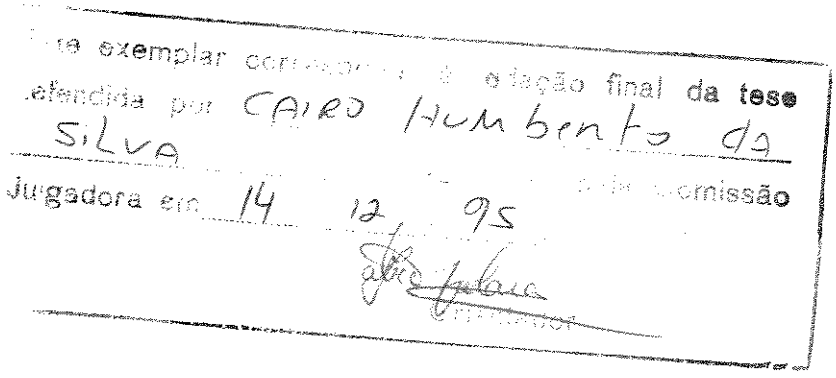
UNIVERSIDADE ESTADUAL DE CAMPINAS
FACULDADE DE ENGENHARIA ELÉTRICA

Modelamento Prosódico Para Conversão Texto-Fala do Português Falado no Brasil

Cairo Humberto da Silva

Orientador: Prof. Dr. Fábio Violaro



Campinas
Dezembro de 1995

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

Si38m

Silva, Cairo Humberto da
Modelamento prosódico para conversão texto-fala do português falado no Brasil / Cairo Humberto da Silva.-- Campinas, SP: [s.n.], 1995.

Orientador: Fábio Violaro.
Dissertação (mestrado) - Universidade Estadual de Campinas, Faculdade de Engenharia Elétrica.

1. Análise prosódica (Linguística). 2. Síntese da voz 3. Fonética. I. Violaro, Fábio. II. Universidade Estadual de Campinas. Faculdade de Engenharia Elétrica. III. Título.

Agradecimentos

Ao **Senhor** que abençoa o meu trabalho e derrama graças sobre minha vida.

Aos meus pais por tudo que fizeram por mim.

A Sandra por me amar apesar da distância e de mim mesmo.

Ao orientador Prof. Dr. Fábio Violaro pela oportunidade da realização deste estudo, pelo apoio e pela amizade.

Ao Centro de Pesquisas e Desenvolvimento da TELEBRAS, pelo suporte material, e aos colegas de trabalho, em especial a Fernando O. Runstein, pelo apoio e pela amizade.

Aos queridos amigos que muito me incentivaram e a quem eu amo.

À Profa. Dra. Eleonora C. Albano pela formação básica na área de fonética acústica, à Profa. Dra. Sandra Madureira pela valiosa colaboração e a todo o grupo do Laboratório de Fonética Acústica e Psicolingüística Experimental do IEL-UNICAMP.

Sumário

Este trabalho descreve um modelo de tratamento prosódico aplicado a um sistema de conversão texto-fala para o português falado no Brasil.

O modelo prosódico é composto de um modelo duracional e de um modelo de freqüência fundamental.

Com base em dados extraídos a partir da análise de realizações de fala natural (frases ditas por pessoas), é proposto um modelo que controla a duração fonética e gera curvas de freqüência fundamental para sentenças declarativas neutras, isto é, sentenças ditas sem qualquer ênfase ou conteúdo emotivo.

O modelo prosódico foi testado usando-se um sintetizador por concatenação que emprega a técnica PSOLA.

Índice

Lista de Figuras vii

CAPÍTULO I - INTRODUÇÃO

1.1 Conversor texto-fala 2

1.2 Sistemas de resposta por voz 3

1.3 Aplicações de sistemas de conversão texto-fala 4

1.4 Interdisciplinaridade 6

1.5 Resumo do trabalho 6

1.6 Estrutura da tese 7

CAPÍTULO II - SÍNTESE

2.1 Categorias de síntese 8

2.2 Técnica PSOLA 13

CAPÍTULO III - AMBIENTE DE TRABALHO

3.1 Explicação geral 18

3.2 Corpus 18

3.3 Ferramenta para análise prosódica 20

3.4 Processo de análise 24

3.5 Conversor texto-fala 25

CAPÍTULO IV - CONSIDERAÇÕES LINGÜÍSTICAS

4.1 Trato vocal 28

4.1.1 Ospulmões 29

4.1.2 Cordas vocais 30

4.2 Fonologia 30

4.2.1 Definição 30

4.2.2 Fonemas 30
4.2.3 Fones 31
4.3 Fonética 32
4.3.1 Definição 32
4.3.2 Fonética articulatória 32
4.3.2.1 Modo de articulação 32
4.3.2.2 Ponto de articulação 33
4.3.3 Fonética acústica 34
4.3.3.1 Caracterização fonêmica do ponto de vista acústico 35
4.4 Transcrição ortográfico-fonética 36
4.5 Coarticulação 39
4.6 Prosódia 39
4.6.1 Definição 39
4.6.2 Funções da prosódia 40
4.6.3 Freqüência fundamental (F0) 42
4.6.4 Duração 42
4.6.5 Energia 43
4.6.6 Acentuação e ritmo 43

CAPÍTULO V - MARCAÇÃO DE FRONTEIRAS PROSÓDICAS
5.1 Estrutura prosódica 45
5.2 Relação entre estrutura sintática e estrutura prosódica simplificada 47
5.3 Derivação da estrutura prosódica simplificada para modelamento prosódico na conversão texto-fala. 48

CAPÍTULO VI - MODELO DE DURAÇÃO
6.1 Aspectos gerais do comportamento duracional dos segmentos 50
6.2 Considerações sobre o modelo duracional construído 51
6.3 Descrição do modelo duracional 53

CAPÍTULO VII - MODELO ENTOACIONAL

7.1 Porque é complicada a tarefa de modelar entonação para um sistema de conversão texto-fala 59

7.2 Tendências gerais das curvas entoacionais de sentenças declarativas 61

7.3 Considerações sobre o modelo entoacional construído 61

7.4 Descrição do modelo entoacional 64

CAPÍTULO VIII - RESULTADOS OBTIDOS

8.1 Avaliação 70

CAPÍTULO IX - CONCLUSÕES

9.1 Considerações sobre o trabalho realizado 73

9.2 Propostas para futuros trabalhos 75

APÊNDICE 76

REFERÊNCIAS BIBLIOGRÁFICAS 86

Lista de Figuras

CAPÍTULO I - INTRODUÇÃO

FIGURA 1.1 - Diagrama básico de um conversor texto-fala 1

CAPÍTULO II - SÍNTESE

FIGURA 2.1 - Diagrama de blocos de um sintetizador por formantes (O'Shaughnessy, pag. 394) 9

FIGURA 2.2 - Diagrama básico de um conversor texto-fala por concatenação 11

FIGURA 2.3 - Aumento de freqüência fundamental pela técnica PSOLA 15

FIGURA 2.4 - Diminuição de frequência fundamental pela técnica PSOLA 15

FIGURA 2.5 - Aumento de duração pela técnica PSOLA 16

FIGURA 2.6 - Diminuição de duração pela técnica PSOLA 16

CAPÍTULO III - AMBIENTE DE TRABALHO

FIGURA 3.1 - Tela do analisador prosódico sendo utilizado para analisar o enunciado referente à frase: "_ Curva perigosa" 24

FIGURA 3.2 - Esquema de entradas e saídas do conversor texto-fala 25

FIGURA 3.3 - Arquivo prosódico gerado pela síntese do texto " O preço da tarifa telefônica foi reduzido." 27

CAPÍTULO IV - CONSIDERAÇÕES LINGÜÍSTICAS

FIGURA 4.1 - Vista da seção transversal do trato vocal (1) lábio inferior, (2) incisivo inferior, (3) ponta da língua, (4) dorso, (5) frente, (6) costas, (7) raiz, (8) lábio superior, (9) incisivo superior, (10) alvéolo, (11) palato duro, (12) velum, (13) úvula, (14) faringe, (15) laringe, (16) cordas vocais e glote. 29

FIGURA 4.2 - Alfabeto Fonético Internacional (IPA). 38

CAPÍTULO VII - ENTOACIONAL

FIGURA 7.1 - Árvore de modelamento entoacional 62

FIGURA 7.2 - Esquema do modelamento entoacional para o enunciado de: "_São Paulo". 64

CAPÍTULO VIII - RESULTADOS OBTIDOS

FIGURA 8.1 - Gráficos de F0 para enunciados relativos à sentença "_É necessário que o convênio permita o intercâmbio". (a) F0 natural, (b) F0 intrínseca e (c) F0 modelada 72

Capítulo 1

Introdução

Neste capítulo procura-se apresentar o conceito e a estrutura de um sistema de conversão texto-fala, algumas de suas aplicações, o caráter interdisciplinar do presente trabalho, seus objetivos e um breve resumo dos capítulos restantes.

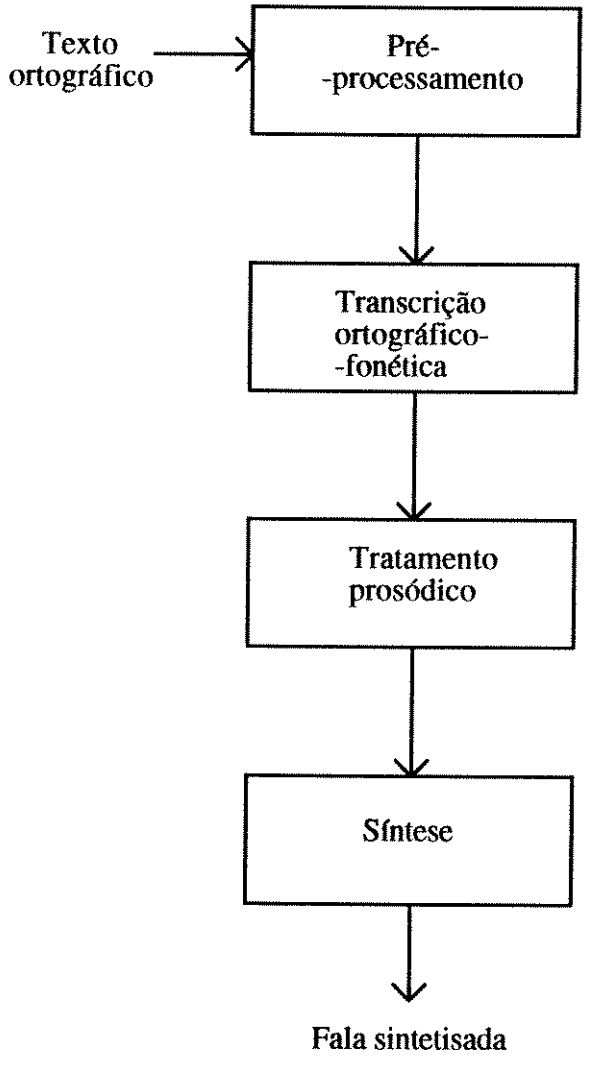


Fig. 1.1 - Diagrama básico de um conversor texto-fala

1. Conversor texto-fala

Um conversor texto-fala é um sistema que aceita como entrada um texto em sua forma ortográfica (isto é, como usualmente se escreve) e produz, como saída, a fala na forma de onda sonora. A figura 1.1 mostra os elementos básicos que compõem o sistema de produção de fala a partir de texto. Abaixo, explica-se a função de cada módulo.

Pré-processamento

Este módulo recebe como entrada o texto em sua forma convencional, contendo símbolos, abreviações e algarismos. Sua função é converter estes símbolos, abreviações e algarismos, em sua forma extensa. Por exemplo, a abreviação "Sr." deve ser traduzida para "senhor"; o número '13142' deve ser transcrito para "treze mil trezentos e quarenta e dois".

Transcrição ortográfico-fonética

Textos ortográficos representam uma maneira eficiente e cômoda de comunicação por escrito. Isto ocorre, em parte, porque o alfabeto empregado possui um número bastante restrito de símbolos (letras). Ao ler-se um texto, no entanto, as palavras apresentam uma variedade muito maior de tipos de sons (fonemas). Acontece que não existe uma correspondência "um a um" entre as letras e os fonemas. Como o objetivo do conversor texto-fala é produzir fala, torna-se necessário representar as palavras de uma forma mais próxima da maneira como são pronunciadas, ou seja, é preciso transcrever as palavras para uma forma gráfica em que os fonemas que as constituem sejam explicitados. Isto significa transcrever a forma ortográfica para a forma fonética. A palavra "casa", por exemplo, possui o "s" com som de "z" e o "c" com som de "k". Este fato é evidenciado ao realizar-se sua transcrição ortográfico-fonética que resulta em KAZA.

Tratamento prosódico

Para que a fala produzida seja compreensível e semelhante à que uma pessoa produziria (apresentando, portanto, naturalidade), é necessário dar um tratamento aos fonemas. Este tratamento consiste em alterar quantitativamente algumas de suas características. Por exemplo,

pode-se fazer com que um fone (realização física de um fonema ao ser pronunciado) tenha uma duração conveniente ao seu contexto na palavra em que ocorre e um apropriado perfil de entonação.

Síntese

De posse da especificação dos fonemas e de seus atributos, parte-se para o modelamento da produção de fala do ponto de vista acústico, isto é, tendo recebido uma descrição fonética, já com tratamento prosódico, o próximo passo é gerar uma onda sonora (ou uma representação desta) correspondente. Esse procedimento é denominado síntese.

2. Sistemas de resposta por voz

É necessário, neste momento, fazer uma distinção entre conversor texto-fala e sistema de resposta por voz. Por conversor texto-fala entende-se um sistema capaz de produzir fala a partir de qualquer texto ortográfico que receba como entrada. Um sistema de resposta por voz não é capaz de produzir fala a partir de texto irrestrito, mas apenas a partir de um conjunto de frases.

A forma mais comum e simples de produzir fala sintética é através de sistemas de resposta por voz. Basicamente, o que estes sistemas fazem nada mais é do que concatenar palavras, ou seqüências de palavras, previamente gravadas por um locutor. Tais sistemas ficam, portanto, limitados a produzir apenas aquelas mensagens que podem ser "montadas" a partir do vocabulário gravado.

Os sistemas de resposta por voz são adequados para várias aplicações tais como brinquedos, mensagens de advertência dadas por máquinas, saldo bancário automático por telefone, etc. Entretanto, tais sistemas possuem a inconveniência de, toda vez que uma nova palavra precisar ser incluída no vocabulário, ser necessário convocar o mesmo locutor, que pronunciou as palavras já gravadas, para uma nova seção de gravação. Isto acarreta três problemas. Primeiro, o locutor pode não estar disponível. Segundo, realizar uma nova seção de gravação requer tempo. Terceiro, os custos financeiros com pagamento de locutor e estúdio de gravação não são baixos. Além disso, apesar de não requerer tanto processamento quanto os sistemas de conversão texto-fala, as sentenças obtidas através de um sistema de resposta por

voz apresentam grande descontinuidade nas junções entre palavras e , portanto, pouca naturalidade.

Devido aos problemas expostos acima, e aos avanços tecnológicos da conversão texto-fala, os sistemas de resposta por voz tendem a cair em desuso.

3. Aplicações de sistemas de conversão texto-fala

Basicamente, as principais aplicações de sistemas de conversão texto-fala referem-se a situações onde a informação que se queira difundir não pode ser prevista ou as atualizações na informação são tão freqüentes que se torne difícil gravar mensagens ditas por um falante. Outra situação bastante propícia à aplicação de um sistema de conversão texto-fala ocorre quando o volume de informações é muito grande. Como exemplo de aplicação onde as informações variam com grande freqüência, e não podem ser previstas, pode-se citar o caso de um sistema de conversão texto-fala que leia um jornal diário para um deficiente visual. Uma aplicação onde o volume de informação seria muito grande poderia ser a consulta à lista telefônica de Campinas.

As aplicações utilizando a rede telefônica são muito atrativas. Isto porque, na telefonia, a comunicação se dá exclusivamente por meio da fala. Assim, informações de banco de dados podem ser enviadas aos assinantes através da fala. Contudo, para que um sistema de conversão possa ser utilizado em todo seu potencial, é mister que haja um sistema de reconhecimento de fala funcionando como meio de entrada de informações. Esta necessidade representa uma limitação que vem sendo rapidamente vencida pelos progressos que estão sendo feitos na área do reconhecimento de fala.

São também muito atrativas as aplicações em informática. A comunicação homem-computador, por meio da fala, representa uma grande comodidade ao usuário de sistemas computacionais. Tal comodidade pode resultar em um aumento de produtividade para usuários uma vez que seja possível haver um diálogo homem-máquina.

A seguir, citam-se alguns exemplos de aplicações:

- Obtenção, pelo telefone, de informações tais como eventos esportivos ou culturais, feiras e exposições, programação de teatro e cinema.

Informações destes tipos precisam ser atualizadas freqüentemente. Como é muito mais cômodo, e rápido, fazer a atualização de um texto escrito do que realizar uma gravação, o recurso da conversão texto-fala seria bastante útil. Além disso, para realizar gravações é necessária uma infraestrutura (estúdio de gravação, técnicos, etc.) cujos custos se somam ao da remuneração de um locutor. Logo, a conversão texto-fala pode ser a opção mais barata.

