especificada o tempo ou o valor associados a um ponto do contorno. Se aplicado ao parâmetro DUR, o comando iguala a duração do fone atual à quantidade especificada, não devendo ser fornecido o segundo argumento.

Argumentos:

- *parâmetro:* parâmetro ou grupo de parâmetros sobre o qual deve atuar o comando.
- *ponto do contorno:* número correspondente ao ponto do contorno do parâmetro sobre o qual deve atuar o comando, precedido por "T" ou "V" conforme deva ser afetado o tempo ou valor associados ao ponto; este argumento não deve ser fornecido quando o argumento anterior for DUR.
- *tempo ou valor:* quantidade à qual deve ser igualado o tempo (em milissegundos) ou o valor (em hertz para freqüências e larguras de banda, em decibéis para amplitudes ou em partes por milhar para *jitter* e *shimmer*) associados ao ponto do contorno; caso o primeiro argumento seja DUR, a duração do fone atual (em milissegundos) é igualada à quantidade especificada.

Exemplo: Dado o contorno mostrado na figura 4.15a para o parâmetro AF, a execução do comando *IGUALE AF T2 80* leva ao contorno mostrado na figura 4.15b.

PARA <*fone*>:

Descrição: A palavra reservada *PARA,* utilizada sempre em conjunto com o comando *CASO,* permite a especificação de um fone ou um grupo de fones associado a um bloco de comandos; o bloco somente será executado se houver coincidência entre esse fone ou grupo de fones e o fone contido na transcrição fonética na posição especificada através do comando *CASO.*

(Para maiores detalhes, ver a descrição do comando CASO*)*

Argumento:

- *fone:* fone ou grupo de fones associado ao bloco de comandos especificado entre as palavras reservadas *PARA* e *FIM;* deve ser seguido pelo símbolo de dois pontos (":").

Exemplo: Ver exemplo referente ao comando *CASO.*

SE <*posição do fone*> <*fone*>

Descrição: Este comando permite a execução condicional de um bloco de comandos com base no contexto em que se insere o fone atual; o bloco somente é executado caso haja coincidência entre o fone indicado pelo argumento *posição do fone* e o fone ou grupo de fones fornecido através do argumento *fone.*

Argumentos:

- *posição do fone:* número que especifica a posição do fone a comparar, relativa ao fone atual, na transcrição fonética; por exemplo, "0" corresponde ao fone atual, "1" corresponde ao fone posterior, "–1" corresponde ao fone anterior, "2" corresponde ao fone localizado duas posições à frente do atual, e assim por diante.
- *fone:* fone ou grupo de fones com o qual deve ser comparado o fone indicado pelo argumento anterior.

Exemplo: Dada a transcrição fonética de entrada /AmosAsa*i*wapresadA/, na qual o fone atual está em itálico, e o seguinte trecho de um programa LDR:

```
SE 1 VOGAL
 SOME AV V1 10
 SOME AF V1 5
 FIM
```

o bloco de comandos definido entre o comando *SE* e a palavra reservada *FIM* somente será executado caso o fone seguinte ao atual (/w/) pertença ao grupo de fones denominado "VOGAL".

SEMOD *<modificação>*

Descrição: Este comando permite a verificação do estado de uma modificação; o bloco de comandos que o segue somente é executado caso a modificação esteja ativa.

Argumento:

- *modificação:* nome associado à modificação cujo estado deve ser verificado.

Exemplo: Dado o seguinte trecho de um programa LDR:

```
SEMOD SOM-INICIAL
 SUB DUR 10
 SOME FONS V1 10
 FIM
```

o bloco de comandos definido entre o comando *SEMOD* e a palavra reservada *FIM* somente será executado caso a modificação "SOM-INICIAL" esteja ativa (isto é, caso o fone atual seja o primeiro fone da frase).

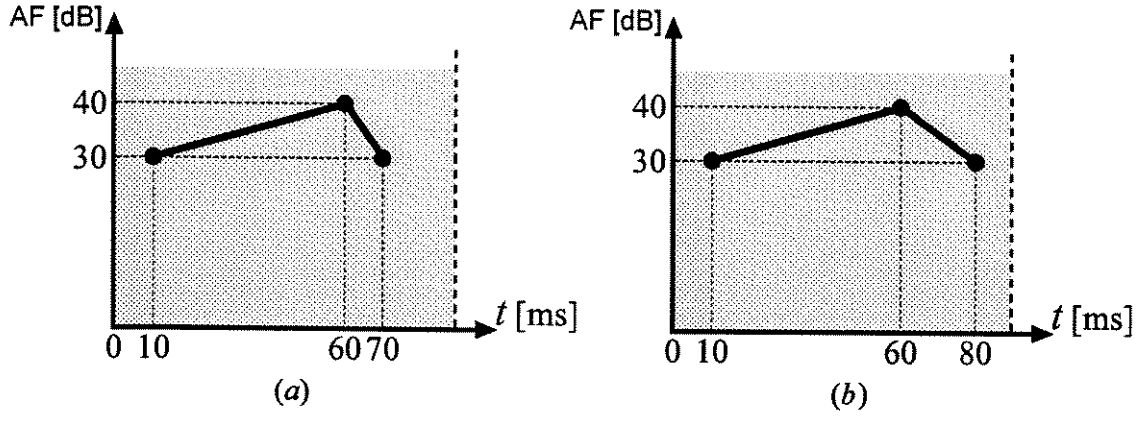

Figura 4.15 Exemplo de utilização do comando IGUALE: (a) contorno original para o parâmetro AF; (b) contorno após a execução do comando *IGUALE AF T2 80*.

SOME *<parâmetro> <ponto do contorno> <tempo ou valor>*

Descrição: Este comando soma uma quantidade especificada ao tempo ou ao valor associados a um ponto do contorno. Se aplicado ao parâmetro DUR, o comando soma a quantidade especificada à duração do fone atual, não devendo ser fornecido o segundo argumento.

Argumentos:

- *parâmetro:* parâmetro ou grupo de parâmetros sobre o qual deve atuar o comando.
- *ponto do contorno:* número correspondente ao ponto do contorno do parâmetro sobre o qual deve atuar o comando, precedido por "T" ou "V" conforme deva ser afetado o tempo ou valor associados ao ponto; este argumento não deve ser fornecido quando o argumento anterior for DUR.
- *tempo ou valor:* quantidade que deve ser somada ao tempo (em milissegundos) ou ao valor (em hertz para freqüências e larguras de banda, em decibéis para amplitudes ou em partes por milhar para *jitter* e *shimmer*) associados ao ponto do contorno; caso o primeiro argumento seja DUR, a quantidade especificada é somada à duração do fone atual (em milissegundos).

Exemplo: Dado o contorno mostrado na figura 4.16a para o parâmetro AV, a execução do comando *SOME AV V0 10* leva ao contorno mostrado na figura 4.16b.

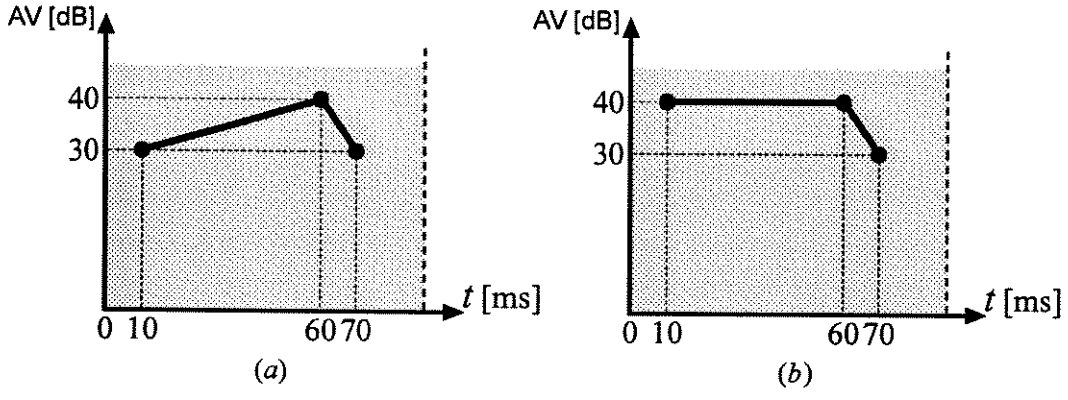

Figura 4.16 Exemplo de utilização do comando SOME: (a) contorno original para o parâmetro AV; (b) contorno após a execução do comando *SOME AV V0 10*

SUB *<parâmetro> <ponto do contorno> <tempo ou valor>*

Descrição: Este comando subtrai uma quantidade especificada do tempo ou do valor associados a um ponto do contorno. Se aplicado ao parâmetro DUR, o comando subtrai a quantidade especificada da duração do fone atual, não devendo ser fornecido o segundo argumento.

Argumentos:

- *parâmetro:* parâmetro ou grupo de parâmetros sobre o qual deve atuar o comando.
- *ponto do contorno:* número correspondente ao ponto do contorno do parâmetro sobre o qual deve atuar o comando, precedido por "T" ou "V" conforme deva ser afetado o tempo ou valor associados ao ponto; este argumento não deve ser fornecido quando o argumento anterior for DUR.
- *tempo ou valor:* quantidade que deve ser subtraída do tempo (em milissegundos) ou do valor (em hertz para freqüências e larguras de banda, em decibéis para amplitudes ou em partes por milhar para *jitter* e *shimmer*) associados ao ponto do contorno; caso o primeiro argumento seja DUR, a quantidade especificada é subtraída da duração do fone atual (em milissegundos).

Exemplo: Dado o contorno mostrado na figura 4.17a para o parâmetro AH, a execução do comando *SUB FONT V1 20* leva ao contorno mostrado na figura 4.17b.

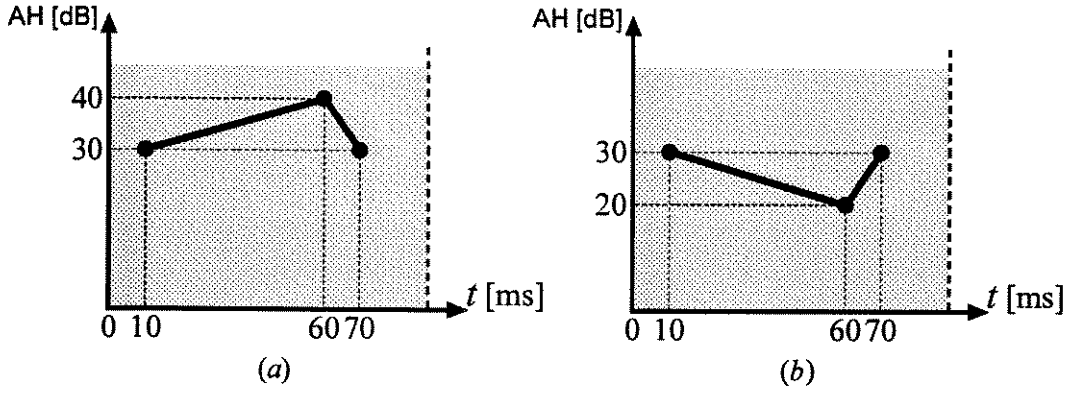

Figura 4.17 Exemplo de utilização do comando SUB: (a) contorno original para o parâmetro AH; (b) contorno após a execução do comando *SUB FONT V1 20*.