- Máquina de leitura para cegos

Uma aplicação bastante humanitária da conversão texto-fala está relacionada ao auxílio para deficientes visuais. Esforços têm sido feitos para desenvolver sistemas que possam reproduzir de maneira falada um determinado texto disponível na forma impressa. Um sistema deste tipo consiste da associação de um "scanner", um OCR ("Optical Character Recognition") e um sistema de conversão texto-fala. Através de um sistema de conversão texto-fala, um deficiente visual poderia ter acesso ao lazer proporcionado por obras literárias. Também poderia estudar em livros didáticos que não possuem edição em Braile.

Como a interação com computadores no trabalho é cada vez mais necessária para profissionais das mais diversas áreas, um sistema de conversão texto-fala possibilitaria também, ao profissional com deficiência visual, o acesso a recursos computacionais. Isso facilitaria o ingresso de deficientes visuais no mercado de trabalho.

- Saldo bancário por telefone.

Como os vocabulários para tais aplicações são limitados, atualmente são empregados sistemas de resposta por voz para produzir mensagens de saldo bancário. Entretanto, a má qualidade da fala obtida com o emprego de tais sistemas e os problemas com acréscimo de novas palavras ao vocabulário, tornam insatisfatório o uso de sistemas de resposta por voz para esta aplicação. Um sistema de saldo pode ser implementado a partir de um conversor texto-fala, sendo que este deve possuir um aprimorado modelo de tratamento prosódico para garantir a naturalidade da fala produzida.

- Auxílio de fala a deficientes vocais .

Um sistema de conversão texto-fala pode auxiliar pessoas impossibilitadas ou com grande dificuldade em falar. Neste tipo de aplicação, um aspecto muito relevante é a maneira pela qual a pessoa passará ao sistema as informações sobre o que deseja falar. É preciso prover uma maneira interativa de formar mensagens de entrada que seja bastante simples e rápida,

evitando que a operação do sistema se torne cansativa e entediante para o usuário. Além disso, o sistema deve ser portátil.

4. Interdisciplinaridade

A construção de um conversor texto-fala exige conhecimentos tanto do campo da lingüística quanto da área de engenharia. Toda parte de processamento de sinais, necessária à etapa de síntese, é normalmente objeto de estudo da Engenharia Elétrica. Já a transcrição ortográfico-fonética e o estudo prosódico exigem conhecimentos que normalmente ficam restritos ao campo da Lingüística.

Sendo o conversor texto-fala um programa de computador (embora existam também elementos físicos para a conversão digital/analógica, amplificação e produção de som), tornam-se necessários, também, conhecimentos de engenharia de "software" para construí-lo. Note-se que um programa desta natureza possui milhares de linhas de código.

5. Resumo do trabalho

O principal objetivo do presente trabalho foi iniciar a construção de um modelo de tratamento prosódico aplicado a um sistema de conversão texto-fala para o português falado no Brasil.

Sendo este um trabalho inicial na área de tratamento prosódico, também tem como objetivo a aquisição de experiência na área. Além disso, buscou-se investir na criação de um ferramental para análises acústicas com o objetivo de auxiliar futuros desenvolvimentos.

O presente trabalho consiste basicamente na análise de realizações de fala natural (frases ditas por pessoas), criação de regras que descrevem o tratamento prosódico a que as pessoas submetem suas próprias falas, uso destas regras para construção de um módulo prosódico aplicado à conversão texto-fala, e emprego do sistema de conversão texto-fala para testar e aprimorar estas regras.

6. Estrutura da tese

No capítulo 2 é feita uma breve descrição das categorias básicas em que são classificados os sistemas de conversão texto-fala. Procurou-se descrever com maior riqueza de detalhes o método de síntese por concatenação, uma vez que o presente trabalho foi desenvolvido utilizando um sistema de síntese por concatenação. Este sistema baseia-se na técnica PSOLA que também é descrita sucintamente.

O capítulo 3 descreve o ambiente de trabalho criado. São apresentados os critérios que levaram à elaboração de um conjunto (corpus) de frases que foram pronunciadas por um falante e gravadas. Também descreve as funções desempenhadas por uma ferramenta de "software" desenvolvida para a análise prosódica dos enunciados previamente gravados a partir da leitura das frases constituintes do corpus.

No capítulo 4 é feito um breve resumo dos conhecimentos lingüísticos empregados na produção deste trabalho. É dada uma pequena explicação a respeito da anatomia do sistema que os seres humanos usam para produção da fala. Também são apresentadas definições de Fonologia, fonemas, alofones e fones. Em seguida são apresentadas definições de Fonética e de suas subdivisões: Fonética Articulatória e Fonética Acústica. São definidos modo e ponto de articulação e são apresentadas classificações dos fonemas com relação a eles. Mostra-se uma breve caracterização fonêmica do ponto de vista acústico e uma explanação sobre transcrição fonética. Em seguida define-se coarticulação. Além disso, são apresentados os conceitos de prosódia e parâmetros prosódicos. Também são citadas as funções que a prosódia desempenha na fala.

No capítulo 5 é apresentado o conceito de estrutura prosódica e sua importância para o tratamento prosódico. Também é descrito neste capítulo como foi tratado, no presente trabalho, o problema de determinar a estrutura prosódica de uma sentença escrita para a sua posterior conversão em fala.

No capítulo 6 é descrito o desenvolvimento do modelo de controle duracional.

O capítulo 7 apresenta o modelo de controle entoacional.

No capítulo 8 é apresentada uma avaliação dos resultados obtidos por este trabalho.

No capítulo 9 é apresentada a conclusão deste trabalho e são feitas sugestões para futuros desenvolvimentos.

Capítulo 2

Síntese

Neste capítulo é feita uma breve descrição das categorias básicas em que são classificados os sistemas de conversão texto-fala. Procurou-se também descrever, sucintamente, a técnica PSOLA de síntese por concatenação.

2.1 Categorias de síntese

Conforme foi dito no capítulo anterior, o módulo de síntese recebe uma descrição fonética, já com informação prosódica, e gera um sinal de fala correspondente a esta entrada.

O presente trabalho foi desenvolvido usando um sintetizador concatenativo que emprega a técnica PSOLA. Abaixo, são dadas descrições sucintas dos principais métodos de síntese.

Existem três grandes categorias de métodos para realizar a síntese sonora a partir das informações fonéticas e prosódicas: a síntese por regras, a síntese articulatória e a síntese por concatenação de unidades acústicas.

- Síntese por regras

  Baseia-se num modelamento paramétrico do sinal de voz e num conjunto de regras que regem a evolução temporal dos parâmetros.

  A maior dificuldade na síntese por regras é a obtenção dos parâmetros. É necessário estudar profundamente as propriedades espectrais da fala natural e determinar aquelas que são perceptualmente relevantes no domínio acústico.

  A técnica da síntese por formantes é a mais adequada para a implementação dos sintetizadores por regras [1]. Na síntese por formantes, a fala é obtida fornecendo-se as freqüências formantes e as respectivas larguras de banda. Logo, os parâmetros de controle

do sintetizador de formantes são obtidos diretamente por meio de análises acústicas. A figura 2.1 mostra um diagrama simplificado de um sintetizador por formantes.

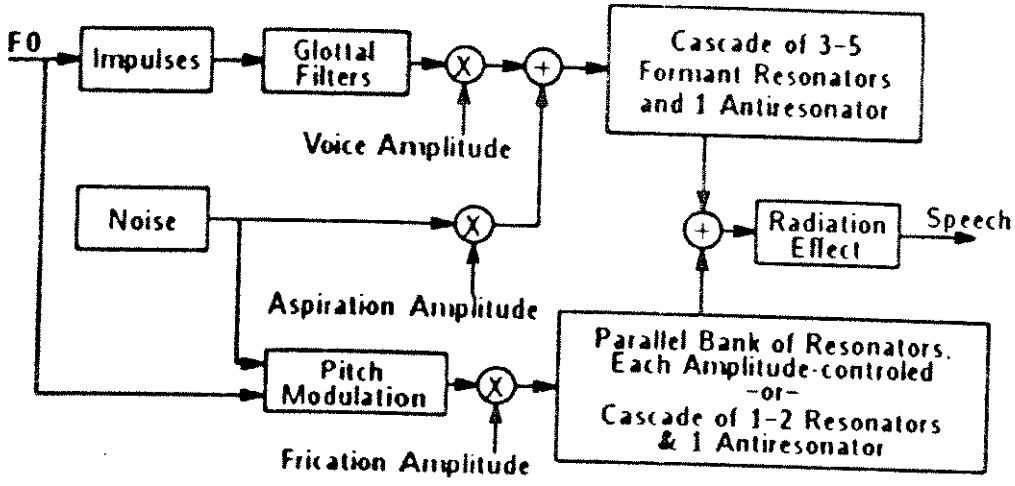


Fig. 2.1 - Diagrama de blocos de um sintetizador por formantes (ref. [2], pag. 394)

- Síntese articulatória

Para produzir fala uma pessoa pensa uma mensagem e envia comandos aos músculos do seu trato vocal que causam a movimentação dos articuladores. Essa movimentação de articuladores faz com que a formato do trato vocal mude. Os modelos articulatórios tentam modelar as configurações que o trato vocal de um falante vai assumindo ao longo do tempo enquanto ele está falando. O objetivo de tais modelos é produzir a fala de maneira análoga ao processo natural, isto é, de maneira semelhante à empregada pelos seres humanos para produzí-la. Para isso, são feitas observações do trato vocal em movimento, enquanto o falante produz enunciados, e coletados dados a respeito da evolução dos articuladores no espaço ao longo do tempo. Tais dados fornecem parâmetros para que, no processo de síntese, um conjunto de regras possa ditar os movimentos sucessivos do trato vocal de modo a que seja produzida a forma de onda de fala. Na síntese coarticulatória o trato vocal é simulado por uma combinação de filtros.

Embora os modelos articulatórios possuam grande potencial e sejam, do ponto de vista teórico, a maneira mais razoável de sintetizar fala, eles apresentam uma dificuldade de

obtenção dos dados de evolução dos articuladores. Os dados de fala para sintetizadores articulatórios geralmente são obtidos com o uso de raios-X sobre tratos vocais durante a produção de enunciados simples. Por essa razão, estes modelos ainda permanecem retritos a laboratórios. Além disso, os modelos articulatórios possuem grande complexidade computacional, o que dificulta ainda mais sua utilização em sistemas de síntese para fins comerciais.

- Síntese por concatenação de unidades acústicas

Sintetizadores por concatenação realizam a conversão de um texto em fala por meio da concatenação de unidades básicas previamente gravadas.

Inicialmente deve-se segmentar trechos de fala natural, codificá-los de acordo com a técnica de síntese adotada e, com eles, construir um dicionário de unidades acústicas. Este dicionário deve conter as unidades necessárias para, através da concatenação das mesmas, gerar o sinal de fala correspondente ao texto que se queira sintetisar. Assim, se o objetivo é a síntese de fala a partir de texto irrestrito para a língua portuguesa, o dicionário deve conter as unidades suficientes para produzir qualquer enunciado do idioma.

A maior dificuldade desta técnica é prover transições entre as unidades. Isto ocorre porque descontinuidades espectrais nas junções produzem efeitos perceptuais indesejados. Por esta razão, é necessário prover algum mecanismo de suavização espectral para as junções de unidades.

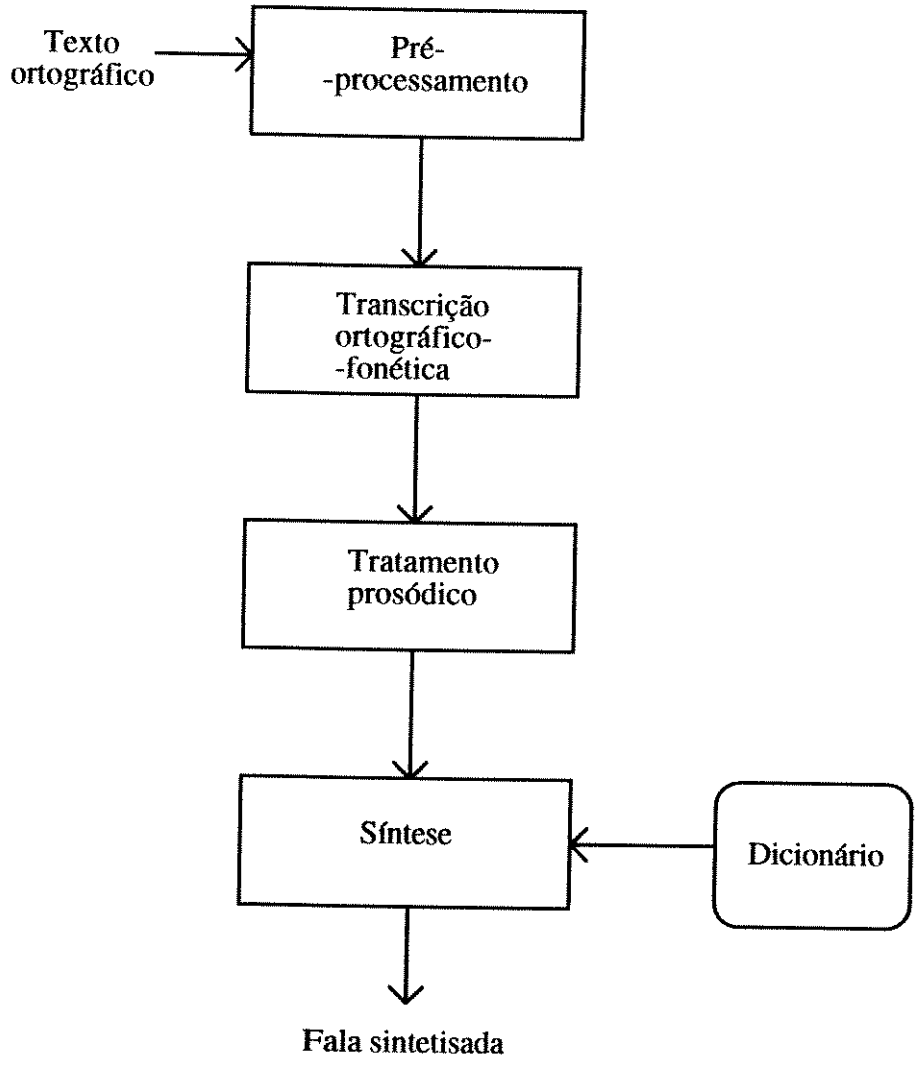


Fig. 2.2 - Diagrama básico de um conversor texto-fala por concatenação

Quanto maior em número de fones for a unidade, menor será o número de junções na fala produzida e, portanto, melhor será sua qualidade. Porém, unidades grandes implicam na necessidade de um enorme dicionário de unidades para prover todas as combinações possíveis. Num caso limite em que as unidades básicas fossem as palavras, o dicionário deveria conter todas as palavras da língua empregada. É preciso, então, buscar uma solução de compromisso entre o problema das transições entre unidades e o inconveniente de se ter um dicionário cujo tamanho torne difícil sua contrução.

A escolha das unidades acústicas para constituição do dicionário deve levar em conta o grau de dificuldade em prover transições espectralmente suaves para as junções de tais unidades. Assim, construir um dicionário de unidades acústicas para síntese por

concatenação requer uma avaliação cuidadosa de cada unidade escolhida e de suas possíveis junções com outras unidades do dicionário quando do processo de síntese.

As unidades acústicas mais usadas para composição dos dicionários dos sistemas de síntese por concatenação (a partir de textos irrestritos) são: sílabas, demissílabas e difones[2]. Uma sílaba consiste de um núcleo, que pode ser uma vogal ou um ditongo, associado a algumas consoantes vizinhas. Demissílabas são unidades de fala obtidas por meio da divisão de sílabas ao meio, sendo que esta divisão ocorre sobre a vogal (ou ditongo) onde os efeitos de coarticulação (veja o capítulo 4) são mínimos.

No sistema de síntese usado no desenvolvimento do presente trabalho, a unidade básica utilizada foi o **difone**. Ele é definido como sendo o segmento de voz que inicia no centro da região espectralmente estável de um fone e termina no centro da região estável do próximo fone, contendo a transição completa entre os dois fones [3]. Por exemplo, o difone "KK_AA" (os símbolos empregados para representação fonética são compostos por duas letras conforme a referência [3]) começa no centro da região espectralmente estável do fone "KK" e termina no centro da região espectralmente estável do fone "AA". Embora a unidade básica utilizada seja o difone, o sistema de síntese também utiliza algumas unidades maiores tais como trifones e polifones para alguns sons cuja representação por difones é problemática (por exemplo, tra, tre, tri, tro, tru). Nesse caso as definições são análogas à de difone no tocante à segmentação em pontos de estabilidade espectral dos fones inicias e terminais, variando apenas com relação ao número de fones contidos na unidade. Os difones foram obtidos a partir de logatomas (palavras sem sentido) em cuja região central estão os difones que se quer segmentar e armazenar. Assim, por exemplo, o difone "KK_AA" foi obtido por segmentação do logatoma PAKAPA.

É necessário, durante o processo de segmentação dos difones, determinar, além dos centros das regiões espectralmente estáveis de cada fone, o ponto de transição entre os fones para que, no momento da síntese, cada fone possa ser identificado a fim de receber as apropriadas alterações prosódicas. Por exemplo, se no difone "KK_AA" não fosse determinado o ponto de transição entre o "KK" e o "AA" não se poderia, no momento da síntese, atribuir durações específicas para o "KK" e o "AA", bem como contornos apropriados de freqüência fundamental.

2.2 Técnica PSOLA

As unidades acústicas que compõem o dicionário precisam ser representadas no computador. Os métodos tradicionais de representação de sinais de voz utilizam modelos paramétricos de codificação, como por exemplo o modelo LPC (Linear Predictive Coding), em que o sinal de fala é modelado como uma fonte de excitação aplicada a um filtro que possui apenas pólos, denominado filtro LPC.

Tais modelos permitem redução de memória exigida em relação à representação em forma de onda do sinal digitalizado. Mas sua grande vantagem, para a síntese de fala, é o fato de possilitarem alterações nas unidades concatenativas, para tratamento prosódico, pela simples alteração de parâmetros.

A maior desvantagem de tais modelos é que, sendo paramétricos, não conservam a forma da onda, o que tende a prejudicar a qualidade da fala.

A técnica PSOLA (Pitch Syncronous OverLap and Add [4]) faz uso de uma representação não paramétrica do sinal de fala de modo a melhor conservar a forma de onda das unidades concatenativas. De fato, a técnica PSOLA permite processar diretamente as unidades concatenativas em forma de onda, possibilitando alterações de duração e pitch diretamente na forma de onda. A melhor conservação da forma de onda resulta numa melhora de qualidade da fala sintetisada.

A dificuldade do emprego desta técnica é que, diretamente sobre a forma de onda, a alteração de duração e freqüência fundamental é bem mais complexa do que em representações paramétricas, onde basta alterar valores de parâmetros para que o sinal produzido pela síntese apresente as durações segmentais e os contornos de freqüência fundamental desejados.

Na síntese PSOLA, blocos de sinal referentes às unidades concatenativas são sobrepostos e adicionados no domínio do tempo para obtenção da forma de onda sintetizada. Inicialmente, os sinais de fala das unidades concatenativas são transformados em uma sequência de blocos elementares de curta duração $s_{m(n)}$, através da multiplicação do sinal original s(n) por uma sequência de janelas de análise $h_m(n)$:

$$s_{m(n)} = h_m(n-t_m) \cdot s(n)$$

onde t_m corresponde a marcas síncronas com o período fundamental das unidades gravadas. Observe-se que as janelas de análise são de tamanho tal que permitem uma superposição entre blocos elementares consecutivos. Também observe-se que $h_m(n)$ são janelas de Hanning assimétricas. Nas porções não sonoras do sinal de fala, as marcas são dispostas regularmente a intervalos de 10 ms.

Na etapa de síntese, haverá um mapeamento entre os intantes t_m e os instantes t_q que correspondem a marcas síncronas com o período fundamental que se queira atribuir aos segmentos que são produzidos pela superposição e adição dos blocos elementares.

A figura 2.3 mostra como é obtido um aumento na frequência fundamental do sinal sintetisado. Basicamente, a alteração da frequência fundamental por um fator **F** é conseguida pela multiplicação, pelo inverso deste fator, do atraso entre os blocos elementares que se juntam no tempo. Se, por exemplo, for desejada uma duplicação da frequência fundamental, deve-se dividir por 2 o atraso entre os blocos elementares. A figura 2.4 mostra como é obtida uma diminuição no valor da frequência fundamental do sinal gerado. Se, por exemplo, for desejado que a frequência fundamental caia pela metade, deve-se multiplicar por 2 o atraso entre os blocos elementares.

A figura 2.5 ilustra o processo de aumento da duração do sinal pela técnica PSOLA. Note que por simples replicação de blocos elementares, pode-se aumentar a duração do sinal produzido. Da mesma forma, por simples eliminação de blocos pode-se diminuir a duração do sinal produzido. A eliminação e a replicação são, evidentemente, proporcionais ao efeito que se queira obter. Por exemplo, se for desejada uma diminuição de duração pela metade, deve-se eliminar metade dos blocos. A figura 2.6 ilustra o processo de diminuição da duração do sinal pela técnica PSOLA.

Apesar da técnica PSOLA apresentar resultados melhores do que os apresentados pelas técnicas paramétricas, por conseguir uma melhor conservação da forma de onda, ela possui uma limitação. Para grandes variações de freqüência fundamental, ou de duração, a técnica PSOLA apresenta resultados insatisfatórios. Por exemplo, caso se aumente a duração de porções não sonoras da fala, elas se tornam periódicas devido à replicação de blocos elementares. Isto faz com que a fala produzida apresente características de "som metálico". Grandes variações de freqüência fundamental também prejudicam a qualidade da fala sintetizada, pois fazem com que o nível de superposição entre janelas varie demais.

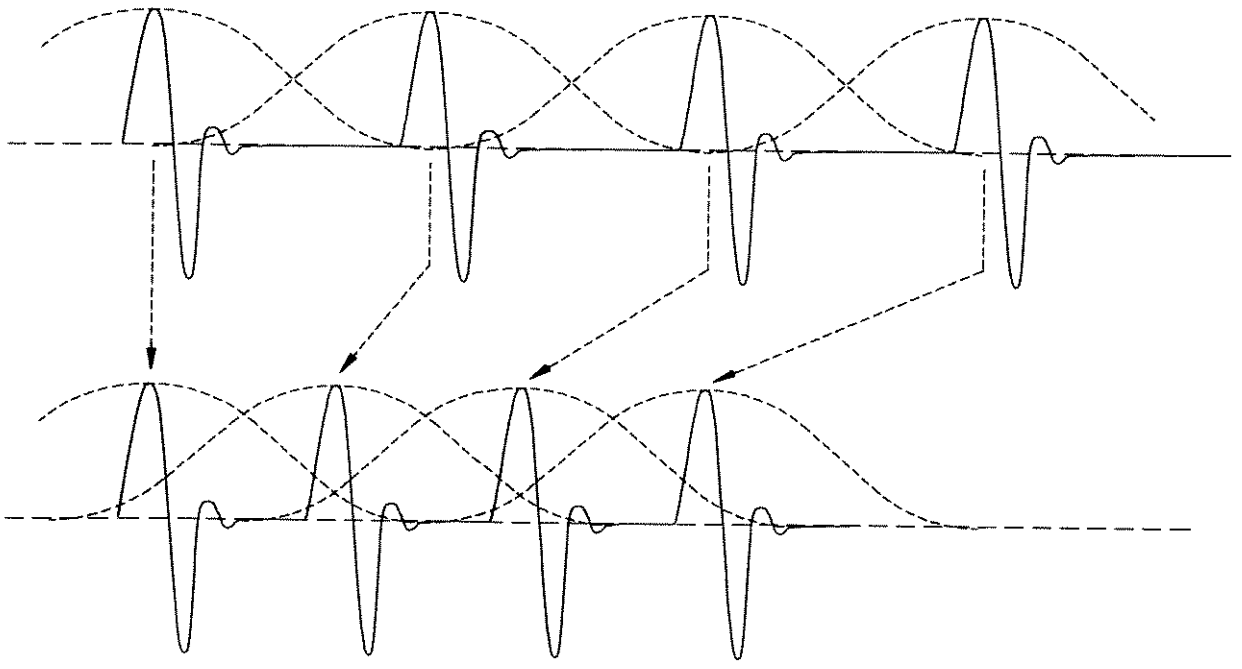


Fig. 2.3 - Aumento de freqüência fundamental pela técnica PSOLA.

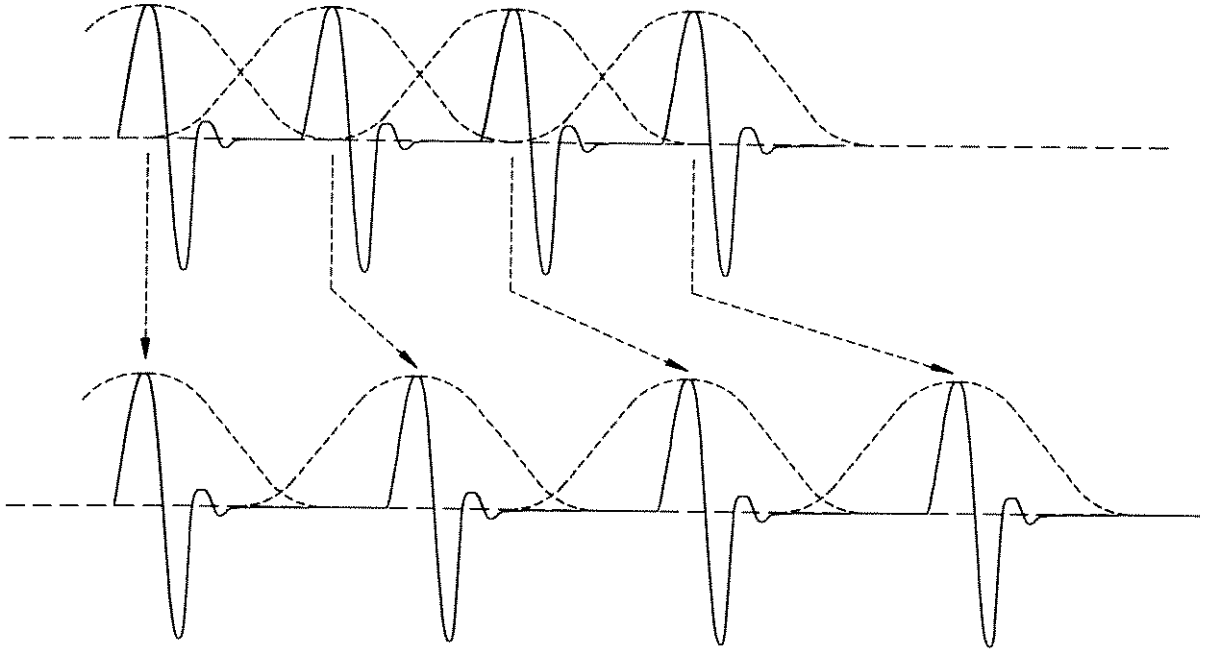


Fig. 2.4 - Diminuição de frequência fundamental pela técnica PSOLA.

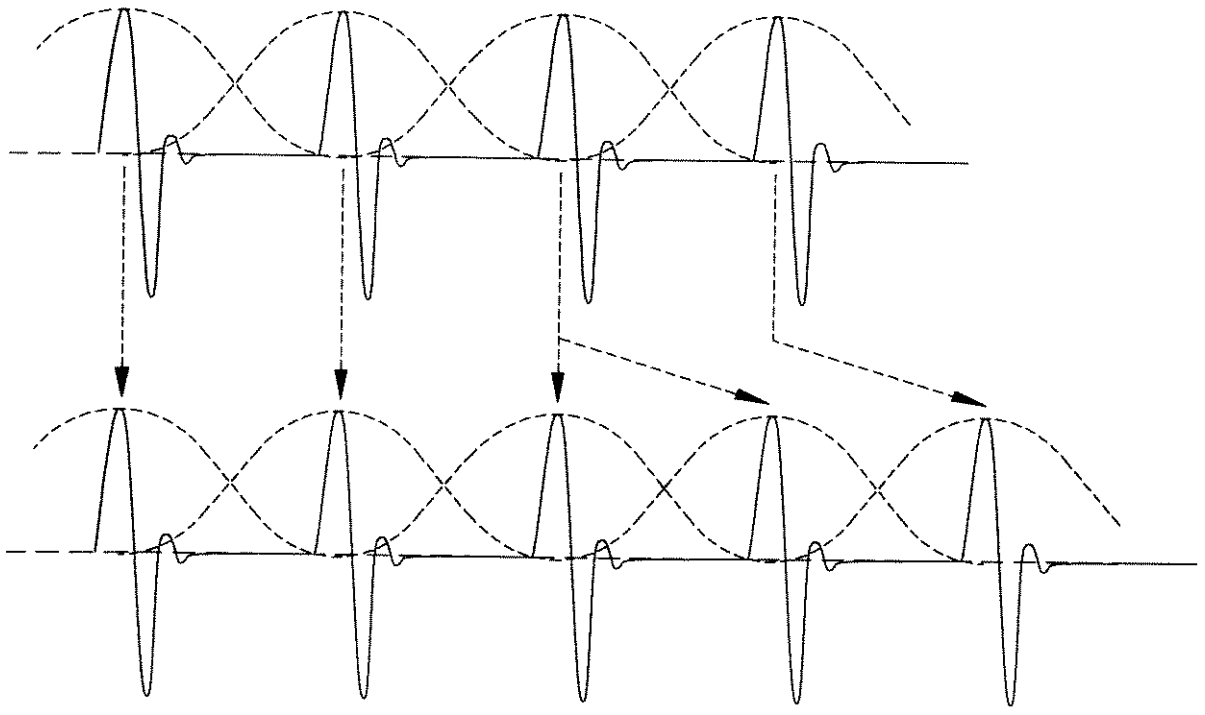


Fig. 2.5 - Aumento de duração pela técnica PSOLA

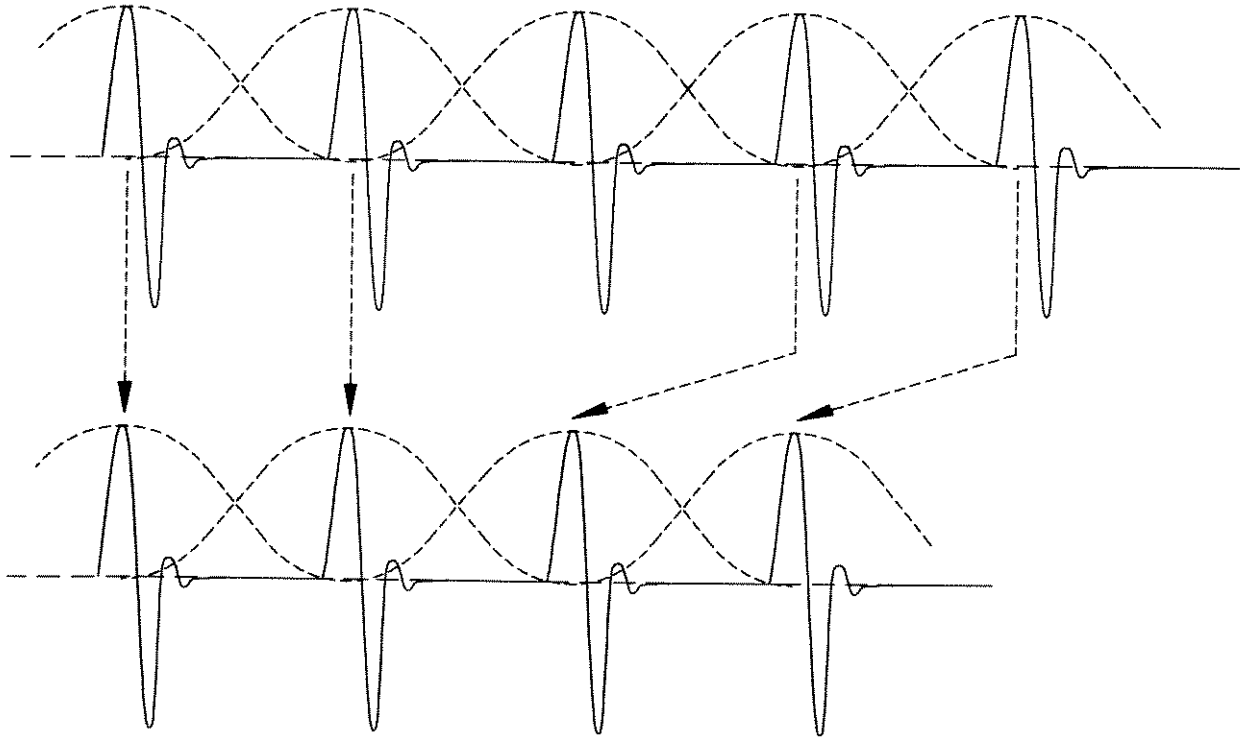


Fig. 2.6 - Diminuição de duração pela técnica PSOLA.

Como na técnica PSOLA as variações de duração são obtidas por meio de replicações e eleminações de blocos elementares, as mudanças de duração ficam quantizadas a valores racionais (...,1/2,3/4,4/5,...,5/4,4/3,2,...), isto é, não se pode aumentar ou diminuir a duração de um fone de um valor menor do que a duração de um bloco elementar. Isto significa que a técnica PSOLA não apresenta grande precisão quanto a variações de duração.

Capítulo 3
Ambiente de trabalho

Este capítulo descreve o ambiente de trabalho criado para o desenvolvimento do módulo prosódico. São apresentados os critérios que levaram à elaboração de um conjunto (corpus) de frases que foram pronunciadas por um falante e gravadas. Também são descritas sucintamente as funções desempenhadas por uma ferramenta de "software" desenvolvida para a análise prosódica de enunciados.

3.1 Explicação geral

Para construir um modelo prosódico aplicado à conversão texto-fala é necessário conhecer o "comportamento prosódico" de pelo menos um falante. Por "comportamento prosódico" entenda-se o controle que o falante exerce sobre os parâmetros prosódicos (duração, freqüência fundamental e energia) enquanto está falando.

Através da análise de um conjunto de frases pronunciadas pelo falante, pode-se chegar a uma descrição do seu comportamento prosódico. Esta análise busca quantificar os valores que os parâmetros prosódicos assumem ao longo dos enunciados. Sua realização exige um ferramental apropriado. Por isso, foi desenvolvido um programa de computador voltado para a análise prosódica.