VA *<rótulo>*

Descrição: Este comando desvia a execução do programa LDR para o ponto onde se localiza o rótulo especificado.

Argumento:

- *rótulo:* palavra idêntica à que está especificada no ponto para o qual se deseja desviar a execução do programa.

Exemplo: Dado o seguinte trecho de um programa LDR:

```
SOME AF T1 10
VA CONT
SUB AF T1 10
CONT:
IGUALE AF 0 10 40
```

quando é encontrado o comando *VA CONT*, a execução do programa é desviada para o comando que segue o rótulo "CONT" (*IGUALE AF 0 10 40*); neste exemplo, o comando *SUB AF T1 10* nunca é executado.

Estrutura de um programa LDR

Um programa LDR se divide em diversas seções finalizadas pela palavra reservada FIM. A primeira seção corresponde à definição de grupos de fones. Em cada linha é definido um grupo, fornecendo-se primeiramente o seu nome e em seguida os símbolos dos fones a ele pertencentes. Como exemplo, o trecho de programa a seguir define dois grupos de fones:

```
% Definição de grupos de fones
% Grupo            Fones
OCLUSIVA-SURDA     p t k
OCLUSIVA-SONORA    b d g
FIM
```

Neste exemplo, são definidos os grupos "OCLUSIVA-SURDA" (com os fones /p/, /t/ e /k/) e "OCLUSIVA-SONORA" (com os fones /b/, /d/ e /g/). O símbolo "%" indica que o texto seguinte é um comentário e deve ser ignorado.

A segunda seção de um programa LDR corresponde à definição de modificações. Para cada modificação é fornecido o seu nome e o caractere modificador associado. No trecho de programa a seguir, duas modificações são definidas:

```
% Definição de modificações e caracteres modificadores
% Modificação     Caractere
TONICA-LEXICAL        '
TONICA-FRASAL         "
FIM
```

As modificações definidas no exemplo acima denominam-se "TONICA-LEXICAL" e "TONICA-FRASAL", tendo associados os caracteres modificadores "'" (apóstrofo ou aspas simples) e " " " (aspas duplas). Para a transcrição fonética /Av'élAySt'aas'ezA/, por exemplo, a modificação "TONICA-LEXICAL" estará ativa durante o processamento dos fones /é/, /a/ e /e/ precedidos por apóstrofo.

A terceira seção de um programa LDR consiste na definição de rótulos que indicam os pontos do programa nos quais tem início o tratamento de fones ou grupos de fones específicos. O exemplo a seguir ilustra esta seção:

```
% Posição de início de tratamento para fones ou
% grupos de fones

% Fone ou grupo    Rótulo
OCLUSIVA-SURDA     INICIO-OCLUSIVA-SURDA
OCLUSIVA-SONORA    INICIO-OCLUSIVA-SONORA
a                  INICIO-a
p                  INICIO-p
FIM
```

Quatro rótulos são definidos no exemplo acima: "INICIO-OCLUSIVA-SURDA", "INICIO-OCLUSIVA-SONORA", "INICIO-a" e "INICIO-p". Os dois primeiros estão associados aos grupos de fones "OCLUSIVA-SURDA" e "OCLUSIVA-SONORA", respectivamente; os dois últimos, por sua vez, associam-se aos fones individuais /a/ e /p/. Como o fone /p/ pertence ao grupo "OCLUSIVA-SURDA", ambos os rótulos "INICIO-p" e "INICIO-OCLUSIVA-SURDA" estão associados a esse fone; nesta situação, prevalece o rótulo definido por último (no exemplo, "INICIO-p").

A última seção do programa LDR consiste na descrição das regras de síntese propriamente ditas. Esta seção é dividida em segmentos dedicados ao tratamento de fones ou grupos de fones específicos, encabeçando-se cada segmento por um dos rótulos definidos na seção anterior. A sua execução tem início no segmento correspondente ao fone atual. A seguir encontra-se um exemplo de uma estrutura possível para esta seção:

```
% Descrição das regras de síntese

INICIO-OCLUSIVA-SURDA:
SOME ...
⋮
FIM

INICIO-OCLUSIVA-SONORA:
SUB ...
⋮
FIM

INICIO-a:
SOME ...
⋮
FIM

INICIO-p:
SUB ...
⋮
FIM
```

Assumindo-se que o fone atual é /p/, a execução desta seção do programa LDR tem início no comando que sucede o rótulo "INICIO-p". Se o fone atual fosse /t/, pertencente ao grupo "OCLUSIVA-SURDA", a execução teria início imediatamente após o rótulo "INICIO-OCLUSIVA-SURDA". A palavra reservada FIM é normalmente utilizada ao final de cada segmento desta seção e indica que a execução do programa LDR deve ser encerrada.

Como as três primeiras seções do programa LDR independem do fone atual, elas são executadas uma única vez pelo sistema de conversão texto-fala; a última seção, no entanto, depende tanto do fone atual como do contexto em que ele se insere, devendo ser executada tantas vezes quantos forem os fones contidos na transcrição fonética. Este processo é coordenado pelo módulo conversor, descrito na seção 4.5.

O programa LDR que contém a versão atual das regras de síntese encontra-se no apêndice B.

Compilador de regras

Para que possa ser utilizado pelo sistema, um programa LDR no formato texto (nome *default* "REGRAS.LDR") deve ser compilado, gerando uma versão binária (nome *default* "REGRAS.BIN"). Esta tarefa é executada por um compilador específico, denominado *compilador de regras* (arquivo "CPREG.CPP"), cuja estrutura se assemelha à de um mon-

tador (*assembler*). Ao ser compilado, um comando LDR é convertido em um número previamente conhecido de bytes, apresentando uma porção fixa, que identifica o comando em si, e uma porção variável que corresponde aos seus argumentos. Espaços em branco e caracteres de tabulação, mudança de linha e retorno de carro são considerados como separadores de palavras, permitindo a livre formatação do programa fonte LDR. A indentação de blocos de comandos, embora não obrigatória, é aconselhável, pois facilita a posterior leitura e alteração do programa.

O compilador permite a utilização de constantes literais em lugar de números nos comandos LDR. Os valores dessas constantes somente serão conhecidos no momento da execução do programa, estando armazenados em um arquivo-texto facilmente editável denominado *arquivo de dados para síntese* ("DADOS.TAB", descrito no próximo item desta seção). Por exemplo, ao encontrar o comando *SOME AF T0 DESLOC,* o compilador considera a cadeia de caracteres "DESLOC" como uma constante cujo valor somente será conhecido em tempo de execução. Este recurso permite que os modelos para classes de fones sejam especificados em função dessas constantes, que podem ser alteradas sem necessidade de recompilação das regras de síntese. A constante literal "DUR" é automaticamente definida pelo sistema, sendo substituída pela duração do fone atual no momento da execução do programa LDR; o comando *ACP AV 2 DUR 50,* por exemplo, insere um ponto no contorno do parâmetro AV com valor igual a 50 e tempo igual à duração do fone atual.

O processo de compilação consiste em uma montagem de dois passos. No primeiro passo, são compiladas as seções iniciais do programa (definições de grupos de fones, modificações e rótulos associados a fones ou grupos de fones) e é efetuada uma primeira varredura nas regras de síntese, compilando-se totalmente os comandos LDR que não necessitem de endereços ainda não conhecidos. O comando de desvio, VA, não pode ser totalmente compilado, pois os endereços associados aos rótulos não se encontram disponíveis nesta etapa do processo. O mesmo ocorre com os comandos de decisão (SE, SEMOD e CASO), para os quais é necessário efetuar um salto para um endereço ainda não conhecido em caso de falha na comparação. Estes comandos são parcialmente compilados, reservando-se o número necessário de bytes para a posterior introdução dos endereços faltantes. Ainda durante a primeira varredura das regras, é gerada uma tabela contendo os rótulos definidos ao longo do programa e seus endereços correspondentes. Outra tabela armazena os endereços de desvio em caso de falha na comparação em um comando de decisão.

No segundo passo da montagem, faz-se uma nova varredura nas regras de síntese, preenchendo-se os espaços reservados aos endereços faltantes (armazenados nas tabelas geradas durante o primeiro passo). Gera-se então o arquivo que contém o programa LDR compilado (nome *default* "REGRAS.BIN"). Além das seções já presentes no programa fonte, a versão compilada inclui uma lista das constantes literais utilizadas ao longo do

programa; essas constantes são substituídas por seus respectivos valores numéricos (obtidos do arquivo de dados para síntese) pelo módulo conversor (descrito na próxima seção).

O compilador de regras efetua uma série de verificações sintáticas no programa fonte LDR. Caso um erro seja localizado, a compilação é abortada e é emitida uma mensagem indicando a natureza do erro e o número da linha na qual ele ocorreu. Os erros mais comuns referem-se a comandos inexistentes, argumentos incorretos ou insuficientes e rótulos não definidos.

Arquivo de dados para síntese

Como explicado anteriormente nesta seção, antes da execução do programa LDR os contornos dos parâmetros para o fone atual são assumidos horizontais, isto é, compostos por dois pontos de valores idênticos e localizados nos limites do fone. O valor atribuído a estes pontos é obtido de uma tabela armazenada em um arquivo-texto denominado *arquivo de dados para síntese* (nome default "DADOS.TAB"). A primeira seção desse arquivo contém os valores dos parâmetros fixos (que não podem ser variados durante a síntese), idênticos para todos os fones. Esses valores são especificados em uma única linha, separados por espaços ou caracteres de tabulação, na seguinte ordem: F0, SR, NWS, NFC, AN, A1, F5, F6, B4, B5, B6, FGP, FGZ, FNP, BGP, BGZ, BNP, BNZ, BGS, G0 e SW. A segunda seção do arquivo contém a tabela com os valores dos parâmetros variáveis para cada fone; estes são os valores utilizados nos contornos horizontais iniciais. São primeiramente especificados os valores *default* desses parâmetros[3], em uma única linha e na seguinte ordem: AV, AF, AH, AVS, F1, F2, F3, F4, B1, B2, B3, FNZ, A2, A3, A4, A5, A6, AB, JI, SH e DUR. O parâmetro DUR corresponde à duração intrínseca de um fone, e é utilizado quando uma duração não é fornecida juntamente com a transcrição fonética de entrada. Em seguida são especificados os valores dos parâmetros variáveis para os vários fones reconhecidos pelo sistema; a cada fone é dedicada uma linha contendo o seu símbolo seguido pela lista de valores dos parâmetros na mesma ordem utilizada para os valores *default.* Esta seção é finalizada pela palavra "FIM".

Além de valores numéricos, podem ser usados na tabela de parâmetros variáveis os símbolos "/", "<" e ">" (exceto para o parâmetro DUR), com significados idênticos aos descritos para a sua aplicação como argumentos de comandos LDR. Pode ainda ser usado o caractere "#", indicando que o valor de um dado parâmetro é igual ao fornecido na linha

[3] Os valores *default* são utilizados para os parâmetros variáveis que permaneceram indefinidos até o final da síntese. Essa situação ocorre, por exemplo, quando todos os sons sintetizados utilizam o modelo de interligação para um mesmo parâmetro.

anterior. Além disso, caso uma linha chegue ao fim sem que tenham sido especificados valores para todos os parâmetros variáveis, assume-se que esses valores são idênticos aos da linha anterior.