Além de conhecer o comportamento prosódico de um falante, é preciso ter um meio para testar as regras do modelo prosódico durante sua evolução. O meio de testar as regras é um conversor texto-fala.

3.2 Corpus

Conforme foi dito anteriormente, para conhecer o comportamento prosódico de um determinado falante, é necessário analisar realizações de sua fala. A maneira mais simples de se fazer isto é gravar enunciados deste falante e posteriormente submetê-los à análise de seus parâmetros prosódicos. Usualmente, tais enunciados correspondem a leituras de sentenças ortográficas. O conjunto destas sentenças constitui o corpus a ser analisado.

No presente trabalho, inicialmente foram analisadas realizações de 78 sentenças declarativas. Ao pronunciá-las, o falante procurou fazê-lo de forma o mais neutra possível, isto é, sem enfatisar trechos em função de seu significado. Posteriormente, foram analisados enunciados relativos a 86 outras sentenças declarativas pronunciadas por outro falante, também de forma neutra. Ambos os falantes são adultos do sexo masculino, sendo que o segundo é um locutor profissional. Os falantes são oriundos dos estados de Minas Gerais e São Paulo, respectivamente. Quanto ao grau de instrução, o falante mineiro possui nível superior completo e o paulista nível secundário. Os enunciados foram digitalizados a uma frequência de amostragem de 16 KHz com 16 "bits" por amostra. Para o processo de digitalização e gravação foi utilizada uma placa "SOUND BLASTER" da CREATIVE LABS.

A escolha das frases que constituem um corpus não deve ser feita ao acaso, mas deve obedecer a critérios [5]. Tais critérios devem garantir que o conjunto seja suficientemente representativo do que acontece no uso do idioma. Pode-se delimitar este uso ao de uma aplicação particular como, por exemplo, um sistema de informações bancárias sobre contas correntes. Neste caso, as frases serão escolhidas dentro do conjunto das que comumente são produzidas neste tipo de aplicação. Porém, quando se pretende realizar síntese a partir de texto irrestrito deve-se buscar uma amostragem satisfatória daquilo que as pessoas dizem nas mais variadas situações.

A escolha das frases constituintes do corpus analisado no presente trabalho levou em conta relações sintáticas, posições relativas e comprimento dos constituintes sintáticos. Por relações sintáticas entenda-se a composição de orações por coordenação e subordinação, as relações entre sujeito e predicado, a existência de complementos verbais e nominais, etc. Os constituintes sintáticos das sentenças (sujeito, verbo, objeto, etc.) se agrupam em diferentes ordens (posições relativas) e variam em dimensão (número de palavras e de sílabas). Além disso, procurou-se frases com conteúdo semântico que não levam o falante a abandonar a neutralidade no falar. Assim,

foram evitadas frases sem sentido ou com uma inerente carga emotiva como, por exemplo, agressividade, ternura, espanto. Também deu-se preferência a frases com grande possibilidade de serem requeridas em aplicações de sistemas de conversão texto-fala. Por exemplo, valores numéricos para um sistema de saldo bancário por telefone. Essa preferência visa um direcionamento do trabalho para aplicações específicas ainda que o objetivo seja fala irrestrita.

As frases constituintes do corpus analisado são apresentadas no apêndice A deste trabalho.

Para elaboração do corpus contou-se com a valiosa colaboração da Dra. Sandra Madureira, lingüista e professora da Pontifícia Universidade Católica de São Paulo (PUC-SP).

3.3 Ferramenta para análise prosódica

Implementou-se um programa de computador para determinar as curvas de freqüência fundamental ao longo de enunciados previamente gravados e as durações de seus fones. Ele foi escrito na linguagem de programação C++, sob o paradigma de orientação ao objeto, para ser executado em computadores pessoais sobre o sistema operacional DOS.

Para determinar a curva de freqüência fundamental e as durações dos fones que constituem um enunciado é necessário saber onde começa e onde termina cada fone. Dada a complexidade de construção de um programa que identificasse os limites de início e término de cada fone, optou-se por realizar esta identificação manualmente. O processo de determinação manual dos limites referentes a cada fone é uma tarefa lenta, cansativa e entediante. Por isso, realizou-se um esforço no sentido de que o programa fosse de fácil uso. Optou-se, assim, por uma ferramenta gráfica que foi denominada **analisador prosódico**.

Basicamente, o analisador prosódico é um sistema que recebe como entrada um arquivo contendo sinal de fala digitalizado a uma freqüência de amostragem de 16 KHz, com 16 bits por amostra. O sinal de fala é processado para obter-se os períodos de freqüência fundamental. Em seguida, o analisador exibe na tela do computador dois gráficos. Um deles é a representação em forma de onda do sinal. O outro gráfico mostra o perfil de freqüência fundamental, calculado através do algoritmo descrito em [6].

Uma vez apresentados os gráficos, são colocadas à disposição do usuário as seguintes funções:

1. Ajuste de amplitude no gráfico do sinal

Esta função permite, na representação gráfica, que o sinal seja ampliado ou atenuado. Isto permite, por exemplo, que trechos de baixa energia sejam examinados com nitidez.

2. Ajuste de amplitude no gráfico de freqüência fundamental

Esta função proporciona uma maior ou menor resolução no gráfico de freqüência fundamental conforme estejam sendo examinadas tendências globais ou locais.

3. Ajuste de tempo

Esta função atua simultaneamente, e de maneira similar, sobre os gráficos de sinal e de freqüência fundamental que são sempre mantidos síncronos. Permite que sejam examinados na tela trechos de fala de maior ou menor duração.

4. Deslocamento no tempo

Esta função atua em ambos os gráficos e permite que sejam examinados trechos da fala em ambas as direções do eixo de tempo. O deslocamento pode ser rápido ou lento conforme a necessidade momentânea do usuário.

5. Duração

Esta função fornece a duração (em milisegundos) do trecho de fala especificado pelo usuário. Esta função pode ser útil, por exemplo, para saber as durações de cada palavra constituinte do enunciado.

6. Inspeção de valor de freqüência fundamental

Esta função permite a inspeção do valor de freqüência fundamental em qualquer ponto de sua curva. Basta que o usuário aponte, com o "mouse", o ponto da curva cujo valor ele queira saber, para que o número correspondente apareça escrito na tela do computador.

7. Edição de freqüência fundamental

Esta função dá ao usuário a possibilidade de traçar com o mouse uma curva de freqüência fundamental estilizada. O processo de traçado desta curva se dá com a determinação de segmentos de reta que unidos entre si constituem uma curva estilizada.

8. Play em forma de sinal

Esta função reproduz o trecho de sinal que o usuário especificar. Para o processo de reprodução é utilizada uma placa "SOUND BLASTER".

9. Play em forma de vocoder com freqüência fundamental natural

Esta função realiza síntese LPC (Linear Predictive Coding) referente a qualquer trecho escolhido pelo usuário. Para o processo de reprodução é utilizada a mesma placa "SOUND BLASTER".

Neste tipo de vocoder, o sinal é representado por um filtro LPC e por uma excitação que pode ser sonora ou não sonora. A excitação sonora é constituída por um trem de pulsos periódicos, espaçados pelo período de freqüência fundamental. Já a exitação não sonora corresponde a um ruído gaussiano branco.

A escolha de um vocoder LPC deveu-se à disponibilidade de um programa já pronto para realizar síntese LPC e que foi empregado no primeiro protótipo do conversor texto-fala desenvolvido no Laboratório de Processamento Digital de Fala do DECOM-FEE-UNICAMP [7]. Foi, portanto, uma decisão tomada com o objetivo de acelerar o trabalho.

10. Play em forma de vocoder com freqüência fundamental estilizada

Esta função também realiza síntese LPC de um trecho escolhido. Porém, a excitação do filtro não corresponde à freqüência fundamental detectada a partir da forma de onda e, sim, da curva de freqüência fundamental estilizada pelo usuário. Assim como as anteriores esta função também utiliza a placa "SOUND BLASTER" para realizar a reprodução sonora.

11. Zoom

Esta função permite ao usuário marcar um trecho de voz que queira analisar e faz com que este trecho passe a ocupar toda a tela. Esta função pode ser repetida quantas vezes o usuário desejar até que na tela seja exibido apenas o trecho em que ele estiver concentrando sua atenção.

12. Back

Esta função permite, caso o usuário tenha usado a função zoom, que ele possa voltar à tela que estava sendo exibida antes que ele usasse a função zoom pela última vez. Da mesma forma que a função zoom, a função back pode ser acionada repetidas vezes. Isto significa que, enquanto a função zoom leva o programa a estar em sucessivos estados de exibição, a função back faz com que estes estados se sucedam em ordem inversa, estabelecendo, por assim dizer, um caminho de volta.

13. Key

A curva de freqüência fundamental estabelecida através de análise acústica do sinal de fala pode ser apresentada de duas maneiras distintas. Esta função permite ao usuário escolher qual a forma de apresentação a ser usada pelo programa.

Uma forma de apresentação consiste na não ligação de valores de freqüência não nulos a valores nulos quando do traçamento da curva. Esta forma se presta melhor ao exame de contornos globais. Na outra forma de apresentação, valores de freqüência não nulos são ligados a valores

nulos de forma que a curva seja contínua. Esta forma se presta melhor ao exame de contornos locais.

A figura 3.1 ilustra uma tela do programa de análise prosódica.

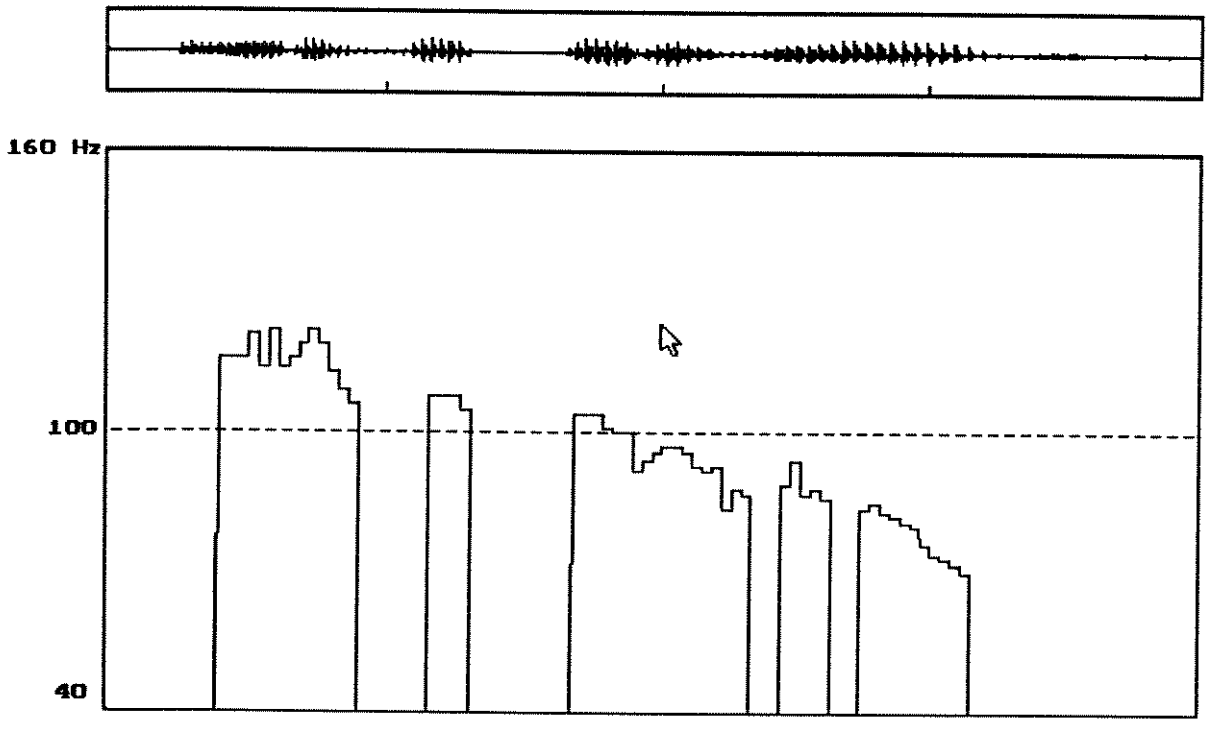


Fig. 3.1 - Tela do analisador prosódico sendo utilizado para analisar o enunciado referente à frase: "_ Curva perigosa".

3.4 Processo de análise

Uma vez concluída a gravação dos 164 enunciados, procedeu-se à análise destes através do analisador prosódico. Um por um, cada enunciado foi analisado e os dados resultantes desta análise foram expressos na forma de gráficos de freqüência fundamental. Assim, foram construídos 164 gráficos representando a evolução da freqüência fundamental ao longo do tempo para cada enunciado.

O processo de análise de cada enunciado consistiu, portanto, na identificação dos instantes de início e término de cada fone constituinte, no registro da duração de cada fone, no registro dos valores de freqüência fundamental correspondentes a estes instantes e no registro do formato da curva de freqüência fundamental entre eles.

3.5 Conversor texto-fala

Foi usado o conversor texto-fala do tipo concatenativo descrito no capítulo II para teste e avaliação das regras do modelo prosódico.

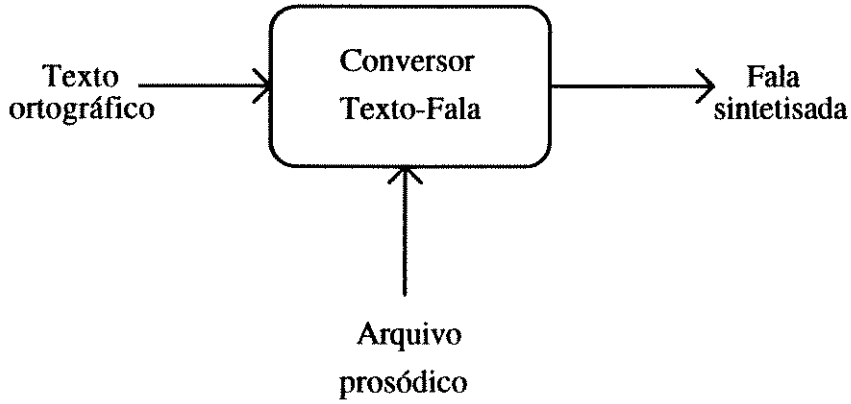


Fig. 3.2 - Esquema de entradas e saídas do conversor texto-fala.

Criou-se dois mecanismos de entrada de dados para o conversor (Fig. 3.2). Um deles é o texto ortográfico. Através dele, pode-se testar rapidamente novas regras. Outro meio de entrada é através de um arquivo prosódico. Tal arquivo contém a especificação dos fones constituintes da frase em formato seqüencial e informações sobre os valores de freqüência fundamental e duração para cada fone. Esta forma de entrada de dados permite o refinamento de regras através de pequenas alterações sucessivas. A cada execução do programa de síntese é gerado um arquivo prosódico correspondente ao texto sintetisado quando a entrada é feita no modo texto ortográfico.

A figura 3.3 ilustra o arquivo prosódico gerado pelo sintetisador ao receber o texto " O preço da tarifa telefônica foi reduzido" .

	FONE	F0 (Hz)	DURAÇÃO (ms)	F0 (Hz)
0:	UW	135	83	131
1:	PP	131	98	131
2:	RX	131	40	131
3:	EE	131	142	132
4:	SS	132	131	125
5:	UW	125	103	121
6:	DD	121	89	119
7:	AA	119	111	118
8:	TT	118	85	117
9:	AA	117	119	115
10:	RX	115	40	116
11:	II	116	146	121
12:	FF	121	126	115
13:	AA	115	120	109
14:	TT	109	85	109
15:	EE	109	139	108
16:	LL	108	81	109
17:	EE	109	127	109
18:	FF	109	126	111
19:	OO	111	126	112
20:	NN	112	85	107
21:	IY	107	84	103
22:	KK	103	102	106
23:	AA	106	120	124
24:	FF	124	145	121
25:	OO	121	148	119
26:	IY	119	108	117

27:	RR	117	93	112
28:	EE	112	133	106
29:	DD	106	86	105
30:	UW	105	103	104
31:	ZZ	104	86	105
32:	II	105	146	105
33:	DD	105	90	100
34:	UW	100	103	95

Fig. 3.3 - Arquivo prosódico gerado pela síntese do texto " O preço da tarifa telefônica foi reduzido."

Capítulo 4
Considerações lingüísticas

Este capítulo visa dar ao leitor um pequeno resumo de alguns conhecimentos lingüísticos para facilitar a compreensão deste trabalho. Inicialmente, é dada uma pequena explicação a respeito da anatomia do sistema que seres humanos usam para falar. Também são apresentadas definições de Fonologia, fonemas, alofones e fones. Em seguida são apresentadas definições de Fonética e de suas subdivisões: Fonética Articulatória e Fonética Acústica. São definidos modo e ponto de articulação, e são apresentadas classificações dos fonemas com relação a esses critérios. Mostra-se uma breve caracterização fonêmica do ponto de vista acústico e uma explanação sobre transcrição fonética. Em seguida define-se coarticulação. Além disso, são apresentados os conceitos de prosódia e parâmetros prosódicos. Também são citadas as funções que a prosódia desempenha na fala.