A seção seguinte do arquivo de dados para síntese consiste em uma lista de constantes literais que podem ser utilizadas no programa LDR. Cada constante é definida em uma linha contendo o nome da constante e o seu valor, seguido opcionalmente por um comentário. Esta seção é finalizada pela palavra "FIM".

O caractere "%" permite a introdução de comentários em qualquer ponto do arquivo de dados para síntese. O texto localizado entre este caractere e o final da linha é ignorado pelo sistema.

A versão atual do arquivo de dados para síntese encontra-se reproduzida no apêndice B.

4.5 O MÓDULO CONVERSOR

Operação do módulo conversor

O módulo conversor (arquivo "CONV.CPP") é responsável pela aplicação das regras de síntese, traçando os contornos dos parâmetros de controle do sintetizador em função do tempo. Além disso, este módulo pode gerar arquivos que permitem visualizar de diferentes formas os resultados da execução do programa LDR. Como o sintetizador de Klatt está incorporado ao módulo conversor, este pode também gerar o arquivo com as amostras do sinal de fala.

A entrada do módulo conversor corresponde à saída da etapa de processamento prosódico, consistindo em uma transcrição fonética acrescida de durações segmentais e valores de freqüência fundamental. Especificando-se as opções "i7" ou "u7" (descritas na seção 1.4) no acionamento do sistema de conversão texto-fala, a entrada do módulo conversor passa a ser fornecida diretamente pelo usuário através do teclado (*default*) ou de um arquivo (caso seja especificada também a opção "f"). Conforme descrito em detalhe na seção 3.4, as durações e os valores de freqüência fundamental são especificados entre colchetes, separados por vírgulas, precedendo os fones aos quais se referem na transcrição fonética.

A primeira tarefa realizada pelo módulo conversor consiste na extração das informações prosódicas presentes na transcrição fonética, a partir das quais é criada uma lista contendo as durações de cada fone e é traçado o contorno de freqüência fundamental (parâmetro F0) para toda a frase. Em seguida, são adicionados silêncios (isto é, períodos

nos quais as fontes têm amplitudes nulas) no começo e no final da frase, evitando, por meio de transições suaves, o início e o término bruscos do sinal de fala.

O ciclo principal de operação do módulo conversor consiste em um laço no interior do qual é acionada a rotina responsável pela execução do programa LDR. A cada iteração deste laço, são gerados os contornos dos parâmetros em função do tempo para um fone da transcrição fonética de entrada. O programa LDR é executado tantas vezes quantos forem os fones presentes na transcrição, incluindo os silêncios adicionados automaticamente pelo sistema.

Ao final do processamento de cada fone, o sintetizador de Klatt é acionado para a geração das amostras correspondentes do sinal de fala. Entretanto, existem situações nas quais nem todos os contornos se encontram disponíveis após o processamento de um fone; por exemplo, quando é usado o modelo de interligação descrito na seção 4.3, o contorno do parâmetro no fone atual passa a depender de informações que somente serão conhecidas após o processamento do fone posterior. Em casos como esse, o acionamento do sintetizador de Klatt é adiado até que seja possível traçar os contornos de todos os parâmetros.

Arquivo de parâmetros

Especificando-se a opção de linha de comando "k"[4] no acionamento do sistema de conversão texto-fala, o módulo conversor gera o *arquivo de parâmetros* (arquivo-texto com extensão *default* ".KLA"), que consiste em uma tabela contendo os valores dos parâmetros de controle do sintetizador em função do tempo. O seu formato é idêntico ao utilizado no programa original de Klatt [Klatt 1980], com exceção da ordem em que são listados os parâmetros. Este arquivo pode também ser utilizado como entrada do sistema (através da opção de linha de comando "v"), gerando-se a partir dele o arquivo de amostras do sinal de fala.

O arquivo de parâmetros apresenta duas seções. Na primeira, são especificados os valores *default* de cada parâmetro, juntamente com um dígito cujo valor indica se o parâmetro variou (dígito diferente de zero) ou permaneceu constante (dígito igual a zero) durante a síntese. Os parâmetros que variaram têm seus valores listados em função do tempo na segunda seção do arquivo. A primeira linha dessa seção contém a duração total da síntese, em milissegundos. Na segunda linha, são listados os nomes dos parâmetros que variaram, na ordem em que estão tabelados. As linhas seguintes contêm os valores dos parâmetros em função do tempo (que é indicado, em milissegundos, no início de cada linha). O intervalo

[4] As opções de linha de comando do sistema de conversão texto-fala encontram-se descritas na seção 1.4.

de tempo entre valores consecutivos de um parâmetro é dado pela relação NWS / SR, como discutido na seção 4.2.

Na maioria dos casos, o tamanho do arquivo de parâmetros é aproximadamente metade do tamanho do arquivo de forma de onda correspondente (gerado pelo sintetizador de Klatt). O armazenamento de um arquivo de parâmetros em lugar de um arquivo de forma de onda implica, portanto, em uma economia de 50% no espaço ocupado em disco, necessitando-se apenas do sintetizador de Klatt para que o sinal de fala possa ser reproduzido.

Arquivo de parâmetros compacto

Como os contornos dos parâmetros são construídos a partir de segmentos de reta, há uma redundância significativa nas informações contidas no arquivo de parâmetros descrito no item anterior. O *arquivo de parâmetros compacto* (arquivo-texto com extensão *default* ".PAR") explora essa propriedade dos contornos, armazenando apenas os pontos extremos de cada segmento de reta; a partir dessa informação, os contornos podem ser reconstruídos de forma exata. A geração deste arquivo é ativada especificando-se a opção de linha de comando "p" no acionamento do sistema. A opção "t", por sua vez, permite a realização da síntese a partir de um arquivo de parâmetros compacto.

As seguinte seções estão presentes no arquivo de parâmetros compacto:

1. Valores dos parâmetros fixos, listados em uma única linha na seguinte ordem: SR, NWS, NFC, AN, A1, F5, F6, B4, B5, B6, FGP, FGZ, FNP, BGP, BGZ, BNP, BNZ, BGS, G0 e SW.
2. Duração total da síntese, em milissegundos.
3. Tempo de início da frase, em milissegundos, seguido da palavra "Frase:" e, opcionalmente, da transcrição fonética da frase.
4. Nomes dos parâmetros que, embora possam ser variados, permaneceram fixos durante a frase, listados em uma única linha.
5. Valores dos parâmetros que não variaram na frase, listados em uma única linha na mesma ordem em que são citados os seus nomes na seção anterior.
6. Pontos dos contornos de parâmetros. A cada linha está associado um instante de tempo indicado no seu início, em milissegundos, seguido alternadamente pelos nomes e pelos valores dos parâmetros cujos contornos apresentam um ponto nesse instante de tempo.

Caso o texto a partir do qual foi produzido o arquivo de parâmetros compacto possua mais do que uma frase, as seções 3, 4, 5 e 6 são repetidas para cada uma delas. Comentários podem ser introduzidos no arquivo, devendo-se precedê-los pelo símbolo "%".

Em geral, o tamanho de um arquivo de parâmetros compacto é aproximadamente igual a 10% do tamanho do arquivo de forma de onda correspondente; desse modo, o seu armazenamento representa uma economia substancial de espaço tanto em relação ao arquivo de forma de onda como em relação ao arquivo de parâmetros descrito no item anterior. O arquivo de parâmetros compacto apresenta ainda a vantagem de ser facilmente editável, constituindo-se em uma ferramenta de grande praticidade para o desenvolvimento e aperfeiçoamento de regras de síntese.

Arquivo de depuração

Especificando-se a opção "d" no acionamento do sistema, o módulo conversor gera o *arquivo de depuração* (arquivo-texto com extensão *default* ".DEP"), cuja função é auxiliar na detecção de erros nas regras de síntese. Esse arquivo lista todos os comandos LDR executados pelo módulo conversor, juntamente com os pontos dos contornos de parâmetros selecionados pelo usuário, denominados *parâmetros rastreados*. A sintaxe para a utilização da opção "d" encontra-se descrita na seção 1.4.

O arquivo de depuração é dividido em seções dedicadas a cada fone processado pelo módulo conversor. A primeira linha de cada seção identifica o fone ao qual ela se refere. Na segunda linha, são mostrados os contornos iniciais de cada parâmetro rastreado. As linhas seguintes identificam os comandos LDR executados no processamento do fone, precedidos pelos endereços (em hexadecimal) a eles associados no programa LDR. Após cada comando, são registrados os pontos dos contornos (tempo e valor) dos parâmetros rastreados, exceto para os comandos que não os afetam diretamente (CASO, FIM, SE, SEMOD e VA). Não são mostrados os nomes de constantes literais e de rótulos, mas sim os seus valores e endereços associados, pois esses nomes não estão disponíveis durante a execução do programa LDR. Ao início de cada frase, a respectiva transcrição fonética é registrada no arquivo de depuração.

Ao permitir a análise passo a passo da execução do programa LDR e da evolução dos contornos dos parâmetros, o arquivo de depuração constitui uma ferramenta de grande valia na detecção de erros nas regras de síntese. Muitos erros comuns ao implementar-se regras novas, como a inversão não intencional da ordem no tempo de dois pontos consecutivos do contorno de um parâmetro, são facilmente detectados através deste arquivo.

Arquivo de forma de onda

A partir das amostras do sinal de fala fornecidas pelo sintetizador de Klatt, o módulo conversor gera o *arquivo de forma de onda.* Esse arquivo, a partir do qual é realizada a reprodução do sinal de fala, pode ser gerado no formato Wave (extensão *default* ".WAV"), adotado pela maioria dos aplicativos que rodam em ambiente Windows, ou no formato NSP (extensão *default* ".NSP"), adotado pelo sistema CSL. Uma descrição do formato NSP pode ser encontrada em [Kay Elemetrics 1994].

A taxa de amostragem utilizada no arquivo de forma de onda é definida através do parâmetro SR do sintetizador de Klatt. O formato NSP aceita qualquer taxa de amostragem dada por um valor inteiro e não negativo de dois bytes. Já o formato Wave a rigor trabalha apenas com três taxas de amostragem: 11025 Hz, 22050 Hz e 44100 Hz. Embora a especificação de outras taxas possa causar problemas de compatibilidade, a maioria dos aplicativos não impõe esse tipo de restrição. Ambos os formatos utilizam amostras de 16 bits, com o byte menos significativo precedendo o mais significativo. Um cabeçalho contendo informações sobre o arquivo (quantidade de amostras, taxa de amostragem etc.) precede as amostras do sinal de fala; esse cabeçalho possui 44 bytes no arquivo Wave e 60 bytes no arquivo NSP.

Especificando-se a opção "r" ao acionar o sistema, o sinal de fala é automaticamente reproduzido após a síntese, no CSL ou em uma placa de som padrão Sound Blaster. Além disso, a opção "h" permite a reprodução de um arquivo de forma de onda previamente gerado. A opção "o", por sua vez, inibe a geração do arquivo de forma de onda. Essas opções encontram-se descritas em detalhe na seção 1.4.

CAPÍTULO 5
Considerações finais

Neste capítulo são retomados alguns tópicos apresentados ao longo da dissertação, destacando as principais contribuições e discutindo o desempenho geral do sistema de conversão texto-fala implementado. São também relacionadas algumas possibilidades de estudos futuros na mesma linha deste trabalho.

5.1 PRINCIPAIS CONTRIBUIÇÕES

Processamento de texto

O módulo de pré-processamento de texto é responsável pelo tratamento inicial de siglas, abreviações, símbolos especiais e algarismos. A principal inovação introduzida neste módulo foi a eliminação de acentos e cedilhas do texto de entrada, convertidos em atributos associados às letras. Como as comparações entre as palavras do texto de entrada e os termos presentes nas várias bases de dados utilizadas pelo sistema são feitas a partir do texto pré-processado, o sistema como um todo se torna mais robusto a erros de acentuação e mesmo à total ausência de acentos.

O processo de classificação gramatical de uma palavra se divide em três etapas: (1) comparação direta da palavra com termos previamente classificados; (2) comparação da terminação da palavra com terminações típicas da língua portuguesa às quais são associadas determinadas classes gramaticais; e (3) classificação da palavra através de regras que analisam as classificações de palavras vizinhas. Esta metodologia de classificação é original e apresenta como principal vantagem a utilização de uma base de dados relativamente pequena.

O módulo de transcrição ortográfico-fonética utiliza um conjunto de regras que atua diretamente sobre o texto pré-processado, convertendo-o em uma seqüência de símbolos fonéticos. O emprego de informações gramaticais torna a transcrição mais confiável, permitindo a diferenciação na pronúncia de palavras homógrafas. Um dicionário de exceções armazena as pronúncias correspondentes a palavras transcritas incorretamente pelas regras. A principal inovação presente neste módulo é a utilização de expressões regulares no dicionário de exceções, evitando o armazenamento de todas as variações de uma palavra (flexões de gênero, número e grau, conjugações etc.).

Processamento prosódico

O módulo para determinação de fronteiras prosódicas é responsável pela quebra das frases de entrada em grupos prosódicos, que correspondem a regiões com características prosódicas até certo ponto independentes das observadas nas demais regiões. Na maioria dos casos, a definição de grupo prosódico coincide com o conceito de oração. A técnica utilizada leva em conta a pontuação do texto de entrada e informações gramaticais; embora bastante simples, ela apresenta em geral um bom desempenho.

O processo utilizado para a geração de contornos de entonação toma por base contornos extraídos de elocuções naturais, armazenados no dicionário de contornos de entonação. Para cada grupo prosódico de entrada, um contorno do dicionário é selecionado com base em informações gramaticais e ajustado às particularidades do grupo prosódico levando-se em conta o posicionamento das sílabas tônicas. Esta técnica foi inspirada no gerador de contornos de entonação para a língua francesa descrito em [Emerard et al. 1992] e apresenta diversos elementos originais.

O módulo para geração das durações dos fones se baseia na suposição de que efeitos globais e locais sobre as durações podem ser tratados separadamente. Os efeitos globais, correspondentes à influência do grupo prosódico como um todo sobre a duração de cada fone, são obtidos a partir de dados extraídos de elocuções naturais, armazenados no dicionário de durações. Para cada grupo prosódico de entrada, um item do dicionário de durações é selecionado e adaptado, levando-se em conta informações gramaticais e o posicionamento das sílabas tônicas. Esta técnica, totalmente original, foi inspirada no tratamento dado pelo sistema aos contornos de entonação. Os efeitos locais, por sua vez, correspondem à influência da vizinhança imediata de um fone sobre a sua duração, sendo gerados por meio de um conjunto de regras similar ao elaborado por Klatt [Allen et al. 1987].

Síntese do sinal de fala

A etapa de síntese do sistema de conversão texto-fala tem como elemento central o sintetizador de formantes de Klatt, responsável pela geração das amostras do sinal de fala. A operação do sintetizador é controlada por um conjunto de parâmetros cujos valores devem ser renovados periodicamente. Além dos 39 parâmetros utilizados pelo sintetizador original, dois outros (*jitter* e *shimmer*) foram acrescentados com o objetivo de aumentar a naturalidade da fala produzida. Outra modificação em relação ao sintetizador original foi a introdução de interpolação linear entre valores consecutivos de coeficientes de filtros (que antes eram alterados abruptamente para cada novo conjunto de valores dos parâmetros de controle), eliminando distorções no sinal de fala sintetizado.

Os valores dos parâmetros de controle do sintetizador são obtidos pelo módulo conversor através das regras de síntese (elaboradas por Edson José Nagle em sua pesquisa de doutorado [Nagle 1991]), tendo como entrada a transcrição fonética acrescida de informações prosódicas. Para a implementação das regras, foi criada uma linguagem específica, denominada Linguagem para Descrição de Regras (LDR), e um compilador associado (o compilador de regras). A utilização de uma linguagem específica para a descrição das regras de síntese constitui a principal inovação introduzida neste módulo em relação a outros sistemas baseados no sintetizador de Klatt [Allen et al. 1987, Klatt 1987].

Após a aplicação das regras de síntese, obtém-se uma tabela contendo os valores assumidos pelos parâmetros de controle do sintetizador ao longo da síntese. A tabela completa pode ser registrada pelo módulo conversor em um arquivo no formato original definido por Klatt [Klatt 1980]. Estes dados podem ser também registrados utilizando um formato compacto que explora o fato de os contornos dos parâmetros serem constituídos por segmentos de reta. Arquivos em ambos os formatos podem ser utilizados como entrada do sistema, gerando-se o sinal de fala correspondente. O módulo conversor pode ainda gerar o arquivo de depuração, que registra passo a passo a execução do programa LDR e permite avaliar a evolução dos contornos dos parâmetros, facilitando a localização de erros nas regras de síntese. O conjunto dos arquivos gerados pelo módulo conversor torna possível visualizar por diferentes ângulos o resultado da aplicação das regras de síntese; além disso, a possibilidade de alteração e reintrodução no sistema de alguns desses arquivos (sobretudo o arquivo de parâmetros no formato compacto) constitui um recurso de grande valia para o aprimoramento das regras.

5.2 AVALIAÇÃO DO DESEMPENHO GERAL DO SISTEMA

A inteligibilidade é o parâmetro mais importante na avaliação do desempenho de um sistema de conversão texto-fala. Um teste formal de inteligibilidade requer a participação de um número relativamente grande de ouvintes nativos da língua, verificando-se o nível de compreensão de frases e palavras isoladas [Allen et al. 1987]. A análise dos dados estatísticos fornecidos por este tipo de teste permite a identificação precisa de pontos do sistema que necessitem de aperfeiçoamento. Para o sistema descrito neste trabalho, foram realizados somente testes informais de inteligibilidade com um número reduzido de ouvintes, sem a preocupação de obter resultados numéricos. Estes testes mostraram que o sistema apresenta um bom desempenho geral. A sua principal deficiência reside na baixa inteligibilidade de alguns sons oclusivos, sobretudo o fone /k/, requerendo aprimoramentos nas regras de síntese. Testes formais de inteligibilidade deverão ser aplicados em estágios futuros do trabalho, após a correção das deficiências mais evidentes.

Outro parâmetro de grande importância na avaliação de um sistema de conversão texto-fala é a naturalidade da fala produzida. Ao contrário do que ocorre para testes de inteligibilidade, um teste de naturalidade se baseia em critérios exclusivamente subjetivos, dificultando a obtenção de valores numéricos precisos que expressem o seu resultado. Como a naturalidade está diretamente relacionada à prosódia, o resultado de um teste deste tipo constitui em certa medida uma avaliação dos módulos de processamento prosódico, embora deficiências nas regras de síntese e na transcrição fonética também possam causar sérios prejuízos à naturalidade da fala sintetizada; além disso, erros na etapa de classificação gramatical podem comprometer o processamento prosódico. Apesar de não haverem sido realizados testes formais de naturalidade para o sistema implementado, avaliações informais indicam que ainda há bastante espaço para o aprimoramento do sistema como um todo e, em particular, dos módulos de processamento prosódico. Algumas sugestões nesse sentido encontram-se enumeradas na próxima seção.

5.3 POSSIBILIDADES DE ESTUDOS FUTUROS

A seguir estão relacionadas algumas possibilidades de estudos futuros seguindo a linha geral deste trabalho.

Processamento de texto

Os módulos de processamento de texto implementados neste trabalho caracterizam-se por utilizar algoritmos relativamente simples e bases de dados de pequenas dimensões. Um desempenho sensivelmente superior seria obtido caso fossem utilizadas técnicas de análise morfológica. Uma base de dados conteria em princípio todos os morfemas da língua portuguesa, juntamente com as respectivas transcrições fonéticas, divisões silábicas e classificações gramaticais características. Conjuntos de regras seriam necessários para tratar mutações sonoras nas fronteiras entre morfemas e para determinar a classificação gramatical de uma palavra com base na sua composição morfológica. Regras posicionais para classificação gramatical seriam requeridas para o tratamento de palavras com múltiplas classificações possíveis; além disso, regras para transcrição ortográfico-fonética e divisão silábica seriam aplicadas a palavras compostas por morfemas não reconhecidos pelo sistema. O custo do melhor desempenho obtido pelo uso da análise morfológica é o grande aumento no tamanho da base de dados empregada.

Uma característica desejável em um sistema de conversão texto-fala de finalidade geral para a língua portuguesa é a robustez à falta de acentos no texto, muito comum em mensagens transmitidas por correio eletrônico. Na implementação de um analisador morfo-

lógico, essa característica poderia ser obtida desprezando-se os acentos durante a busca por morfemas, levando-os em conta apenas quando fosse necessário decidir entre dois ou mais morfemas possíveis.

Processamento prosódico

O desempenho dos módulos de processamento prosódico pode ser melhorado por meio do acréscimo de novos itens aos dicionários de durações e de contornos de entonação. Passando-se dos atuais 30 itens para 50, por exemplo, certamente seria obtida uma fala de qualidade bastante superior. Frases tratadas inadequadamente pelos módulos de processamento prosódico são candidatas naturais à inserção nos dicionários; os testes de naturalidade constituem, portanto, a ferramenta mais apropriada para a seleção das elocuções naturais a utilizar na criação de novos itens.

As regras atualmente em uso para a determinação de efeitos locais sobre as durações dos fones foram elaboradas com base em trabalhos similares que enfocam a língua inglesa [Allen et al. 1987]. Um estudo sistemático do comportamento das durações no português permitiria a criação de regras melhor adaptadas às particularidades desta língua, aumentando a naturalidade da fala produzida. Técnicas de otimização poderiam ser empregadas na obtenção de valores para os vários coeficientes presentes nas regras, procurando minimizar o erro quadrático médio das durações geradas em relação a durações medidas em elocuções naturais.