4.1 Trato vocal

O trato vocal é uma estrutura tubular pela qual passa o fluxo de ar vindo dos pulmões e modulado nas cordas vocais. Sua principal função é moldar o espectro de freqüência da onda sonora que vem das cordas vocais e promover constricções para a geração de certos tipos de som. Tais funções são exercidas por meio dos articuladores que se movem dentro do trato vocal.

Os órgãos usados para a produção da fala não possuem exclusivamente esta função. Ao contrário, evolutivamente falando, o homem parece ter tomado emprestados para a produção da fala certos órgãos cujas funções mais básicas são a respiração, deglutição e olfação. Os órgãos constituintes do trato vocal empregados na produção da fala , e que se movem neste processo, são denominados articuladores da fala. A figura 4.1 corresponde a uma vista da seção transversal do trato vocal onde são indicados os articuladores da fala..

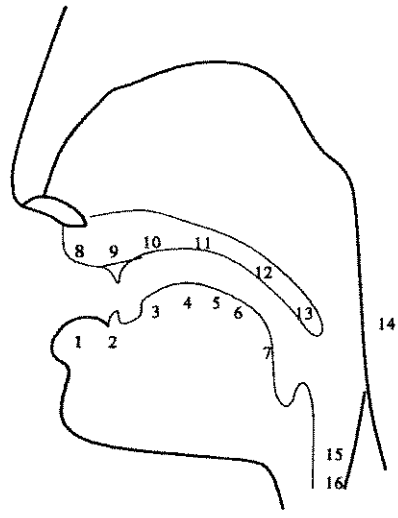


Fig. 4.1 - Vista da seção transversal do trato vocal (1) lábio inferior, (2) incisivo inferior, (3) ponta da língua, (4) dorso, (5) frente, (6) costas, (7) raiz, (8) lábio superior, (9) incisivo superior, (10) alvéolo, (11) palato duro, (12) velum, (13) úvula, (14) faringe, (15) laringe, (16) cordas vocais e glote.

Além dos articuladores mostrados na fig. 4.1, os pulmões também desempenham importante papel no processo de produção da fala, conforme será descrito logo mais.

Através de comandos neuromotores, enviados aos músculos do trato vocal, o falante controla os movimentos dos articuladores da fala. Tais movimentos são responsáveis pela produção do sinal de fala.

4.1.1 Os pulmões

Cabe aos pulmões produzir o fluxo de ar que, passando através da laringe e do trato vocal, dá origem à onda sonora denominada sinal de fala. Situados na cavidade torácica, a função principal dos pulmões é promover o processo respiratório por meio de inspirações e expirações de ar que ocorrem por ação de um músculo denominado diafragma e que está situado na porção inferior do tórax. Normalmente, a fala é produzida somente durante as expirações, embora existam sons em diversas línguas que são produzidos durante inspirações.

Durante a fala, a pressão de ar produzida pelos pulmões não sofre grandes variações. Os valores absolutos desta pressão permanecem ligeiramente superiores à pressão atmosférica.

4.1.2 Cordas vocais

Sendo muito baixa a pressão do fluxo de ar produzido pelos pulmões, ele por si só não é suficiente para produzir sons audíveis. Sons são produzidos, entretanto, quando o caminho pelo qual segue o fluxo de ar, que vai dos pulmões até a saída do trato vocal, sofre estreitamentos ou totais oclusões. A principal fonte de origem do sinal sonoro são as cordas vocais que podem obstruir parcial ou completamente o fluxo de ar que vem dos pulmões.

Situadas na laringe, as cordas vocais são um par de estruturas elásticas constituídas por tendões, músculos e membrana mucosa. Seu comprimento médio é de 15 mm para homens e 13 mm para mulheres, podendo, pela ação de contrações musculares, sofrer alterações de comprimento, espessura e posicionamento em várias configurações.

Durante a respiração normal, as cordas vocais permanecem afastadas entre si para evitar a geração de sons durante a passagem do ar. Uma aproximação entre as cordas vocais leva à geração de sons. Os chamados sons sonoros são produzidos somente quando as cordas vocais estão vibrando. A freqüência com que as cordas vocais vibram em determinado instante denomina-se freqüência fundamental.

4.2 Fonologia

4.2.1 Definição

Fonologia é o ramo da lingüística que estuda os sons constituintes da fala segundo seu aspecto funcional, isto é, segundo o papel destes sons na comunicação através da fala, sem preocupar-se com suas propriedades acústicas e articulatórias.

4.2.2 Fonemas

Entre a infinita variedade de sons que podem ser produzidos pelo trato vocal humano, cada língua escolhe um conjunto de sons através dos quais é codificada a informação que se quer enunciar através da fala. Estes sons escolhidos são agrupados em classes que podem ser discernidas entre si pelos ouvintes. A estas classes, com papel distintivo na comunicação, dá-se o nome de fonemas.

Um fonema é uma abstração, não possui natureza acústica. Pode-se pensar um fonema como sendo um tipo de som com características gerais que permitem diferenciá-lo de outros fonemas. Por exemplo, na Língua Portuguesa temos /p/ e /b/ como fonemas distintos porque diferenciam palavras tais como pata e bata .

Ao especificar-se características mais precisas para um fonema, além das que são suficientes para diferenciá-lo de outros fonemas, obtém-se uma sub-classe denominada alofone. Por exemplo, na Língua Portuguesa tem-se o fonema /t/ que forma palavras tais como tia, tatu, tábua,etc...Porém a maneira como os cariocas pronunciam o /t/ é diferente da maneira como os paulistas o fazem. Tanto cariocas como paulistas dão ao /t/ características próprias, criando portanto diferentes tipos de /t/ que são chamados alofones.

4.2.3 Fones

Um fone é a realização acústica de um fonema. Não é uma classe de som, como um fonema, mas sim um sinal sonoro. Para um mesmo fonema pode-se produzir um número virtualmente infinito de fones. Por exemplo, se um falante pronunciar um /a/ cem vezes, a cada pronúncia ele produzirá um sinal sonoro distinto dos correspondentes às demais pronúncias. Estes sinais, apesar de distintos entre si, apresentarão um grau de semelhança suficiente para caracterizá-los como realizações acústicas de um mesmo fonema.

Para entender a relação entre fonema, fone e alofone pode-se recorrer a uma analogia com a programação orientada ao objeto. Imagine que fonema é uma classe com características bem gerais e alofone é outra classe que herda as características de fonema além de possuir outras que lhe são próprias. Os fones correspondem, nesta analogia, a instâncias da classe alofone.

4.3 Fonética

4.3.1 Definição

Fonética é o ramo da Lingüística que estuda a fala buscando determinar suas propriedades acústicas e articulatórias.

4.3.2 Fonética articulatória

A fonética articulatória busca descrever as relações que existem entre fonemas e as posições e movimentos dos articuladores do trato vocal que ocorrem durante sua produção.

4.3.2.1 Modo de articulação

O modo de articulação refere-se ao caminho pelo qual segue o fluxo de ar ao longo do trato vocal e ao grau de impedimento a este fluxo gerado pelas constricções do trato vocal. Os diferentes graus de obstrução resultam na produção de diferentes sons.

Na produção de vogais e ditongos, o fluxo de ar segue diretamente através do trato vocal, não encontrando nenhuma constricção estreita o suficiente para criar um efeito de turbulência.

Durante a produção de sons nasais, o abaixamento do velum faz com que o fluxo de ar seja totalmente obstruído oralmente e escape pelas fossas nasais. Como exemplo cita-se o /n/.

Na produção das consoantes líquidas a língua é utilizada para obstruir a passagem de ar no centro do trato vocal de modo que o fluxo escape pelas laterais da língua, razão pela qual as líquidas são também chamadas de laterais. Como exemplo cita-se o /l/ e o /r/.

Na produção das consoantes plosivas (oclusivas), ocorre uma total obstrução do trato vocal seguida pela liberação da passagem de ar. Com a obstrução ocorre um aumento na pressão do ar que é repentinamente liberado criando uma explosão sonora. Como exemplo cita-se o /k/ e o /p/.

As consoantes fricativas são produzidas quando uma constricção do trato vocal gera uma turbulência no fluxo de ar. Como exemplo pode-se citar o /s/ e o /f/.

Tanto as plosivas como as fricativas podem ser produzidas com ou sem vibração das cordas vocais. No caso de haver vibração elas são chamadas de sonoras e, no caso de não haver vibração, de surdas. O /v/ e o /z/ são exemplos de fricativas sonoras enquanto que o /s/ é uma fricativa surda. O /p/ é um exemplo de plosiva surda, enquanto que o /b/ é um exemplo de plosiva sonora.

4.3.2.2 Ponto de articulação

O ponto de articulação refere-se ao ponto do trato vocal onde a constricção deixa a menor abertura para a passagem do fluxo de ar, ou seja, ao ponto onde é máxima a constricção. Os principais pontos de articulação são: os lábios, os dentes, os alvéolos, o palato duro, o palato mole, a úvula, a faringe e a glote.

Conforme o ponto de articulação, as consoantes do Português podem ser classificadas em seis categorias: bilabiais, labiodentais, dentais, alveolares, palatais e velares.

Ocorre constricção dos lábios inferior e superior na produção de fonemas tais como o /p/ e o /b/, razão pela qual estes fonemas são classificados como consoantes bilabiais.

A produção de uma consoante labiodental ocorre pela aproximação do lábio inferior aos dentes superiores. Como exemplo pode-se citar o /f/ e o /v/.

Na produção de consoantes dentais a língua toca os dentes incisivos superiores. Como exemplo cita-se o /d/ e o /t/.

A articulação das consoantes alveolares dá-se com a língua tocando os alvéolos como ocorre na produção de /l/ e /n/.

Nas consoantes palatais o dorso da língua realiza uma constricção com o palato duro. O /nh/ é um exemplo de consoante palatal.

Nas consoantes velares o dorso da língua aproxima-se do palato mole. Exemplos: /k/ e /g/.

Devido ao fato de serem produzidas com o trato vocal relativamente aberto, as vogais são classificadas segundo três principais aspectos articulatórios. O primeiro deles é o grau de abertura da boca, critério pelo qual as vogais são classificadas como fechadas ou abertas. Como exemplo de vogal fechada pode-se citar o /u/ e como exemplo de vogal aberta pode-se citar o /a/. O segundo aspecto é a posição dos lábios que podem estar arredondados ou não arredondados. O /a/, por exemplo, é produzido com os lábios não arredondados enquanto o /u/ é produzido com os lábios arredondados. O terceiro aspecto é o ponto de constricção máxima pelo qual as vogais são classificadas como anteriores, centrais ou posteriores. A vogais /e/, /i/ e /u/ são, respectivamente, exemplos de vogais anteriores, centrais e posteriores.

4.3.3 Fonética acústica

A fonética acústica estuda a fala do ponto de vista do sinal sonoro emitido e o relaciona a aspectos lingüísticos. Ela não se preocupa com o processo de produção da fala como a fonética articulatória, mas sim com a onda sonora produzida.

Como um mesmo fonema pode ser articulado de diferentes maneiras e por diferentes tratos vocais, existe uma grande variedade de sinais de fala para um mesmo fonema. A fonética acústica procura, então, caracterizar cada fonema em termos de aspectos acústicos comuns a diferentes realizações.

As relações estabelecidas pela fonética acústica entre fonemas e suas realizações acústicas fornecem uma base para a codificação, síntese e reconhecimento de fala, estudos de percepção, etc.

Devido ao fato de a maior parte das características acústicas dos fonemas ser mais aparente no domínio da freqüência do que no domínio do tempo, a fonética acústica usa principalmente informações espectrais extraídas do sinal sonoro. Essas informações espectrais comumente são apresentadas na forma de gráficos de amplitude versus freqüência e de espectrogramas que nada mais são do que uma representação tridimensional (amplitude X freqüência X tempo) do sinal de fala. Através dos gráficos de amplitude versus freqüência ou de

espectrogramas pode-se visualizar os formantes que correspondem a freqüências de ressonância do trato vocal.

4.3.3.1 Caracterização fonêmica do ponto de vista acústico

Conforme foi dito anteriormente, a fonética acústica procura caracterizar os fonemas em termos das características acústicas comuns a diferentes realizações.

Vogais

Vogais normalmente correspondem a sons vozeados (cuja produção se dá com as cordas vocais vibrando) e têm grandes amplitudes se comparadas aos demais fonemas. Assim como todo som produzido a partir de uma excitação puramente periódica, a energia das vogais se concentra abaixo de 1 KHz e cai cerca de 6 dB por oitava. Devido ao fato de serem produzidas a partir de uma excitação periódica, os picos observados em um gráfico de amplitude versus freqüência de vogais estão espaçados de F0 Hz, isto é, estão espaçados entre si por distâncias cujo valor é o mesmo da freqüência fundamental com que vibram as cordas vocais.

Vogais são distinguidas entre si principalmente pelas localizações de seus primeiros três formantes.

Ditongos

Ditongos são semelhantes a vogais no sentido de que eles são produzidos a partir de excitação sonora e têm padrões de formantes bem definidos. Porém os ditongos diferem das vogais por não poderem ser caracterizados por um padrão estático de formantes. Ditongos são sons dinâmicos em que os articuladores do trato vocal se movimentam suavemente durante sua produção levando a uma mudança gradativa do padrão de formantes. Como exemplo de ditongo, cita-se o /ai/ da palavra "pai".

Basicamente, do ponto de vista fonético acústico, pode-se definir vogais e ditongos como sons caracterizados por padrões de formantes bem definidos.

Líquidas

Líquidas são consoantes sonoras. Assim como as vogais, as líquidas possuem uma grande amplitude de sinal e a maior parte da energia concentra-se nas baixas freqüências. Porém o espectro das líquidas geralmente possui um pouco menos de energia do que o de vogais.

Nasais

As formas de onda de consoantes nasais lembram formas de onda de vogais, porém são significativamente menos intensas devido ao efeito de atenuação do som na cavidade nasal. Os formantes estão tipicamente abaixo de 850 Hz e não de 1 KHz como nas vogais .

Fricativas

As formas de onda de sons fricativos mostram aperiodicidade e baixa intensidade. Espectralmente a maior parte da energia está nas altas freqüências.

Plosivas

Ao contrário de outros tipos de consoantes que podem ser descritas em termos de padrões espectrais, as plosivas são fonemas transientes e, portanto, são acusticamente complexas.

4.4 Transcrição ortográfico-fonética

Os símbolos usados para se escrever não representam apropriadamente os sons gerados ao se falar. Isto faz com que se necessite fazer uma transcrição ortográfico-fonética de um texto que se queira converter em fala. Fazer uma transcrição ortográfico-fonética de um texto significa representar graficamente cada fonema constituinte de um possível enunciado que um falante faria a partir deste texto. Fazer a transcrição fonética de um enunciado é representar graficamente cada

fonema constituinte deste enunciado. Gera-se, portanto, uma seqüência de símbolos fonéticos referentes à seqüência de fones que constitui o enunciado.

O processo de transcrição fonética leva em conta o nível de detalhes a que se quer chegar. Assim, pode-se transcrever um enunciado sem considerar muitos detalhes na diferenciação entre fones, de modo que fones com poucas diferenças acústicas sejam representados pelo mesmo símbolo. Este tipo de transcrição, feita sem considerar muitos detalhes, é chamado de transcrição larga. Por outro lado, denomina-se transcrição estreita aquela que é feita buscando levar em conta o maior número possível de detalhes acusticamente perceptíveis.

Uma vez que seja escolhido um conjunto de símbolos e estabelecida uma correspondência biunívoca entre símbolos e fones (isto é, um fone é sempre representado por um símbolo e um símbolo sempre representa um fone) deve ser possível transcrever enunciados de qualquer língua. A fim de que foneticistas possam transcrever os fones de qualquer língua, e representá-los de forma inambígua, foi criado o Alfabeto Fonético Internacional (AFI) pela Associação Fonética Internacional. Na figura 4.2 é mostrada uma representação do Alfabeto Fonético Internacional.

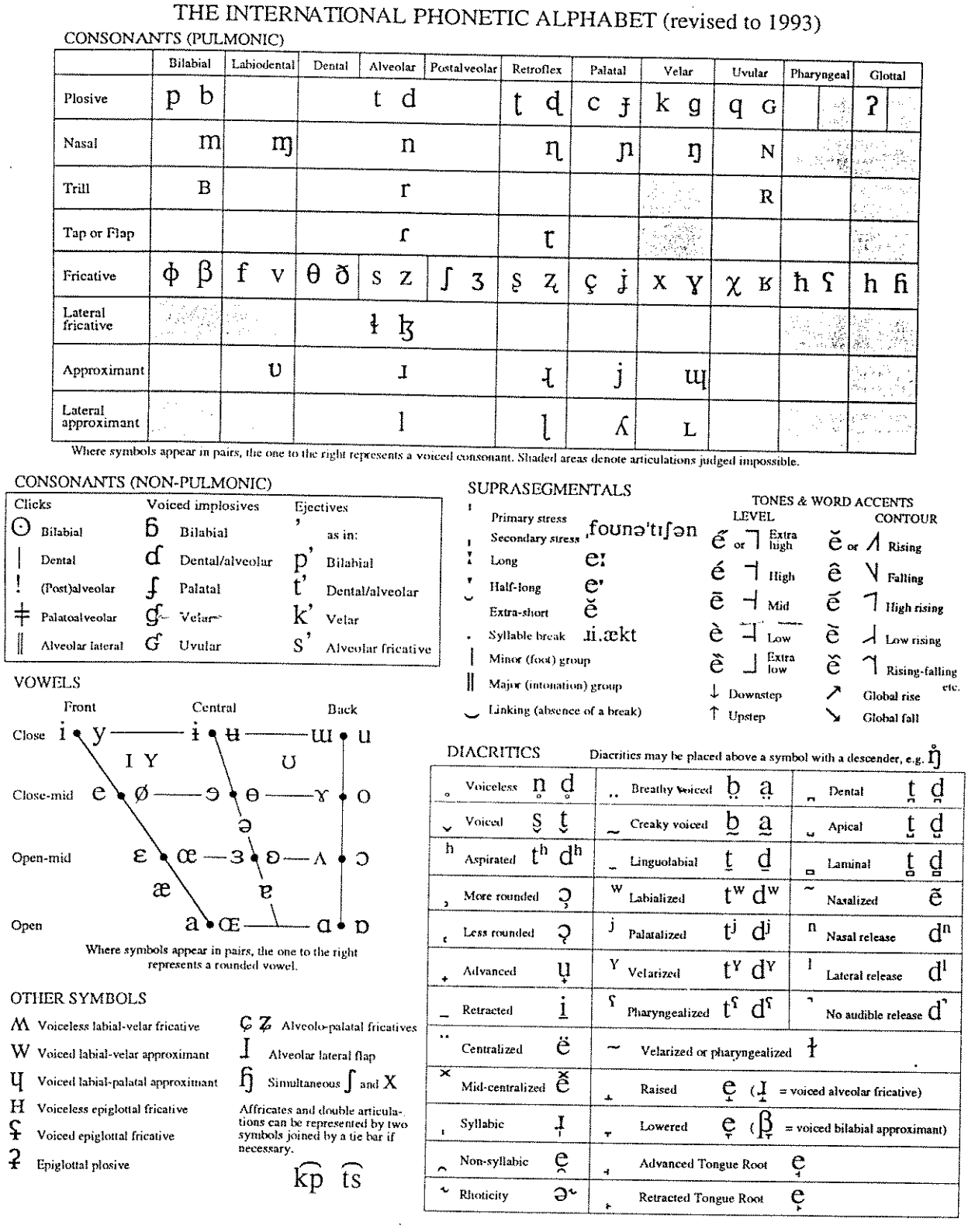


Fig 4.2 - Alfabeto Fonético Internacional (IPA) [8].

4.5 Coarticulação

No processo de produção da fala os articuladores do trato vocal realizam movimentos denominados gestos. Ao serem produzidos sucessivos fones, observa-se que os gestos que se sucedem no tempo influenciam uns aos outros. Isto significa que as características articulatórias e, portanto, também as características acústicas dos fones produzidos, são influenciadas pelos contextos fonéticos em que eles ocorrem.

Um fone pode ser influenciado tanto pelo contexto fonético que o segue no tempo como pelo contexto que o precede. No primeiro caso, diz-se que há coarticulação antecipatória, ou da direita para a esquerda, e, no segundo caso, que há coarticulação da esquerda para a direita. Enquanto a coarticulação da esquerda para a direita está relacionada com a inércia mecânica dos articuladores, a coarticulação antecipatória corresponde a um planejamento dos movimentos dos articuladores. O objetivo é conseguir produzir uma seqüência de fones com um mínimo de esforço, ou seja, gestos são sobrepostos no tempo para que seja auferida comodidade no falar.

Fenômenos de coarticulação podem levar um determinado segmento a ter uma duração muito reduzida ou mesmo a ser omitido durante uma fala rápida. Por exemplo, se a palavra "partes" for pronunciada rapidamente, a última vogal será tão reduzida que praticamente desaparecerá.

4.6 Prosódia

4.6.1 Definição

Conforme foi dito anteriormente, os fonemas correspondem a categorias sonoras distintivas entre si e é com as realizações físicas de fonemas, isto é, fones, que o ser humano produz a fala. Sendo assim, a fala pode ser pensada como sendo uma seqüência fonética. Mas seria a fala apenas uma seqüência de sons? Para responder a essa pergunta pode-se recorrer a uma suposição. Imagine um sistema de conversão texto-fala por concatenação. Imagine que este sistema consiga concatenar de forma perfeita as unidades sonoras mas que não possua nenhuma

espécie de tratamento prosódico. Conseguiria este sistema produzir enunciados? Sem dúvida que sim. Mas a fala produzida não seria de modo nenhum semelhante à que um falante produziria, ou seja, não apresentaria naturalidade, e, quase sempre, resultaria em enunciados de difícil compreensão. Seria uma fala monótona, semelhante à fala dos robôs dos filmes de ficção.

Costuma-se chamar os fonemas de segmentos e, por extensão, cadeias fonéticas são ditas estar no nível segmental da fala. Acima do nível segmental, ou seja, fonêmico, está o nível suprasegmental que também é chamado de prosódico. Desse modo, define-se a prosódia como as características da fala que estão num nível imediatamente acima da cadeia fonética. Essa caracterização nada mais é do que um meio de prover mais informação para a fala do que a codificada através da seqüência fonética; informação esta que desempenha importantes funções na comunicação.

4.6.2 Funções da prosódia

A principal função da prosódia é ajudar o ouvinte no processo de compreensão da mensagem codificada foneticamente. Caracterizações diferenciadas de diferentes trechos de um enunciado permitem que o falante confira ao enunciado uma estruturação que facilitará o processo de compreensão para o ouvinte. Esta estruturação se dá em todo e qualquer enunciado realizado por um falante, mas seu papel cresce em importância para enunciados longos. De fato, quanto mais longo for o enunciado, mais necessário será dividí-lo em partes para que o ouvinte consiga tratá-lo mentalmente. A título de ilustração, imagine a seguinte frase sendo pronunciada na forma de uma cadeia de sons sem qualquer estruturação: "_À mesa de pinho-de-riga, que convivia com o sofá importante, cadeiras de palhinha sem história e um tamborete rústico, sem qualquer vexame para a peça nobre do mobiliário ou constrangimento das demais, estava sentado um homem de meia idade, magro e louro, que sorriu para o recém-chegado e, com um aceno leve, dispensou a presença dos acompanhantes"(Carlos Drumond de Andrade, "Os dias lindos"). Seria fácil compreendê-la?

Através da estruturação prosódica pode-se ainda evitar ambigüidades de significado. Imagine um sistema de conversão texto-fala lendo a expressão "A*(B+C)" sem nenhum

tratamento prosódico. Um ouvinte entenderia "(A*B)+C" ou "A*(B+C)"? De fato, poderia entender de qualquer uma das formas. Mas se a produção do enunciado for feita de modo a que características prosódicas confiram a ele estruturação, desaparecerá a ambigüidade.

Outra função da prosódia é comunicar o estado emocional do falante. Através da fala de uma pessoa pode-se saber se ela está calma ou irritada, alegre ou triste, desanimada ou entusiasmada, etc.

O nível suprasegmental pode também fornecer informação sobre a identidade do falante. De fato, o "jeito de falar" muitas vezes é suficiente para identificar uma pessoa.

A intensão do falante pode ser expressa através da prosódia. Por exemplo, uma frase dita com ironia dá a entender o oposto do que a mesma frase resultaria caso fosse dita sem ironia. Mas o que é o "tom irônico" senão uma informação veiculada pela prosódia? Outro exemplo ocorre quando o falante quer fazer rir o ouvinte e, para isso, dá a sua voz um "tom engraçado".

A personalidade do falante também é evidenciada pela maneira com que ele fala. É comum ouvir-se dizer que determinada pessoa possui uma maneira arrogante de falar, ou humilde, ou tímida, ou extrovertida,...

Informação geográfica também é transmitida pela prosódia. Ao ouvir uma pessoa falar é possível distinguir, por exemplo, de qual região do país ela provem.

A prosódia desempenha uma função estética na fala, isto é, através da riqueza de características prosódicas a fala adquire uma beleza que de modo nenhum teria caso ficasse restrita ao nível segmental.

O falante pode realçar certas palavras de um enunciado através da prosódia. Este realce é conhecido como acentuação e desempenha um importante papel na comunicação. Através dele o ouvinte sabe em que partes do enunciado ele deve concentrar sua atenção, isto é, o ouvinte sabe quais palavras são mais importantes para a compreensão do enunciado.

Todas as funções prosódicas são desempenhadas por meio de variações das características da fala. Mas que características são essas que definem o nível suprasegmental? Elas são chamadas de parâmetros prosódicos. Freqüência fundamental, duração e energia constituem os parâmetros prosódicos que também são chamados de traços prosódicos. Uma outra definição de prosódica

poderia ser, então, o conjunto de efeitos produzidos pela freqüência fundamental, pela duração e pela energia sobre uma seqüência de fones.

A energia é, dos três parâmetros prosódicos, o que menos influência exerce sobre a fala. Razão pela qual os modelos prosódicos aplicados a sistemas de conversão texto-fala usualmente não consideram este parâmetro.

4.6.3 Freqüência fundamental (F0)

Freqüência fundamental é o nome dado à taxa de vibração das cordas vocais de uma pessoa enquanto ela está falando. Do ponto de vista acústico, a periodicidade do sinal de fala corresponde à freqüência fundamental, ou seja, o número de vezes que as cordas vocais se abrem e fecham por segundo é o mesmo número de repetições que a forma de onda apresenta por segundo.

O correlato da freqüência fundamental no plano perceptual denomina-se "pitch" e, embora um não se relacione linearmente ao outro, ambos são usualmente, no contexto da conversão texto-fala e da análise acústica da fala, empregados indistintamente. Assim, é comum ouvir-se alguém dizer que a freqüência de pitch detectada através de um programa de análise acústica foi de, por exemplo, 120 Hz, embora o mais correto fosse dizer que a F0 detectada foi de 120 Hz. Enquanto a freqüência fundamental é uma grandeza acústica expressa por medidas numéricas (usualmente em Hz), a freqüência de pitch se relaciona com o julgamento que o ouvinte faz se o som que ele está ouvindo é baixo ou alto, ou se está subindo ou descendo.

F0 é o parâmetro prosódico mais importante, ou seja, o modelamento correto da curva de freqüência fundamental produz efeitos perceptuais mais relevantes do que o modelamento da duração e da energia.

4.6.4 Duração

O termo duração refere-se ao intervalo de tempo que vai desde o início de um fone até o seu término na fala. Usualmente, chama-se a grandeza física, expressa em unidades de tempo, de

comprimento do fone e emprega-se o termo duração para designar o seu correlato perceptual. Neste trabalho, entretanto, os dois termos são empregados indistintamente.

A produção de fones durante a fala, é um processo que envolve, conforme visto anteriormente, a atuação de um sistema mecânico. Sendo assim, devido a razões de inércia, os fones necessitam um certo intervalo de tempo para serem produzidos. Já do ponto de vista do ouvinte, cada som deve ter uma duração suficiente para que possa ser discernido.

Uma das maneiras pela qual o nível suprasegmental codifica suas informações é através da determinação da duração de cada fone constituinte da cadeia fonética de um enunciado.

Do ponto de vista acústico, não é fácil saber onde termina um fone e começa o seguinte ao analisar-se um enunciado. Isto deve-se ao fato de o sinal de fala ser contínuo. A ocorrência do fenômeno de coarticulação vem dificultar ainda mais o reconhecimento das fronteiras entre fones. Muitas vezes, sobretudo em encontros vocálicos, é praticamente impossível dizer onde acaba um som e começa o seguinte.

4.6.5 Energia

A energia presente no sinal acústico relaciona-se com a amplitude do sinal de fala. Esta, por sua vez, varia segundo a pressão do fluxo de ar vindo dos pulmões.

O correlato perceptual da energia chama-se intensidade. A intensidade é medida através do julgamento feito pelo ouvinte se um som é forte ou fraco. Este julgamento não é influenciado apenas pela amplitude do sinal, mas também, ainda que em menor grau, pela duração e pela freqüência fundamental.

4.6.6 Acentuação e ritmo

Ao ouvir-se o enunciado de uma palavra, geralmente percebe-se que uma de suas sílabas foi enfatizada em relação às demais. A esta ênfase dá-se o nome de acento e diz-se, com relação à sílaba que o recebe, que é acentuada.

O acento é produzido por um aumento dos valores dos parâmetros prosódicos durante a sílaba que o recebe (em relação às demais sílabas) e por uma melhor articulação dos fones que compõem a sílaba acentuada.

Existem dois tipos de acento: o acento lexical, que ocorre a nível de palavra, e o acento frasal que ocorre a nível de frase. Ao ouvir-se uma palavra pronunciada isoladamente, percebe-se que uma de suas sílabas recebeu acento (esta é a chamada sílaba tônica); trata-se do acento lexical que, na ortografia de uma palavra, pode ser indicado por um símbolo gráfico (por exemplo, acento circunflexo). Mas quando a palavra é pronunciada dentro de uma frase, o acento é dito ser frasal e não cai necessariamente sobre a mesma sílaba de sua pronúncia isolada, podendo mesmo não ocorrer acento sobre nenhuma das sílabas que constituem a palavra. Por exemplo, na pronúncia da frase "A menina é bela", o acento frasal sobre a palavra menina muitas vezes não recai na sílaba "ni" (onde acontece o acento lexical quando a palavra é pronunciada isoladamente) e sim na sílaba "me".

A existência, em um enunciado, de palavras que possuem, e de outras que não possuem sílaba com acento frasal, corresponde à intenção do falante de realçar certas palavras, em detrimento de outras, de modo a melhor transmitir para o ouvinte as informações que deseja. Trata-se de um recurso usado para colocar "em evidência" as palavras que carregam mais informações com a finalidade de tornar mais fácil para o ouvinte a sua compreensão.

O fato da fala receber acentos em certos pontos confere a ela uma certa musicalidade intrínseca. Esta musicalidade intrínseca consiste em uma cadência, denominada ritmo, que desempenha o importante papel de tornar a fala agradável de ser ouvida.

Na verdade, a fala não é constituída apenas por sílabas que recebem acento e por sílabas que não recebem nenhum acento. De fato, podem existir acentos secundários em palavras. Entretanto, a ocorrência de um acento secundário sobre determinada sílaba não faz dela uma sílaba tônica. Em uma mesma palavra pode haver apenas uma sílaba tônica. Por exemplo, na palavra "romana" a sílaba tônica é a segunda, mas a primeira sílaba pode receber um acento secundário.

Capítulo 5

Marcação de fronteiras prosódicas

Neste capítulo é apresentado o conceito de estrutura prosódica e descrito sua importância para o tratamento prosódico. Também é descrito como foi tratado, no presente trabalho, o problema de se determinar a estrutura prosódica de uma sentença escrita ao convertê-la em fala.

5.1 Estrutura prosódica

Conforme foi dito no capítulo anterior, a principal função da prosódia é ajudar o ouvinte a compreender a mensagem codificada foneticamente. Foi dito ainda que esta ajuda se dá por meio de caracterizações diferenciadas de diferentes trechos de um enunciado de modo que ele possa ser quebrado mentalmente em partes pelo ouvinte. Este processo de divisão em partes de um enunciado, por meio de variações apropriadas dos parâmetros prosódicos, confere a ele uma estruturação. Portanto, pode-se dizer que um enunciado possui uma estrutura prosódica [9].

A estrutura prosódica é composta por domínios prosódicos. Existem dois tipos de domínios prosódicos: frase fonológica e frase entoacional. Estes domínios formam um estrutura hierárquica em forma de árvore que denomina-se estrutura prosódica de sentença.

Conforme a referência [9], um domínio prosódico do tipo frase fonológica é construído ao redor de uma cabeça lexical, isto é, uma palavra de conteúdo. Palavras de conteúdo são aquelas que possuem significado mesmo quando isoladas, ou seja, ditas sem nenhum contexto. Podem ser substantivos, verbos, adjetivos ou advérbios. Um domínio do tipo frase fonológica inclui também especificadores da palavra de conteúdo conhecidos como palavras funcionais. Estes especificadores são palavras que não podem ser classificados como de conteúdo (preposições, conjunções, artigos, etc.).

Ainda conforme a referência [9], acima do nível dos domínios do tipo frase fonológica, estão os domínios prosódicos do tipo frase entoacional. Tais domínios são construídos pelo agrupamento de domínios do tipo frase fonológica que estão adjacentes. Portanto, um domínio

do tipo frase fonológica está sempre completamente incluído em um, e apenas um, domínio do tipo frase entoacional.

Na referência [9] é citado o exemplo transcrito abaixo, onde "##" indica uma fronteira entre domínios do tipo frase entoacional e "#" indica uma fronteira entre domínios do tipo frase fonológica: "## Kasypa's great war elephant # turned aside ## to avoid # a patch # of marshy ground ##.".

No presente trabalho foi criada uma simplificação da estrutura prosódica descrita acima. Tal simplificação baseou-se em observações de dados extraídos a partir da análise de enunciados, conforme é descrito no capítulo 3, e em testes em que procurou-se modelar a prosódia de frases para conversão texto-fala segundo diferentes maneiras de estruturação prosódica da sentença. Obteve-se, assim, uma estrutura prosódica simplificada.

A estrutura prosódica simplificada não faz distinção entre domínios do tipo frase fonológica e domínios do tipo frase entoacional. De fato, para ela existem apenas constituintes prosódicos que não podem estar incluídos uns nos outros.