Como o principal critério para a seleção de itens do dicionário de contornos de entonação é a similaridade gramatical com a frase de entrada, é pequena a probabilidade de que coincidam as quantidades de sílabas tônicas na frase e no item escolhido. O ajuste à frase de entrada de um contorno cuja sentença de origem apresenta uma quantidade diferente de sílabas tônicas prejudica a naturalidade da fala sintetizada, uma vez que a freqüência fundamental geralmente apresenta picos locais nessas regiões. Uma solução seria a divisão do processo de geração de contornos em duas etapas: (1) tratamento de efeitos globais, causados pela sentença como um todo, utilizando a técnica de ajuste de contornos de entonação obtidos de sentenças naturais; e (2) tratamento de efeitos locais, decorrentes do contexto em que se situa cada trecho do contorno, por meio de um conjunto de regras. Esta técnica, semelhante à utilizada para a geração de durações, requereria que os contornos de entonação armazenados no dicionário fossem previamente tratados de forma a eliminar variações decorrentes de efeitos locais, mantendo-se apenas o contorno-base característico da estrutura sintática. Variações de freqüência fundamental associadas a sílabas tônicas ou a quaisquer outros efeitos locais, determinadas por meio de regras, seriam sobrepostas a esse contorno-base. Uma vantagem desta abordagem é a liberdade na alteração do contorno,

permitindo, por exemplo, a manipulação da freqüência fundamental de modo a enfatizar determinadas palavras. Uma desvantagem, por outro lado, é a dificuldade na formulação de regras que reproduzam com detalhes as variações observadas em contornos naturais, exigindo um estudo minucioso dos efeitos locais que agem sobre a entonação na língua portuguesa.

Técnicas de processamento prosódico baseadas em redes neurais tem apresentado bons resultados tanto para a geração de durações de segmentos como de contornos de entonação [Traber 1992, Violaro et al. 1996]. Esta abordagem constitui mais uma possibilidade para o aperfeiçoamento dos módulos de processamento prosódico do sistema.

Síntese do sinal de fala

Conforme comentado na seção anterior, são necessários aperfeiçoamentos nas regras de síntese, sobretudo para alguns sons oclusivos cuja inteligibilidade está aquém do desejável. Em sua versão atual, a maioria das regras não leva em conta o contexto em que se insere cada fone, desprezando efeitos de coarticulação com influência significativa tanto na inteligibilidade como na naturalidade da fala produzida. Trabalhos relacionados a regras de síntese para a língua portuguesa utilizando o sintetizador de formantes de Klatt estão sendo atualmente conduzidos por Edson José Nagle em sua pesquisa de doutorado.

A linguagem LDR, descrita no capítulo 4, permite implementar as regras de síntese de maneira relativamente simples; no entanto, conforme as regras se tornem mais elaboradas, recursos adicionais provavelmente se farão necessários. Dois possíveis aprimoramentos na linguagem LDR são:

- Possibilidade de definição de vetores, permitindo a associação de mais do que um valor a uma constante literal; o valor a utilizar seria escolhido durante a execução do programa em função do contexto em que se insere cada fone.
- Introdução de recursos para a formulação de regras que atuem em domínios superiores ao do fone individual, como, por exemplo, a sílaba e a palavra [Hertz 1982, Klatt 1987].

Uma vez atingido um nível satisfatório de qualidade para a voz atualmente disponível no sistema, pode-se partir para a incorporação de novas vozes, eventualmente incluindo vozes femininas e infantis. A versão em uso do sintetizador de Klatt, no entanto, não apresenta bons resultados para a síntese dessas classes de vozes. Modificações no modelo de síntese visando a produção de vozes femininas foram propostas em [Klatt e Klatt 1990]; a implementação destas e de outras modificações pode vir a ser efetuada em etapas futuras do trabalho.

Interface com o usuário

A interface atualmente em uso para comunicação entre o sistema e o usuário, baseada em opções de linha de comando e na leitura e escrita de arquivos-texto, foi criada com o objetivo de facilitar a realização de testes durante o desenvolvimento do sistema. Aplicações práticas, no entanto, exigiriam uma interface que tornasse as operações internas transparentes ao usuário e permitisse a obtenção de um sinal de fala a partir de um texto em português utilizando comandos simples. O estudo e implementação desta interface é uma etapa necessária para uma eventual comercialização do sistema.

Outro desenvolvimento importante é a criação de uma interface específica para a utilização do sistema por deficientes visuais. Em sua forma mais simples, esta interface consistiria em um leitor de textos, possivelmente utilizando um *scanner* e um software de OCR para a captura de textos impressos. Um exemplo de interface mais elaborada seria um sistema para auxílio ao uso de computadores por deficientes visuais, empregando recursos de conversão texto-fala para facilitar o acesso a programas genéricos (como editores de texto, por exemplo) e à Internet (lendo o conteúdo de páginas em português).

APÊNDICE A

Regras utilizadas na etapa de processamento de texto

Regras posicionais para classificação gramatical

Nas regras a seguir, são utilizadas as abreviações definidas no capítulo 2 para as classes e subclasses gramaticais. A notação "(*subclasse*) = *número*" indica que o coeficiente da subclasse (ou grupo de subclasses) entre parênteses deve ser igualado ao número fornecido; já as notações "(*subclasse*) + +" e "(*subclasse*) – –" indicam que o coeficiente da subclasse (ou grupo de subclasses) entre parênteses deve ser incrementado ou decrementado, respectivamente, respeitando-se os valores mínimo (0) e máximo (3) admitidos. Exceto quando indicado, todos os testes realizados pelas regras referem-se à palavra atual (isto é, à palavra que está sendo analisada). Testes do tipo "a palavra é *classe ou subclasse*" verificam se a palavra em questão possui a subclasse citada (ou alguma subclasse da classe citada) com coeficiente não nulo. Testes do tipo "a palavra provavelmente é *classe ou subclasse*", por outro lado, verificam se a classe ou subclasse citada é a mais provável (isto é, aquela com maior coeficiente) na palavra em questão.

Se (PRO-OA) e (está associada à palavra anterior)
→ PRO-OA = 3, (subclasses que não pertençam a PRO) = 0

Se (VER) e (a próxima palavra é PRO-OA) e (está associada à próxima palavra)
→ (subclasses que não pertençam a VER) = 0

Se (PRO-RE) e (PRO-OT)
→ Se (a palavra anterior é PRE)
  → PRO-OT + +, PRO-RE – –
  Senão
  → PRO-RE + +, PRO-OT – –

Se (PRO-OA) e ((a palavra anterior é PRO-RE) ou (a palavra anterior é SUB) ou (a palavra anterior é ADJ)) e (a próxima palavra é VER) e (não é a primeira da frase)
→ PRO-OA + +, (subclasses que não pertençam a PRO) – –

Se (ART) e ((é a primeira palavra da frase) ou ((a palavra anterior provavelmente é VER) e (a palavra seguinte provavelmente é SUB ou ADJ)))
→ ART + +, (subclasses que não pertençam a ART) – –

Se ("se") e (é a primeira palavra da frase)
→ CON-CD + +, (subclasses que não pertençam a CON) – –

Se ("se") e (a próxima provavelmente não é VER) e (não está associada à palavra anterior)
→ CON-CD ++, (subclasses que não pertençam a CON) --
Senão
→ PRO-OA ++, (subclasses que não pertençam a PRO) --

Se ("que") e (a palavra anterior é "de")
→ PRO-RL ++, (subclasses que não pertençam a PRO) --
  Se (a palavra anterior é a primeira da frase) e (a frase é interrogativa)
  → ADV-IT ++

Se ("que") e (a palavra anterior é "do")
→ CON-CR ++, (subclasses que não pertençam a CON) --
  Se (a palavra anterior é a primeira da frase) e (a frase é interrogativa)
  → ADV-IT ++

Se ("que") e (a palavra anterior é "para")
→ CON-CS ++, (subclasses que não pertençam a CON) --
  Se (a palavra anterior é a primeira da frase) e (a frase é interrogativa)
  → ADV-IT ++

Se ("que") e ((a palavra anterior é SUB) ou (a palavra anterior é ADJ))
→ PRO-RL ++, (subclasses que não pertençam a PRO) --

Se ("que") e (a palavra anterior é "por")
→ ADV-IT ++

Se ("qual") e (a palavra anterior é "tal")
→ CON-CR ++, (subclasses que não pertençam a CON) --

Se ("qual") e ((a palavra anterior é "do") ou (a palavra anterior é "da"))
→ PRO-RL ++

Se (VER-SB) e ((VER-IM) ou (VER-IE)) e (há uma conjunção subordinativa antes da palavra atual)
→ Se (as palavras entre a conjunção subordinativa e a palavra atual são SUB, ADJ, ART ou PRO-RE)
  → VER-SB ++, VER-IM --, VER-IE --

Se (VER-IM) e (é a primeira palavra da frase)
→ VER-IM ++

Se (VER-IE) e (a palavra anterior é "não", "nunca" ou "jamais") e (a palavra anterior é a primeira da frase)
→ VER-IE ++
  Se (VER-IM)
  → VER-IM = 0

Se (VER) e (a palavra anterior é VER) e (a palavra anterior à anterior é VER)
→ (subclasses de VER) --

Se (VER) e (a palavra anterior é PRO-RE)
→ (subclasses de VER para pessoa e número correspondentes às subclasses não nulas de PRO na palavra anterior) ++, (demais subclasses) --

Se (VER) e (a palavra anterior provavelmente é SUB ou ADJ)
→ VER-EL ++, (subclasse de VER para número correspondente à subclasse não nulas de SUB ou ADJ na palavra anterior) ++, (demais subclasses de pessoa e número de VER) --

Se (SUB) e (a palavra anterior é ART)
→ (subclasses de SUB) = (subclasses de ART), (subclasses de ADJ) = (subclasses de ART), (subclasses que não pertençam a SUB) --

Se (a palavra anterior é PRO-PO)
→ (subclasses de SUB) = (subclasses de PRO relativas ao objeto),
(subclasses de ADJ) = (subclasses de PRO relativas ao objeto),
(subclasses que não pertençam a SUB) --

Se (SUB) e (inicia com letra maiúscula) e (não está no começo da frase)
→ (subclasses não nulas de SUB) ++, SUB-PR = 3, SUB-CM --,
(subclasses que não pertençam a SUB) --

Se (ADJ) e (a palavra anterior é SUB ou ADJ)
→ ADJ ++, (subclasses que não pertençam a ADJ) --

Algoritmo para divisão silábica

No algoritmo para divisão silábica mostrado abaixo, as seguintes notações foram empregadas:

- Variáveis estão representadas em itálico.
- Um número, variável ou expressão entre colchetes corresponde ao caractere da palavra na posição indicada, sendo o primeiro caractere associado à posição zero; por exemplo, para a palavra "disco", "[3]" corresponde à letra "c".
- Letras são representadas entre aspas e grupos de letras são definidos entre chaves e utilizando a vírgula como separador ({"a", "e", "i"}, por exemplo).

```
pos = 0
ini = 0
final = 0
Enquanto final != posição da última letra  // Repete enquanto não atingir o final da palavra
   conc = falso
   aeo = falso
   Enquanto conc = falso
      Se pos+1 > posição da última letra
         final = pos
         conc = verdadeiro
      Fim
      Senão
```

```
Procurar a próxima vogal da palavra a partir de pos e atribuir sua posição a pos
Se pos+1 > posição da última letra
   final = posição da última letra
   conc = verdadeiro
Fim
Senão
   Se [pos+1] está em VOGAL
      Se ([pos+1] está em {"i" sem acento, "u" sem acento}) e
      ([pos+3] não está em VOGAL) e
      ((([pos+2] está em {"b", "d", "f", "k", "p", "t", "v"}) e
      ([pos+3] não está em {"h", "l", "r"})) ou (([pos+2] = "c") e
      ([pos+3] não está em {"h", "l", "r", "k"})) ou ([pos+2] = "h")
      ou (([pos+2] está em {"j", "l", "m", "r", "x", "z"} e ([pos+3] != "h"))
      ou (([pos+2] = "g") e ([pos+3] não está em {"l", "r"})) ou ([pos+2] = "q"}))
         final = pos
         conc = verdadeiro
      Fim
      Se (conc = falso)
         Se [pos] está em {"a", "e", "o"}
            Se ([pos+1] está em {"a", "e", "o", "i" com acento, "u" com acento})
               final = pos
               conc = verdadeiro
            Fim
            Senão
               aeo = verdadeiro
            Fim
         Fim
         Senão // [pos] está em {"i", "u"}
            Se ([pos] tem acento) ou ([pos+1] tem acento) ou
            ((pos+1 = posição da última letra) e
            ([pos+1] está em {"a", "e", "o"}))
            ou (aeo = verdadeiro) ou ([pos] = [pos+1])
               final = pos
               conc = verdadeiro
            Fim
         Fim
      Fim
      pos = pos+1
   Fim
   Senão // [pos+1] está em CONSOANTE
      Enquanto (pos+2 <= posição da última letra) e
      ([pos+2] não está em VOGAL) e
      ((([pos+1] está em {"b", "d", "f", "k", "p", "t", "v"}) e
      ([pos+2] não está em {"h", "l", "r"}))
      ou (([pos+1] = "c") e ([pos+2] não está em {"h", "l", "r", "k"}))
      ou ([pos+1] = "h")
      ou (([pos+1] está em {"j", "l", "m", "n", "r", "s", "x", "z"} e
      ([pos+2] != "h"))
      ou (([pos+1] = "g") e ([pos+2] não está em {"l", "r"}))
      ou ([pos+1] = "q"}))
```

```
                pos = pos+1
              Fim
              Se (pos+2 > posição da última letra)
                final = pos+1
              Fim
              Senão
                Procurar a próxima vogal da palavra a partir de pos+2
                Se não encontrou vogal
                  final = posição da última letra
                  pos = posição da última letra + 1
                Fim
                Senão
                  final = pos
                  pos = pos+1
                Fim
              Fim
              conc = verdadeiro
            Fim
          Fim
        Fim
      Fim
      Armazena início (ini) e final (final) de sílaba
      ini = pos
Fim
```

Regras para transcrição ortográfico-fonética

Nas regras de transcrição ortográfico-fonética mostradas a seguir, os caracteres ortográficos são representados em itálico e os símbolos fonéticos são representados em tipo comum. A notação "*x* = y" indica que o caractere ortográfico "*x*" deve ser substituído pelo símbolo fonético "y". Quando dois caracteres ortográficos consecutivos devem ser substituídos por um único símbolo fonético, emprega-se a notação "*x*+ = y". As regras devem ser aplicadas sucessivamente a cada caractere ortográfico do texto de entrada, iniciando pelo caractere mais à esquerda e avançando sempre para o próximo caractere ainda não substituído. Exceto quando indicado, as comparações referem-se à letra atual (isto é, à letra cuja pronúncia está sendo determinada). Os termos "anterior" e "próxima" referem-se às letras anterior e posterior à atual, respectivamente.

Vogais: *a, e, i, o, u*

Consoantes: *b, c, d, f, g, h, j, k, l, m, n, p, q, r, s, t, v, x, z*

Consoantes sonoras: *b, d, g, j, l, m, n, v*

Sonoras: vogais + consoantes sonoras

Letra *a*:

Se (tem til) ou (tem circunflexo) ou ((a próxima é *m* ou *n*) e (a próxima pertence à sílaba atual)) ⇒ *a* = ä
Senão ⇒ Se é sílaba pós-tônica ⇒ *a* = A
Senão ⇒ *a* = a

Letra *b*:

Se é final de sílaba ⇒ *b* = by
Senão ⇒ *b* = b

Letra *c*:

Se (tem cedilha) ou ((a próxima é *e* ou *i*) e (a próxima pertence à sílaba atual))
Se é dobrada ⇒ *c* = ks
Senão ⇒ *c* = s
Senão ⇒ Se é final de sílaba ⇒ *c* = ky
Senão ⇒ Se (a próxima é *h*) e (a próxima pertence à sílaba atual) ⇒ *c*+ = x
Senão ⇒ *c* = k

Letra *d*:

Se é final de sílaba ⇒ *d* = dJy
Senão ⇒ Se (a próxima é *i*) ou ((a próxima é *e*) e ((a próxima é final de palavra) ou
((a seguinte à próxima é *s*) e (a seguinte à próxima é final de palavra)))) ⇒ *d* = dJ
Senão ⇒ *d* = d

Letra *e*:

Se (a próxima é *m* ou *n*) e (a próxima pertence à sílaba atual) ⇒ *e* = ë
Senão ⇒ Se tem acento agudo ⇒ *e* = é
Senão ⇒ Se ((é final de palavra) ou ((a próxima é *s*) e (a próxima é final de palavra))) e
(não tem acento circunflexo) ⇒ *e* = y
Senão ⇒ Se ((((a palavra é verbo indicativo presente singular) ou
(a palavra é verbo indicativo presente plural de terceira pessoa) ou
(a palavra é verbo subjuntivo presente singular) ou
(a palavra é verbo subjuntivo presente plural de terceira pessoa) ou
(a palavra é verbo imperativo singular) ou
(a palavra é verbo imperativo plural de terceira pessoa)) e
(a primeira letra da próxima sílaba não é *m* ou *n*)) ou
(a palavra é substantivo feminino) ou (a palavra é adjetivo feminino) ou
(a palavra é pronome feminino)) e (a próxima não é vogal) e
((a anterior não pertence à sílaba atual) ou (a anterior não é vogal)) e
(é a penúltima sílaba da palavra) ⇒ *e* = é
Senão ⇒ *e* = e

Letra *f*:

f = f

Letra *g*:

Se é final de sílaba ⇒ *g* = gy
Senão ⇒ Se a próxima é *e* ou *i* ⇒ *g* = j
Senão ⇒ Se (a próxima é *u* sem trema) e (a próxima pertence à sílaba atual) e
(a seguinte à próxima é *e* ou *i*) ⇒ *g*+ = g
Senão ⇒ *g* = g

Letra *h*:

Ignorar esta letra (*h* mudo).

Letra *i*:

Se (a próxima é *m* ou *n*) e (a próxima pertence à sílaba atual)
Se é sílaba pós-tônica ⇒ *i* = ÿ
Senão ⇒ *i* = ï
Senão ⇒ Se (é sílaba pós-tônica) ou ((a próxima é vogal) e (a próxima pertence à sílaba atual)) ou ((a anterior é vogal) e (a anterior pertence à sílaba atual)) ⇒ *i* = y
Senão ⇒ *i* = i

Letra *j*:

j = j

Letra *k*:

Se é final de sílaba ⇒ *k* = ky
Senão ⇒ *k* = k

Letra *l*:

Se é final de sílaba ⇒ *l* = w
Senão ⇒ Se (a próxima é *h*) e (a próxima pertence à sílaba atual) ⇒ *l*+ = L
Senão ⇒ *l* = l

Letra *m*:

Se é final de sílaba ⇒ *m* = M
Senão ⇒ *m* = m

Letra *n*:

Se tem til ⇒ *n* = ñ
Senão ⇒ Se (a próxima é *h*) e (a próxima pertence à sílaba atual) ⇒ *n*+ = ñ
Senão ⇒ Se (é final de sílaba) ou ((a próxima é *s*) e (a próxima é final de sílaba))
Se a próxima é *t* ou *d* ⇒ *n* = N
Senão ⇒ *n* = Ñ
Senão ⇒ *n* = n

Letra *o*:

Se (tem til) ou ((a próxima é *m* ou *n*) e (a próxima pertence à sílaba atual)) ⇒ *o* = ö
Senão ⇒ Se tem acento agudo ⇒ *o* = ó
Senão ⇒ Se ((é final de palavra) ou ((a próxima é *s*) e (a próxima é final de palavra))) e (não tem acento circunflexo) ⇒ *o* = w
Senão ⇒ Se ((((a palavra é verbo indicativo presente singular) ou
(a palavra é verbo indicativo presente plural de terceira pessoa) ou
(a palavra é verbo subjuntivo presente singular) ou
(a palavra é verbo subjuntivo presente plural de terceira pessoa) ou
(a palavra é verbo imperativo singular) ou
(a palavra é verbo imperativo plural de terceira pessoa)) e
(a primeira letra da próxima sílaba não é *m* ou *n*)) ou
(a palavra é substantivo feminino) ou (a palavra é adjetivo feminino) ou
(a palavra é pronome feminino)) e (a próxima não é vogal) e
((a anterior não pertence à sílaba atual) ou (a anterior não é vogal)) e
(é a penúltima sílaba da palavra) ⇒ *o* = ó
Senão ⇒ *o* = o

Letra *p*:

Se é final de sílaba ⇒ *p* = py
Senão ⇒ Se (a próxima é *h*) e (a próxima pertence à sílaba atual) ⇒ *p*+ = f
Senão ⇒ *p* = p

Letra *q*:

Se é final de sílaba ⇒ *q* = ky
Senão ⇒ Se (a próxima é *u* sem trema) e (a próxima pertence à sílaba atual) e
(a seguinte à próxima é *e* ou *i*) ⇒ *q*+ = k
Senão ⇒ *q* = k

Letra *r*:

Se (é dobrada) ou (é início de palavra) ou ((a anterior é *m*, *n*, *l*, *s*, *x* ou *z*) e
(a anterior não é início de palavra)) ⇒ *r* = R
Senão ⇒ Se é final de sílaba ⇒ *r* = P
Senão ⇒ *r* = r

Letra *s*:

Se é dobrada ⇒ *s* = s
Senão ⇒ Se (não é início de palavra) e (((é final de palavra) e (a próxima é sonora)) ou
((a anterior é vogal) e (a próxima é vogal)) ou (a próxima é consoante sonora) ou
(as letras anteriores são *tran*)) ⇒ *s* = z
Senão ⇒ Se (a próxima é *c*) e (a seguinte à próxima é *e* ou *i*) ⇒ *s*+ = s
Senão ⇒ Se (a próxima é *h*) e (a próxima pertence à sílaba atual) ⇒ *s*+ = x
Senão ⇒ Se é a última letra da sílaba ⇒ *s* = S
Senão ⇒ *s* = s

Letra *t*:

Se é final de sílaba ⇒ *t* = tXy
Senão ⇒ Se (a próxima é *i*) ou ((a próxima é *e*) e ((a próxima é final de palavra) ou
((a seguinte à próxima é *s*) e (a seguinte à próxima é final de palavra)))) ⇒ *t* = tX
Senão ⇒ *t* = t

Letra *u*:

Se (a próxima é *m* ou *n*) e (a próxima pertence à sílaba atual)
Se é sílaba pós-tônica ⇒ *u* = Ü
Senão ⇒ *u* = ü
Senão ⇒ Se (é sílaba pós-tônica) ou ((a próxima é vogal) e (a próxima pertence à sílaba atual)) ou
((a anterior é vogal) e (a anterior pertence à sílaba atual)) ⇒ *u* = w
Senão ⇒ *u* = u

Letra *v*:

v = v

Letra *x*:

Se é final de palavra
Se (a anterior é *e*) e (a anterior é início de palavra) ⇒ *x* = S
Senão ⇒ *x* = kS
Senão ⇒ Se a anterior é *e*
Se (a anterior é início de palavra) e (a próxima não é *c*) ⇒ *x* = z
Senão ⇒ Se a próxima é *c* ⇒ *x*+ = s
Senão ⇒ Se a próxima é consoante ⇒ *x* = S
Senão ⇒ *x* = ks

Senão ⇒ Se (é início de palavra) ou ((a anterior é vogal) e (a anterior à anterior é vogal) e (a próxima é vogal)) ⇒ *x* = x
Senão ⇒ *x* = ks

Letra *z*:

Se a próxima não é sonora ⇒ *z* = S
Senão ⇒ *z* = z

APÊNDICE B

Regras de síntese e arquivo de dados para síntese

Regras de síntese

```
% Regras de síntese - arquivo "REGRAS.LDR"

% Definição de grupos de fones
% Grupo           Fones
VOGAL             i e é a ó o u   ï ë ä ö ü   y w A ÿ Ü
FRICATIVA         s z x j f v R   X J S
LIQUIDA_E_NASAL   l m n   ñ   L   M N Ñ
OCLUSIVA          p b t d k g
OCLUSIVA-SURDA    p t k
OCLUSIVA-SONORA   b d g
FIM

% Definição de modificações e caracteres modificadores
% Modificação     Caractere
TONICA-LEXICAL        '
TONICA-FRASAL         "
FIM

% Posição de início de tratamento para fones
% ou grupos de fones
% Fone ou grupo    Rótulo
$                  INICIO_$
VOGAL              INICIO_VOGAL
FRICATIVA          INICIO_FRICATIVA
f                  INICIO_f
LIQUIDA_E_NASAL    INICIO_LIQUIDA_E_NASAL
r                  INICIO_r
P                  INICIO_r
OCLUSIVA           INICIO_OCLUSIVA
FIM

% Descrição das regras de síntese

INICIO_$:  % Silêncio

CASO 1  % Faz transição entre o silêncio e o fone seguinte.
  PARA VOGAL:
    DUIP FONS T$V-FONS
    % Se for o primeiro fone da frase, toda a transição deve estar no próximo
    % fone.
    SEMOD SOM-INICIAL
      DUIP FONS T$V-FONS
      FIM
    FIM

  PARA FRICATIVA:
    DUIP FONT T$F-FONT
    % Se for o primeiro fone da frase, toda a transição deve estar no próximo
    % fone.
    SEMOD SOM-INICIAL
      DUIP FONT T$F-FONT
      FIM
    FIM

  PARA LIQUIDA_E_NASAL:
    DUIP FONT T$L-FONT
```

```
      % Se for o primeiro fone da frase, toda a transição deve estar no próximo
      % fone.
      SEMOD SOM-INICIAL
        DUIP FONT T$L-FONT
        FIM
      FIM

    PARA OCLUSIVA:
      DUIP FONT T$P-FONT
      % Se for o primeiro fone da frase, toda a transição deve estar no próximo
      % fone.
      SEMOD SOM-INICIAL
        DUIP FONT T$P-FONT
        FIM
      FIM

    PARA r:
      DUIP FONT T$r-FONT
      % Se for o primeiro fone da frase, toda a transição deve estar no próximo
      % fone.
      SEMOD SOM-INICIAL
        DUIP FONT T$r-FONT
        FIM
      FIM
    FIM

FIM  % Finaliza a execução.

INICIO_f:  % f

% Faz com que a amplitude do ruído no fone 'f' tenha uma inclinação positiva.
SOME AF V1 DUPf-AF
VA INICIO_FRICATIVA  % Prossegue com as demais fricativas.

INICIO_FRICATIVA:  % Fricativas e africadas

CASO 1  % Faz transição entre a fricativa ou africada e o fone seguinte.
    PARA VOGAL:
      DUIP FBn TFV-FBn
      DUIP FONT TFV-FONT
      FIM

    PARA LIQUIDA_E_NASAL:
      DUIP FBn TFL-FBn
      DUIP FONT TFL-FONT
      FIM

    PARA OCLUSIVA:
      DUIP FBn TFP-FBn
      DUIP FONT TFP-FONT
      FIM

    PARA r:
      DUIP FBn TFr-FBn
      DUIP FONT TFr-FONT
      FIM

    PARA $:
      DUIP FONT TF$-FONT
      FIM
    FIM

FIM  % Finaliza a execução.

INICIO_VOGAL:  % Vogais

CASO 1  % Faz transição entre a vogal e o próximo fone.
    PARA VOGAL:
      DUIP FBn TVV-FBn
      FIM

    PARA FRICATIVA:
      DUIP FBn TVF-FBn
      DUIP FONT TVF-FONT
```

```
    FIM
  PARA LIQUIDA_E_NASAL:
    DUIP FBn TVL-FBn
    DUIP FONT TVL-FONT
    FIM
  PARA OCLUSIVA:
    DUIP FBn TVP-FBn
    DUIP FONT TVP-FONT
    FIM
  PARA r:
    DUIP FBn TVr-FBn
    DUIP FONT TVr-FONT
    FIM
  PARA $:
    DUIP FONS TV$-FONS
    FIM
  FIM
FIM  % Finaliza a execução.
INICIO_r:  % 'r' e 'P'
CASO 1  % Faz transição entre o 'r' e o próximo fone.
  PARA VOGAL:
    DUIP FBn TrV-FBn
    DUIP FONT TrV-FONT
    FIM
  PARA FRICATIVA:
    DUIP FBn TrF-FBn
    DUIP FONT TrF-FONT
    FIM
  PARA OCLUSIVA:
    DUIP FBn TrP-FBn
    DUIP FONT TrP-FONT
    FIM
  PARA LIQUIDA_E_NASAL:
    DUIP FBn TrL-FBn
    DUIP FONT TrL-FONT
    FIM
  PARA $:
    DUIP FONT Tr$-FONT
    FIM
  FIM
FIM  % Finaliza a execução.
INICIO_LIQUIDA_E_NASAL:  % Líquidas e nasais
CASO 1  % Faz transição entre a líquida ou nasal e o próximo fone.
  PARA VOGAL:
    DUIP FBn TLV-FBn
    DUIP FONT TLV-FONT
    FIM
  PARA FRICATIVA:
    DUIP FBn TLF-FBn
    DUIP FONT TLF-FONT
    FIM
  PARA OCLUSIVA:
    DUIP FBn TLP-FBn
    DUIP FONT TLP-FONT
    FIM
  PARA $:
    DUIP FONT TL$-FONT
```

```
    FIM
  FIM

FIM  % Finaliza a execução.

INICIO_OCLUSIVA:  % Oclusivas

% Desenha contorno das fontes de voz.

SUB FONS T1 TPX-FONT

% Posiciona a explosão nas fontes de voz simultaneamente à explosão na
% fonte de ruído de fricação.
ACP FONS 2 DUR 55
SUB FONS T2 TPX-FONT

% Verifica se é uma oclusiva sonora; se for, faz com que a explosão nas fontes
% sonoras ocorra juntamente com a explosão na fonte de ruído de fricação.
SE 0 OCLUSIVA-SONORA

  SUB FONS T1 DRF
  SOME FONS T1 DTRF1
  SUB FONS T2 DTRF2
  FIM

% Desenha contorno da fonte de ruído.

% Retarda a explosão do ruído de fricação.
IGUALE AF T0 DUR
SUB AF T0 TPX-FONT
SUB AF T0 DRF
SOME AF T0 DTRF1

% Insere ponto no início da subida de AF.
ACP AF 0 DUR 0
SUB AF T0 TPX-FONT
SUB AF T0 DRF

% Insere ponto para que o ruído de fricação seja inicialmente nulo.
ACP AF 0 TVP-FONT 0

% Define a duração do período estável do ruído de fricação.
SUB AF T3 TPX-FONT
SUB AF T3 DTRF2

% Anula o ruído de fricação ao final do fone.
ACP AF 4 DUR 0
SUB AF T4 TPX-FONT

% Transição dos formantes.
CASO 1
  PARA VOGAL:
    DUIP FBn TPV-FBn
    FIM
  PARA FRICATIVA:
    DUIP FBn TPF-FBn
    FIM
  PARA r:
    DUIP FBn TPr-FBn
    FIM
  PARA LIQUIDA_E_NASAL:
    DUIP FBn TPL-FBn
    FIM
  FIM

FIM  % Finaliza a execução.
```

Arquivo de dados para síntese

```
% Dados para síntese
% Valores dos parâmetros fixos
%     F0    SR   NWS   NFC    AN    A1    F5    F6    B4    B5    B6   FGP   FGZ   FNP   BGP   BGZ   BNP   BNZ   BGS    G0    SW
     120 20000   100     6     0     0  5600  6700   200   300   300     0  1500   270   100  6000   100   100   200    55     0
% Valores dos parâmetros variáveis
%     AV    AF    AH   AVS    F1    F2    F3    F4    B1    B2    B3   FNZ    A2    A3    A4    A5    A6    AB    JI    SH   DUR
       0     0     0     0   700  1350  2600  3700   110   130   140   270     0     0     0     0     0     0    10    20   100
 $     0     0     0     0     /     /     /     /     /     /     /     /     /     /     /     /     /     /
 i    55     0    50    45   270  2300  3100  3700    60   120   150   270
 e    55     0    50    45   360  2190  2750  3800   130   100   300
 é    55     0    50    45   540  1950  2650  3800    70   130   200
 a    55     0    50    45   700  1350  2600  3700   110   130   130
 ó    55     0    50    45   530   900  2650  3450    90   130   130
 o    55     0    50    45   350   750  2650  3450    80    50   150
 u    55     0    50    45   280   670  2600  3400    65   110   200
 I    55     0    50    50   330  2350  3150  3800    80   150   150   310
 ë    55     0    50    50   460  2250  2750  3900   180   130   300   370
 ä    55     0    50    50   650  1450  2700  3800   210   200   130   440
 ö    55     0    50    50   450   750  2650  3450   130    90   150   360
 ü    55     0    50    50   380   670  2650  3450   125   180   200   330
 A    50     0    50    40   650  1450  2600  3700   110   130   130   270     #     #     #     #     #     #     #     #    80
 y    50     0    50    40   310  2250  2950  3700    90   110   200     #     #     #     #     #     #     #     #     #    70
 w    50     0    50    40   320   750  2600  3450    75    80   170
 ỹ    55     0    50    50   330  2350  3150  3800    80   150   150   310
 Ũ    55     0    50    50   380   670  2650  3450   125   180   200   330
 z    40    40    40    40   250  1250  2700  4000    70    60   180   270     0     0    30    50    40     0
 j    40    40    40    40   350  1650  2600  3400    70    80   150     #    30    40    50    40    30
 v    40    40    40    40   300  1000  2000  3400    60    90   120     #     0     0     0     0     0    50
 s     0    50     0     0   250  1250  2700  4000   200    80   200     #     0     0    30    50    50     0     #     #   130
 x     0    50     0     0   450  2000  2700  3400   120   100   250     #    30    40    50    40    30
 f     0    30     0     0   300  1000  2000  3400   200   120   150     #     0     0     0     0     0    50
 R    30    30    50    40     /     /     /     /    70    80   150     #    30    40    50    40    30     0     #     #   120
 X     0    40     0     0   450  2000  2700  3400   120   100   250     #    30    40    50    40    30     0     #     #    40
 J    30    35    35    40   350  1650  2600  3400    70    80   150
 S     0    50     0     0   250  1250  2700  4000   200    80   200     #     0     0    30    50    50     0     #     #    80
 p     0    50     0     0   200  1000  2000  3400   300   150   220     #     0     0     0     0     0    60     #     #   100
 b    40    50    45    45   200  1000  2000  3400    60   110   130     #     0     0     0     0     0    60
 t     0    30     0     0   200  1700  2600  3600   300   120   150     #     0    30    45    55    60     0
 d    40    30    45    45   200  1700  2600  3600    60   100   170     #     0    45    55    60    60     0     #     #    85
 k     0    40     0     0   200  2100  2800  3600   250   160   330     #     0    55    50    45    45
 g    40    40    45    45   200  2100  2800  3600    60   150   280
 r    30     0    30    40   400  1700  2600  3600    60   100   170     #    45    35    45    45    45     0     #     #    30
 P    30     0    30    40   400  1700  2600  3600    60   100   170     #    45    35    45    45    45     0     #     #    25
 l    55     0    40    40   250  1400  2700  3600    40   200   150     #     0     0     0     0     0     0     #     #   120
 L    55     0    40    40   250  1400  2700  3600    40   200   150     #     0     0     0     0     0     0     #     #    40
 m    40     0    40    50   480  1100  2200  3400    40   200   250   440     0     0     0     0     0     0     #     #    70
 n    40     0    40    50   480  1700  2600  3500    40   300   300   440     0     0     0     0     0     0     #     #    60
 M    40     0    40    50   480  1100  2200  3400    40   200   250   440     0     0     0     0     0     0     #     #    25
 N    40     0    40    50   480  1700  2600  3500    40   300   300   440     0     0     0     0     0     0     #     #    25
 Ñ    40     0    40    50   480  1900  2700  3500    40   300   300   440     0     0     0     0     0     0     #     #    25
 ñ    45     0    40    50   370  2350  3150  3800    80   150   150   320
FIM
```

```
% Variáveis
% Nome          Valor     Comentário
T$V-FONS        25        Metade da transição silêncio-vogal para AV, AVS e AH
T$F-FONT        20        Metade da transição silêncio-fricativa para as fontes
T$L-FONT        15        Metade da transição silêncio-líquida para as fontes
T$P-FONT        10        Metade da transição silêncio-oclusiva para as fontes
T$r-FBn         10        Metade da transição silêncio-r para Fn e Bn
T$r-FONT         5        Metade da transição silêncio-r para as fontes
TV$-FONS        35        Metade da transição vogal-silêncio para AV, AVS e AH
TVV-FBn         35        Metade da transição vogal-vogal para Fn e Bn
TVF-FBn         30        Metade da transição vogal-fricativa para Fn e Bn
TVF-FONT        10        Metade da transição vogal-fricativa para as fontes
TVL-FBn         25        Metade da transição vogal-líquida para Fn e Bn
TVL-FONT         5        Metade da transição vogal-líquida para as fontes
TVP-FBn         30        Metade da transição vogal-oclusiva para Fn e Bn
TVP-FONT        10        Metade da transição vogal-oclusiva para as fontes
TVr-FBn         10        Metade da transição vogal-r para Fn e Bn
TVr-FONT         5        Metade da transição vogal-r para as fontes
TF$-FONT        30        Metade da transição fricativa-silêncio para as fontes
TFV-FBn         30        Metade da transição fricativa-vogal para Fn e Bn
TFV-FONT        20        Metade da transição fricativa-vogal para as fontes
TFL-FBn         15        Metade da transição fricativa-líquida para Fn e Bn
TFL-FONT        15        Metade da transição fricativa-líquida para fontes
TFP-FBn         30        Metade da transição fricativa-oclusiva para Fn e Bn
TFP-FONT        10        Metade da transição fricativa-oclusiva para as fontes
TFr-FBn         10        Metade da transição fricativa-r para Fn e Bn
TFr-FONT         5        Metade da transição fricativa-r para as fontes
DUPf-AF         20        Diferença entre o último e o primeiro valor de AF no fone 'f'
TL$-FONT        15        Metade da transição líquida-silêncio para as fontes
TLV-FBn         25        Metade da transição líquida-vogal para Fn e Bn
TLV-FONT         5        Metade da transição líquida-vogal para as fontes
TLF-FBn         30        Metade da transição líquida-fricativa para Fn e Bn
TLF-FONT        10        Metade da transição líquida-fricativa para as fontes
TLP-FBn         30        Metade da transição líquida-oclusiva para Fn e Bn
TLP-FONT        10        Metade da transição líquida-oclusiva para as fontes
TPV-FBn         20        Metade da transição oclusiva-vogal para Fn e Bn
TPX-FONT        10        Transição entre oclusiva e qualquer fone para as fontes
TPF-FBn         20        Metade da transição oclusiva-fricativa para Fn e Bn
TPL-FBn         20        Metade da transição oclusiva-líquida para Fn e Bn
TPr-FBn         10        Transição oclusiva-r para Fn e Bn
Tr$-FONT         5        Metade da transição líquida-silencio para as fontes
TrV-FBn         10        Metade da transição líquida-vogal para Fn e Bn
TrV-FONT         5        Metade da transição líquida-vogal para as fontes
TrF-FBn         10        Metade da transição líquida-fricativa para Fn e Bn
TrF-FONT         5        Metade da transição líquida-fricativa para as fontes
TrP-FBn         10        Metade da transição líquida-oclusiva para Fn e Bn
TrP-FONT         5        Metade da transição líquida-oclusiva para as fontes
TrL-FBn         10        Metade da transição líquida-oclusiva para Fn e Bn
TrL-FONT         5        Metade da transição líquida-oclusiva para as fontes
DTRF1            5        Tempo de subida de AF na explosão
DTRF2            5        Tempo de descida de AF na explosão
DRF             10        Período em que há ruído de fricação na oclusiva
FIM
```

Referências bibliográficas

Allen, J., Hunnicut, M.S. e Klatt, D.H. *From text to speech: the MITalk System.* Cambridge University Press, 1987.

Aquino, P.A. *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, Unicamp, novembro de 1997.

Comerford, R., Makhoul, J. e Schwartz, R. *The voice of the computer is heard in the land (and it listens too).* IEEE Spectrum, pp. 39-47, dezembro de 1997.

Egashira, F. *Síntese de voz a partir de texto para a língua portuguesa.* Dissertação de mestrado. Faculdade de Engenharia Elétrica, Unicamp, julho de 1992.

Egashira, F. e Violaro, F. *Conversor texto-fala para a língua portuguesa.* 13° Simpósio Brasileiro de Telecomunicações. Campinas/Águas de Lindóia, SP, pp. 71-76, setembro de 1995.

Emerard, F., Mortamet, L. e Cozannet, A. *Prosodic processing in a text-to-speech synthesis system using a database and learning procedures. In: Talking machines: theories, models and designs.* Bailly, G., Benoît, C. e Sawallis, T.R. (eds.). Elsevier Science Publishers, 1992.

Fant, C.G.M. *Acoustic theory of speech production.* The Hague: Mouton, 1960.

Gomes, L.C.T., Nagle, E.J. e Chiquito, J.G. *Interface entre processamento de texto e de sinal para a síntese de fala por regras.* 14° Simpósio Brasileiro de Telecomunicações. Curitiba, PR, pp. 355-360, julho de 1996.

Gomes, L.C.T., Nagle, E.J. e Chiquito, J.G. *Text-to-speech conversion system for Brazilian Portuguese using a formant-based synthesis technique.* 4th International Telecommunications Symposium. São Paulo, SP, agosto de 1998 (no prelo).

Hertz, S. *From text to speech with SRS.* Journal of Acoustical Society of America, vol. 72, pp. 1155-1170, 1982.

Kay Elemetrics Corp. *CSL (Computerized Speech Lab.) modelo 4300B, versão 5.X – manual de instruções.* Kay Elemetrics, fevereiro de 1994.

Klatt, D.H. *Software for a cascade/parallel formant synthesizer.* Journal of Acoustical Society of America, vol. 67, no. 3, pp. 971-995, março de 1980.

Klatt, D.H. *Review of text-to-speech conversion for English.* Journal of Acoustical Society of America, vol. 82, nº 3, pp. 737-893, setembro de 1987.

Klatt, D.H. e Klatt, L.C. *Analysis, synthesis and perception of voice quality variations among female and male talkers.* Journal of Acoustical Society of America, vol. 87, nº 2, pp. 820-857, fevereiro de 1990.

Lima, R. *Gramática normativa da Língua Portuguesa.* José Olympio Editora, 1991.

Nagle, E.J. *Síntese inicial de vogais e fricativas.* Relatório técnico. Faculdade de Engenharia Elétrica da Unicamp, agosto de 1991.

Nagle, E.J. e Chiquito, J.G. *Síntese de sinais de fala usando o sintetizador de formantes de Klatt.* 11° Simpósio Brasileiro de Telecomunicações. Natal, RN, pp. 718-723, setembro de 1993.

Silva, C.H. *Modelamento prosódico para conversão texto-fala do português falado no Brasil.* Dissertação de mestrado. Faculdade de Engenharia Elétrica, Unicamp, novembro de 1995.

Silva, C.H. e Violaro, F. *Determinação de fronteiras entre constituintes prosódicos para conversão texto-fala do português falado no Brasil.* 14° Simpósio Brasileiro de Telecomunicações. Curitiba, PR, pp. 349-354, julho de 1996.

STR (Speech Technology Research Ltd.) *CSL 4300 Hardware Interface Library – manual de referência.* STR, 1994.

Traber, C. *F_0 generation with a database of natural F_0 patterns and with a neural network. In: Talking machines: theories, models and designs.* Bailly, G., Benoît, C. e Sawallis, T.R. (eds.). Elsevier Science Publishers, 1992.

Violaro, F., Barbosa, P.A., Albano, E.C. e Françozo, E. *Um conversor texto-fala para o português brasileiro com processamento lingüístico de alta qualidade.* 14° Simpósio Brasileiro de Telecomunicações. Curitiba, PR, pp. 361-366, julho de 1996.