Não existe uma correspondência direta entre os domínios da estrutura prosódica completa e os constituintes da estrutura simplificada. A grosso modo pode-se dizer que foram abolidos os domínios do tipo frase fonológica e poupados os domínios do tipo frase entoacional que, em certos casos, são divididos. Esta divisão se dá, por exemplo, entre grupos de palavras que sintaticamente correspondem a sujeito e predicado. O mesmo exemplo transcrito acima, a partir da referência [9], teria a seguinte estrutura prosódica simplificada: "# Kasypa's great war elephant # turned aside # to avoid a patch of marshy ground #.".

Mas o que são constituintes prosódicos? A grosso modo eles podem ser definidos como grupos de palavras adjacentes na sentença, onde cada grupo possui a propriedade de influenciar a evolução dos parâmetros prosódicos ao longo das palavras que o constituem. Por exemplo, a frase "As crianças de rua são o principal problema brasileiro" pode ser dividida em dois constituintes prosódicos: "As crianças de rua" e "são o principal problema brasileiro.". Outro exemplo pode ser a frase "Vieram bastante apressados até o terceiro estágio.", onde "Vieram bastante apressados" e "até o terceiro estágio" correspondem a constituintes prosódicos distintos. Os constituintes prosódicos, então, nada mais são do que o resultado da divisão que o falante faz de um enunciado, por meio de variações dos parâmetros prosódicos, para ajudar o ouvinte a compreender a mensagem que ele quer transmitir.

Os constituintes prosódicos de um enunciado tendem a ter o mesmo comprimento, comprimento esse que varia em função da fala ser mais ou menos rápida. Mas essa tendência a apresentarem o mesmo comprimento não pode ser classificada como forte. Trata-se mais de um objetivo secundário dos falantes para tornar a fala esteticamente mais agradável para os ouvintes. Esta tendência é facilmente negligenciada quando a estrutura sintática da sentença impõe um comportamente distinto.

5.2 Relação entre estrutura sintática e estrutura prosódica simplificada

Qualquer sentença produzida pela língua pode ser submetida a uma análise sintática, onde os constituintes sintáticos são identificados, obtendo-se, assim, a estrutura sintática da sentença. Os constituintes sintáticos são sujeitos, predicados, orações, complementos, etc.

De maneira análoga à análise sintática, pode-se submeter qualquer sentença a uma análise prosódica onde o objetivo é identificar os constituintes prosódicos da sentença que, juntos, compõem a sua estrutura prosódica simplificada.

Existe uma estreita relação entre estrutura sintática e estrutura prosódica. De maneira geral, pode-se dizer que cada constituinte prosódico é formado por um ou mais contituintes sintáticos adjacentes, embora seja possível haver parte de um constituinte sintático pertencendo a um constituinte prosódico e o restante pertencendo a outro. Por exemplo, na frase "O problema, infelizmente, da economia brasileira são os maus governantes" tem-se quatro constituintes prosódicos: "O problema", "infelizmente", "da economia brasileira" e "são os maus governantes". Do ponto de vista sintático, "O problema da economia brasileira" é um só constituinte (sujeito da frase), embora esteja dividido entre dois constituintes prosódicos: "O problema" e "da economia brasileira".

Apesar da relação entre as estruturas sintática e prosódica ser estreita, não existe uma correspondência biunívoca entre elas, ou seja, uma sentença que possui determinada estrutura sintática pode ter várias estruturas prosódicas possíveis.

Mas o quê, então, faz com que um falante escolha produzir um enunciado obedencendo a uma determinada estrutura prosódica simplificada e não a outra? A resposta é que esta

escolha depende do significado do que ele esteja dizendo (semântica) e do uso específico que ele esteja fazendo das palavras (pragmática).

5.3 Derivação da estrutura prosódica simplificada para modelamento prosódico na conversão texto-fala

Dentro do contexto do tratamento prosódico aplicado a um sistema de conversão texto-fala para texto irrestrito, surge a necessidade de se derivar a estrutura prosódica, a partir do texto, para produzir um enunciado. Esta necessidade pode ser facilmente justificada pela definição de constituinte prosódico dada acima. Nela afirma-se que um constituinte prosódico é um grupo de palavras adjacentes na sentença que tem a propriedade de influenciar a evolução dos parâmetros prosódicos ao longo das palavras que o constituem. Fica evidente, portanto, que para modelar satisfatoriamente a evolução dos parâmetros prosódicos (freqüência fundamental, duração e energia) ao longo de um enunciado sintetizado a partir de uma dada sentença escrita, é preciso primeiro determinar uma estrutura prosódica apropriada para ela. Rigorosamente falando, não se trata de derivar a estrutura prosódica simplificada referente a um texto, mas sim de determinar uma estruturação possível de ser realizada por um falante ao produzir o enunciado correspondente a este texto.

Optou-se por derivar a estrutura prosódica simplificada da sentença a partir apenas do conhecimento das classes gramaticais das palavras e dos sinais de pontuação. Esta opção deveu-se à dificuldade de obter-se informações semânticas e pragmáticas através de processamento automático de texto irrestrito.

No estágio atual, a delimitação dos contituintes prosódicos e, portanto, a determinação da estrutura prosódica simplificada é feita manualmente. Esta limitação deve-se a dois fatores. Primeiramente, no atual estágio do conversor texto-fala em desenvolvimento, ainda não se dispõe de meios automáticos para determinar as classes gramaticais das palavras constituintes das sentenças. Segundo, uma marcação prosódica manual é, em princípio, isenta de erros. Isto possibilita a concentração de esforços nas etapas posteriores do processamento prosódico. Como exemplo de marcação prosódica, podemos ter "A cotação do ouro # no mercado paralelo # sofreu um queda significativa # nos últimos dezesseis meses.", onde o símbolo "#" é usado para indicar fronteiras entre constituintes prosódicos adjacentes na sentença.

Atualmente, encontra-se em fase de desenvolvimento um programa de computador para realizar automaticamente, a partir apenas do conhecimento das classes gramaticais das palavras e dos sinais de pontuação, a determinação das marcas prosódicas que hoje são colocadas manualmente.

No modelo prosódico desenvolvido no presente trabalho, a identificação dos constituintes prosódicos é feita por meio da determinação das fronteiras entre eles. Tais fronteiras prosódicas são atribuídas a pontos que coincidem com fronteiras entre constituintes sintáticos.

Capítulo 6

Modelo de duração

Neste capítulo, procura-se inicialmente explicar os fatores que determinam qual duração cada fone deve possuir. Em seguida são feitas algumas considerações sobre o modelo construído e, finalmente, descreve-se a estrutura do modelo de duração desenvolvido no presente trabalho.

6.1 Aspectos gerais do comportamento duracional dos segmentos

Conforme foi dito no capítulo 4, o termo duração refere-se ao intervalo de tempo que vai desde o início de um fone até o seu término na fala. Foi dito ainda que uma das maneiras pela qual o nível suprasegmental codifica suas informações é através da determinação da duração de cada fone constituinte da cadeia fonética de um enunciado.

São vários os fatores que determinam qual duração cada fone deve ter em um enunciado. Estes fatores podem ser divididos em três grandes grupos: os de natureza segmental, os de natureza coarticulatória e os de natureza suprasegmental, ou seja, prosódica.

Os fatores de natureza segmental referem-se à natureza acústico-articulatória do segmento, ou seja, do fone. Diferentes fonemas tendem a ter diferentes durações em suas realizações fonéticas. Isto significa que as durações que porventura se observem nos fones de um enunciado serão em parte devidas a suas respectivas naturezas fonêmicas. Assim, pode-se dizer que cada fonema tem uma duração inerente ou intrínseca.

Os fatores de natureza coarticulatória referem-se ao fato da duração de um fone ser influenciada tanto pelo contexto fonético que o segue no tempo como pelo contexto que o precede. Isto significa que as durações que os fones possuem nos enunciados são resultado também de fenômenos de coarticulação. Note que seria possível, então, falar a respeito de uma duração contextual inerente a cada fonema, em que não mais haveria a idéia de um fonema possuir uma duração intrínseca, mas sim de cada fonema possuir uma duração inerente a si em função de onde (contexto fonético) ele ocorre.

Os fatores de natureza suprasegmental referem-se a como a prosódia utiliza o parâmetro duração para exercer suas funções. Todas as funções que a prosódia desempenha na fala envolvem variações de todos os parâmetros prosódicos (freqüência fundamental, duração e energia). Isto significa que o conceito de duração inerente, ou de duração contextual inerente, perde o sentido a menos que seja estendido para duração contextual prosódica. Ou seja, pode-se falar que um fonema possua uma duração conjuntamente inerente a sua natureza, ao contexto fonético e ao efeito prosódico buscado no instante em que ele ocorra no enunciado. Por efeito prosódico buscado quando da ocorrência do fonema, pode-se imaginar, por exemplo, que esta ocorrência se dê ao final de um constituinte prosódico. Neste exemplo, pode-se ter uma duração relativamente grande para o fonema com a finalidade de ajudar a demarcar a fronteira entre o final do constituinte prosódico e o início do próximo constituinte.

6.2 Considerações sobre o modelo duracional construído

Por modelo duracional aplicado à sintese de fala entende-se qualquer tratamento automático pelo qual as durações dos fones de um enunciado a ser sintetizado possam ser determinadas.

Um modelamento de duração será tão mais eficiente quanto mais próximas sejam as durações determinadas pelo modelo, para a síntese de um enunciado, das respectivas durações dos fones constituintes do mesmo enunciado quando dito por um falante humano. Logo, o caminho natural para criar um modelo de duração é baseá-lo nas durações que um falante humano dá aos fones em sua fala.

No presente trabalho, optou-se por construir um modelo duracional formado por regras. O processo de estabelecimento das regras deste modelo consistiu numa escolha de um subconjunto das regras de um modelo semelhante, porém construído para a língua inglesa por Klatt [1]. Foram escolhidas regras que foneticamente pareciam pertinentes também à língua portuguesa. Também foram criadas outras regras específicas para o português.

Para determinar a duração dos fones nas sentenças, o melhor caminho seria realizar um extenso estudo descritivo. Este estudo, porém, seria muito complexo e demorado, o que tornaria sua realização impraticável dentro dos limites temporais a que ficou restrito o presente trabalho. Por esta razão, optou-se por uma abordagem mais simples em que busca-se estudar

os efeitos duracionais que carregam informação lingüística. Esta é a metodologia utilizada por Klatt.

Basear a construção do modelo duracional no modelo que Klatt desenvolveu para a Língua Inglesa foi uma decisão tomada principalmente em virtude da simplicidade de seu modelo. Outra razão que levou a esta decisão foi a existência de um estudo realizado por Simões [10] para as vogais {*a, i, u*} do Português brasileiro que baseou-se no trabalho de Klatt.

Segundo Egashira [3], o modelo desenvolvido por Klatt baseia-se nas seguintes suposições:

a) Cada segmento possui uma duração intrínseca. Esta duração intrínseca corresponde ao valor médio da distribuição de valores que a duração daquele segmento pode assumir.

b) Cada regra tenta prever uma variação percentual a fim de efetuar um aumento ou diminuição da duração do segmento.

c) Os segmentos não podem ser reduzidos a valores menores do que uma certa duração mínima.

O modelo pode ser expresso por:

$$Do = Dmin + K \times (Di - Dmin)$$

onde

Do é a duração prevista em qualquer ponto no texto

Di é a duração intrínseca de cada segmento

Dmin é a duração mínima que cada segmento pode assumir

K é o fator que produz a variação da duração de acordo com a aplicação das regras.

O primeiro passo para a aplicação do modelo é a obtenção da duração intrínseca do segmento em questão. Em seguida aplicam-se as regras para variação da duração.

O modelo duracional construído no presente trabalho possui regras que buscam determinar a duração de cada segmento com base em sua natureza, no seu contexto fonético e no efeito prosódico que se queira produzir quando de sua ocorrência no enunciado, ou seja, o modelo busca contemplar os fatores de natureza segmental, coarticulatória e prosódica.

6.3 Descrição do modelo duracional

Como já foi dito, o modelo duracional foi construído por meio de regras que determinam a duração de cada fone no momento da síntese. A escolha de criar um modelo baseado em regras deve-se ao fato desta abordagem não exigir um conjunto de dados de grande dimensão.

Foi criado um modelo cujas regras atuam de forma independente entre si. Esta atuação se dá por meio de um produtório de coeficientes, onde cada regra entra com seu fator. Para cada fone de um enunciado a ser sintetizado é calculado o produtório dos coeficientes determinados pelas regras que incidem sobre ele. O valor obtido para este produtório multiplicado pela "duração média do fone" resulta no valor de duração a ser utilizado em sua síntese.

O modelo pode ser expresso por:

$$D = Dm \times K$$

onde

D é a duração calculada para o fone

Dm é a duração média do fone

K é o valor resultante do produtório de coeficientes.

Mas o que vem a ser essa "duração média do fone"? Para responder a essa pergunta é necessário lembrar que o presente trabalho desenvolveu-se sobre um sintetizador concatenativo. Neste tipo de síntetizador é preciso extrair as unidades concatenativas de realizações de fala natural. Essas unidades para concatenação são constituídas por fones cujos limites devem ser precisados. Uma vez que se disponha das durações de um conjunto de fones para cada fonema, pode-se, por simples média aritmética, determinar a duração média de cada fonema. É interessante observar que a duração média de um fone, como foi obtida neste trabalho, corresponde a uma aproximação da duração intrínseca tal como é definida por Klatt.

Deve-se observar adicionalmente que foram estipulados valores máximos e mínimos para a duração de cada fone. Para estipular estes valores observou-se as durações geradas para um conjunto de frases sintetizadas. As durações de segmentos cuja audição mostrou-se

desagradável serviram como referencial para o estabelecimento de limites duracionais. A partir deste referencial, a estipulação de valores exatos para os limites foi feita por tentativa e erro. Se, por exemplo, o valor de duração calculado para um fone ultrapassar a duração máxima estipulada para ele, este valor de duração passará a ser igual ao valor máximo estipulado. A limitação dos valores de duração para cada fone objetiva eliminar possíveis distorções geradas pelo modelo e também leva em conta a limitação da técnica de síntese PSOLA em processar variações elevadas de duração, principalmente em segmentos não sonoros.

Note que há uma discrepância entre o modelo apresentado neste trabalho e o modelo de Klatt no que se refere à existência de um limite máximo para a duração de cada segmento. O fato do modelo de Klatt não impor uma duração máxima a cada segmento é um indício de que os resultados obtidos por seu modelo não apresentam distorções a serem corrigidas.

Uma vez estabelecido um conjunto de regras para o modelo duracional, os valores dos coeficientes relativos a elas foram determinados principalmente através de sucessivos ajustes orientados pela percepção de sentenças sintetizadas e, em menor grau, pela comparação das durações obtidas com o modelo com as obtidas pela análise do corpus.

O método de determinação dos valores dos coeficientes, por comparação com os dados obtidos pela análise do corpus, apesar de aparentemente ser o mais eficaz, apresenta um problema. A velocidade da fala (expressa, por exemplo, pelo número de palavras ditas por minuto) varia significativamente de um enunciado para o outro. Uma possível solução para este problema seria a normalização da velocidade de fala através da multiplicação das durações de todos os enunciados analisados por fatores que tornem suas velocidade próximas entre si. Como o emprego deste recurso poderia causar distorções e certamente seria um elemento a mais para aumentar a complexidade do modelamento duracional, optou-se por não empregá-lo na realização do presente trabalho.

As regras atuam segundo a estrutura prosódica de cada sentença e de maneira hierárquica. Existem regras que atuam a nível de sentença, isto é, influenciando os fones da sentença. Existem outras regras que atuam a nível de constituinte prosódico de modo a influenciar os fones do constituinte. Outras agem a nível de palavra, de sílaba e, por último, de fones.

Os coeficientes de algumas regras são expressos por uma faixa de valores e não por valores exatos. Isto deve-se ao fato de que qualquer opção por um valor dentro de uma faixa

seria aceitável. A escolha de valores ótimos implicaria na realização de testes de avaliação subjetiva que ainda não foram realizados.

A seguir são apresentadas as regras que compõem o modelo duracional construído.

Nível de sentença:

1) Uma pausa é inserida entre dois constituintes prosódicos adjacentes.

Esta regra deve-se ao fato de pausas serem importantes mecanismos empregados pelos falantes para estabelecer fronteiras entre constituintes prosódicos. A duração da pausa é, de certo modo, arbitrária, pois podem ser usados valores maiores ou menores conforme for pretendida uma fala mais lenta ou mais rápida. Para uma fala não muito lenta, e nem muito rápida, pode-se usar valores entre 100 e 300 ms.

2) Quando uma consoante é seguida por outra consoante, a sua duração é reduzida. O valor do coeficiente desta regra é de 0.79.

3) Quando uma consoante é precedida por outra consoante, a sua duração é reduzida. O valor do coeficiente desta regra é de 0.83.

4) São aumentadas as durações dos fones constituintes de palavras que recebem ênfase em função do significado. O valor do coeficiente desta regra é de 1.4.

Esta regra ainda não atua no modelo duracional em virtude de ainda não se poder identificar as palavras enfatizadas.

Como exemplo de palavra que pode receber ênfase em função do significado, pode-se citar a palavra **corruptos,** presente no seguinte enunciado: "Brasília está cheia de corruptos".

5) Os fones que formam sentenças curtas (com menos de oito sílabas) têm suas durações aumentadas. O valor do coeficiente desta regra pode variar de 1.1 até 1.4.

Esta regra visa modelar o fato de falantes tenderem a desacelerar a fala durante sentenças curtas.

Nível de constituinte prosódico:

6) O primeiro fone de um constituinte prosódico tem sua duração aumentada. O valor do coeficiente desta regra pode variar de 1 até 1.3.

Aparentemente, as pessoas falam mais lentamente no início de um enunciado para evitar que o ouvinte perca informação pelo fato de estar distraído.

Nível de palavra:

7) Os fones constituintes de uma palavra têm suas durações alteradas conforme o número de sílabas da palavra. A tabela 6.1 expressa os valores de coeficientes para esta regra.

Esta regra visa modelar o fato de falantes tenderem a acelerar a fala durante palavras longas.

Número de sílabas	Coeficiente	Número de sílabas	Coeficiente
1	1.2	5	0.92
2	1	6	0.9
3	0.97	7	0.9
4	0.94	mais de 7	0.85

Tabela 6.1

8) As consoantes que não iniciam palavra têm suas durações reduzidas. O valor do coeficiente desta regra é de 0.92.

9) Os fones constituintes de palavras funcionais têm suas durações reduzidas. O valor do coeficiente desta regra é de 0.87.

Esta regra deve-se à tendência que falantes possuem de colocar as palavras funcionais num plano secundário em relação às palavras de conteúdo.

10) Os fones constituintes de palavras de conteúdo têm suas durações aumentadas. O valor do coeficiente desta regra pode variar de 1.02 até 1.2.

Esta regra deve-se à tendência que falantes possuem de ressaltar as palavras de conteúdo em relação às funcionais.

11) Uma vogal possui sua duração alterada em função da natureza do fone que a sucede, ou seja, do seu contexto fonético direito. A tabela 6.2 expressa os valores de coeficientes para esta regra.

contexto fônico direito	coeficiente
fricativa	1.05
oclusiva	1.05
nasal	0.7

Tabela 6.2

12) Os fones constituintes de sílabas pré-tônicas têm suas durações reduzidas. O valor do coeficiente desta regra pode variar de 0.8 até 0.98.

13) Os fones constituintes de sílabas tônicas têm suas durações aumentadas. O valor do coeficiente desta regra pode variar de 1.1 até 1.6.

14) Os fones constituintes de sílabas postônicas têm suas durações reduzidas. O valor do coeficiente desta regra pode variar de 0.7 até 0.95.

Nível de sílaba:

15) Quando uma vogal é seguida por outra vogal, a sua duração é reduzida. O valor do coeficiente desta regra é de 0.82.

16) Quando uma vogal é precedida por outra vogal, a sua duração é reduzida. O valor do coeficiente desta regra é de 0.86.

17) Se uma vogal é precedida por uma plosiva, a duração da vogal deve ser aumentada. O valor do coeficiente desta regra pode variar de 1.05 até 1.25.

Nível de fone:

18) O limite para aumento da duração de uma vogal será um fator igual a 2.

19) O limite para diminuição da duração de uma vogal será um fator igual a 0.5.

20) O limite para aumento da duração de uma consoante será um fator igual a 1.8.

21) O limite para diminuição da duração de uma consoante será um fator igual a 0.6.

22) A duração do fonema RX (prato, cretino, crase, etc.) não será alterada por nenhuma das regras.

A regra de número 8 foi obtida diretamente do modelo de Klatt. Já as regras de números 1, 2, 3, 7, 11, 15, 16, 19 e 21 correspondem a adaptações feitas sobre regras do modelo de Klatt. As demais regras são originais.

A **Tabela 6.3** ilustra o resultado da aplicação do modelo duracional sobre a sentença: "O preço da tarifa telefônica foi reduzido.". Os símbolos que constam na coluna "Fone" correspondem à representação fonética da sentença.

Fone	Duraçao (ms)	Fone	Duração (ms)	Fone	Duração (ms)	Fone	Duração (ms)
UW	52	AA	58	FF	157	IY	108
PP	132	RX	40	OO	101	RR	62
RX	40	II	125	NN	42	EE	79
EE	138	FF	67	IY	56	DD	55
SS	118	AA	48	KK	72	UW	63
UW	55	TT	76	AA	69	ZZ	110
DD	57	EE	82	##	100	II	170
AA	65	LL	37	FF	174	DD	51
TT	58	EE	78	OO	154	UW	56

Tabela 6.3

Capítulo 7
Modelo entoacional

Neste capítulo, são feitas considerações sobre a problemática do modelamento de F0 para a conversão texto-fala e sobre as tendências gerais do comportamento da curva de freqüência fundamental ao longo de enunciados declarativos. Em seguida, é feita uma descrição da estrutura do modelo de controle entoacional que foi desenvolvido no presente trabalho.

7.1 Porque é complicada a tarefa de modelar entonação para um sistema de conversão texto-fala.

A maior dificuldade em converter um texto para fala talvez seja a determinação de uma entonação semelhante à que se observaria no enunciado que um falante humano produziria a partir deste texto. Ou seja, modelar corretamente a entonação com que uma pessoa produz fala é um problema de grande complexidade.

A complexidade do modelamento entonacional deve-se principalmente à necessidade de se determinar uma curva de freqüência com base apenas em um texto escrito. Para que se entenda o porquê disto ser um problema, basta observar como as pessoas atribuem curvas de F0 a seus enunciados. Imagine a frase " _ O Fluminense ganhou a decisão no Maracanã." sendo pronuncida por dois falantes. O primeiro deles é um apaixonado torcedor do Flamengo e o segundo é um também apaixonado torcedor, porém, do Fluminense. Seriam observados contornos semelhantes de freqüência fundamental caso fossem analisados os enunciados de ambos? A resposta é que seria bastante difícil encontrar semelhança entre as duas curvas. Mas porquê acontece isto se a frase dita por eles é a mesma? Apesar da frase, ou seja, do texto ser o mesmo, os falantes produzem curvas de entonação bastante distintas porque o "modelamento entoacional" com que as pessoas falam leva em conta o significado daquilo que estão falando. Mais do que isso, ele leva em conta a maneira com que o falante relaciona-se com a realidade expressa através da fala (alegria, decepção, entusiasmo, ...), informações estas de que ainda não se dispõe para o modelamento entoacional.

Considerando que um modelo de controle entoacional para a converão texto-fala deve produzir um contorno de F0 sem dispor das mesmas informações que as pessoas usam para "modelar" sua entonação, surge uma questão: o que fazer para suprir esta deficiência? Basicamente, existem duas alternativas. A primeira é olhar de frente o problema e tentar construir algoritmos para obter as informações que não ficam explicitadas na forma de texto. Isto seria indubitavelmente uma ótima opção caso atualmente fosse possível extrair informação semântica de texto irrestrito. Como isto em geral ainda não é possível, sobra apenas a segunda alternativa, que é de tentar contornar o problema por meio de um artifício. Este artifício consiste em buscar uma maneira neutra de produzir contorno entoacional. Para compreender o que seria esta maneira neutra, imagine um terceiro falante dizendo a frase "_ O Fluminense ganhou a decisão no Maracanã.". Imagine também que este terceiro falante não seja torcedor de nenhuma das equipes, que esteja se referindo a uma disputa hipotética e nem sequer tenha grande interesse por futebol. Não haveria nenhuma emoção na sua forma de falar, ele simplesmente estaria dizendo algo. Sua fala apresentaria , portanto, uma neutralidade bem maior do que a observada nos enunciados produzidos pelos dois primeiros falantes (torcedores).

Além do que foi exposto acima, existe uma forte razão para o modelamento entoacional ser um problema complexo. Esta razão é o fato dos contornos de freqüência fundamental serem dependentes do idioma. De fato, não se pode pretender que um modelo construído para uma língua funcione bem quando aplicado a uma outra. Enquanto que a duração dos fones é altamente dependente da natureza segmental destes, a freqüência fundamental apresenta um comportamento que evidencia um grau de liberdade bem maior em relação ao nível segmental. Isto não significa que um mesmo modelo duracional possa ser aplicado a diferentes idiomas sem que haja uma degradação na qualidade. Mas significa que esta degradação tenderia a ser menor do que a que seria observada caso se tentasse aplicar o mesmo conjunto de regras para modelar a entonação de um outro idioma além daquele para o qual as regras foram elaboradas.

Ainda fazendo um paralelo entre o modelamento duracional e o de entonação, pode-se dizer que a neutralidade, imposta pela impossibilidade atual de se extrair as mesmas informações de natureza semântica que uma pessoa usa para "modelar sua prosódia", causa maiores transtornos a quem dedica-se à tarefa de modelar contornos de freqüência fundamental

do que para quem constrói um modelo de duração. Isto acontece porque a freqüência fundamental é bem menos condicionada à natureza dos fonemas do que a duração.

7.2 Tendências gerais das curvas entoacionais de sentenças declarativas

O padrão de comportamento da curva de freqüência fundamental, ao longo de uma sentença declarativa, é tipicamente um declínio gradual. Esta tendência ao declínio gradual é tão mais verdadeira quanto mais neutra for a sentença.

Apesar da tendência global ser de decrescer, a curva de freqüência fundamental apresenta, a nível de sílabas e fones, tanto trechos onde sua derivada primeira é negativa quanto trechos onde esta derivada é positiva. Os trechos onde a primeira derivada da curva é positiva correspondem geralmente a sílabas acentuadas relativamente a suas vizinhas.

Também a nível de tendência geral de sentenças declarativas, pode-se afirmar que os vales e picos da curva de freqüência fundamental vão assumindo, ao longo de um enunciado, valores cada vez mais próximos entre si. Isto significa que a variação da curva de freqüência fundamental tende a se tornar menor ao longo do enunciado.

Outra tendência geral que se observa é que o máximo valor de freqüência fundamental dentro de um enunciado ocorra sobre a sílaba tônica de sua primeira palavra de conteúdo (substantivo, adjetivo, verbo, advérbio ou numeral).

A figura 3.1 do capítulo 3 ilustra essas tendências ao mostrar a curva natural de F0 de uma sentença declarativa.

7.3 Considerações sobre o modelo entoacional construído

Um modelo entoacional aplicado à síntese de fala deve prover um tratamento automático pelo qual sejam determinadas curvas de freqüência fundamental para os fones de um enunciado a ser sintetizado, ou seja, ele deve prover uma curva de F0 para cada fone, de modo que a junção dessas pequenas curvas corresponda ao contorno entoacional do enunciado.

Um modelamento de entonação será tão mais eficiente quanto mais se aproxime do comportamento que uma curva entoacional produzida por um falante humano apresente. Logo, o caminho natural para criar um modelo entoacional é baseá-lo no comportameto que a frequência fundamental tem ao longo da fala de uma pessoa. Esta pessoa preferencialmente deve ser a mesma usada para a gravação do dicionário de unidades acústicas caso esteja-se usando um sistema de síntese por concatenação para o teste das regras, que é o caso do presente trabalho. Assim, procurou-se analisar enunciados produzidos pela mesma pessoa cuja fala foi utilizada para a construção do dicionário de unidades acústicas.

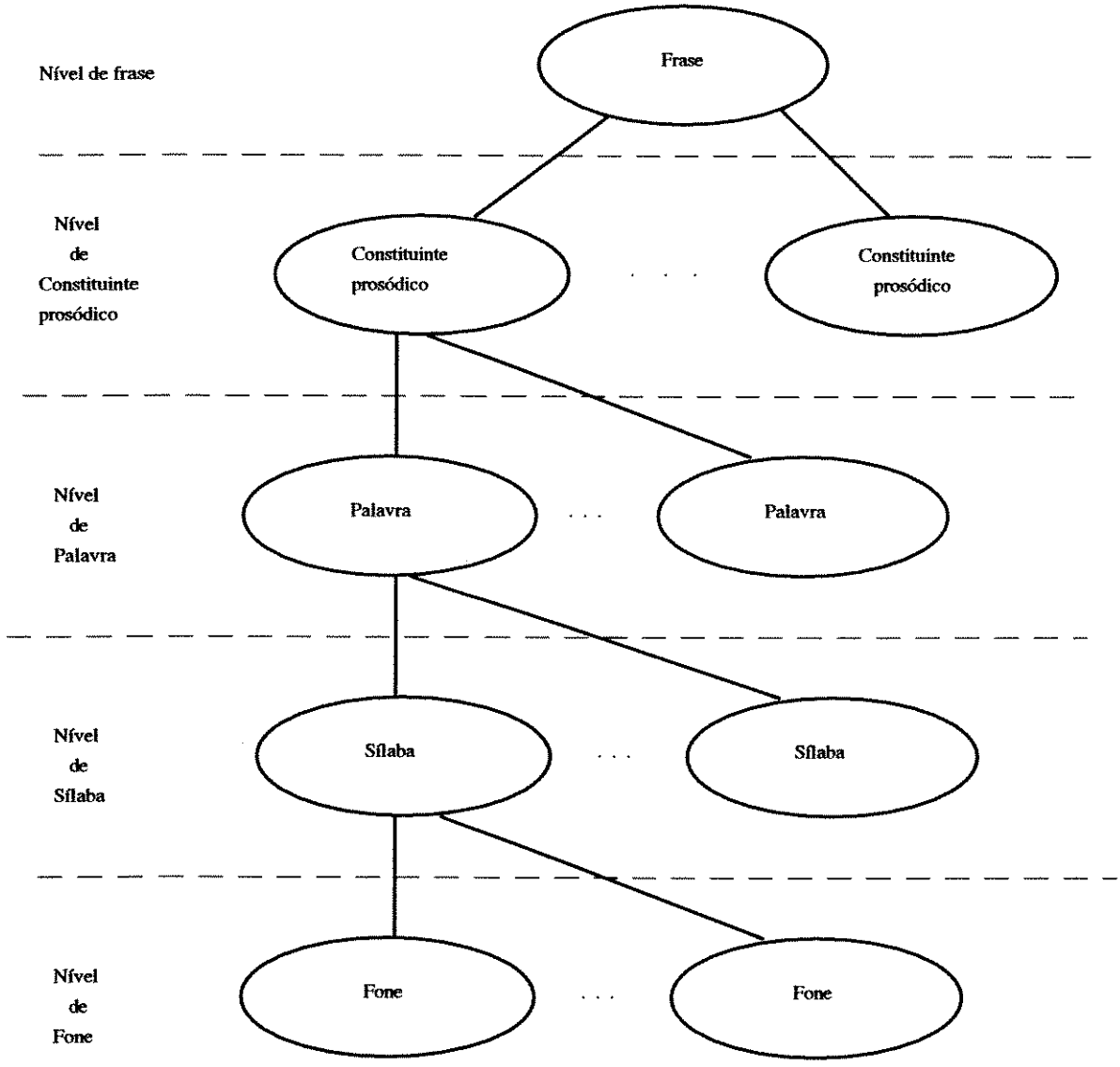


Fig 7.1 - Árvore de modelamento entoacional.

Para a análise da entonação na fala natural, foram utilizados dois locutores, sendo que para cada um deles dispunha-se de um dicionário de unidades para concatenação. Analisou-se 78 enunciados de um locutor e 86 enunciados do outro, conforme é descrito no capítulo 3.

Uma vez que foram extraídos dados sobre a fala das pessoas escolhidas, foi necessário criar um mecanismo para, através dele, determinar as curvas de freqüência fundamental ao longo dos enunciados que se quisesse sintetisar. Este mecanismo é o modelo entoacional criado.

Semelhantemente ao modelo duracional, o modelo entoacional foi construído por meio de regras que, baseadas em análise de dados coletados, determinam uma curva de freqüência fundamental para cada fone constituinte do enunciado a ser sintetisado.

A construção de um modelo baseado em regras requer um volume de dados relativamente modesto, sendo esta a principal causa da escolha desta abordagem. Conforme foi explicado no capítulo 3, determinar manualmente onde começa e onde termina cada fone de um enunciado, por meio de análise acústica do mesmo, é uma tarefa bastante lenta e cansativa.

Foi criado um modelo entoacional cuja principal característica é basear-se em uma estrutura hierárquica para representação de sentenças (Fig. 7.1). Essa forma de representar uma sentença possui relação estreita com a noção de estrutura prosódica simplificada que foi apresentada no capítulo 5. Por esta abordagem, cada nível hierárquico deve obedecer às determinações do nível superior e, por sua vez, gerar determinações para o nível inferior. Para cada enunciado que se queira sintetizar é derivada uma estrutura hierárquica em forma de árvore. A nível de frase, é definido, segundo a modalidade da mesma, um comportamento macroscópico para a freqüência fundamental. Lembre-se, entretanto, que o presente trabalho limitou-se a frases declarativas neutras.

A decisão tomada no presente trabalho de modelar o contorno de freqüência fundamental por meio de uma abordagem em que os níveis hierárquicos inferiores obedecem a determinações dos níveis superiores, deve-se ao fato do contorno de F0 poder ser visto como uma sobreposição de efeitos. Sobreposição esta que vai desde o nível global (a sentença) até o nível local (o fonema), passando sucessivamente pelo níveis de constituinte prosódico, palavra e sílaba.

7.4 Descrição do modelo entoacional

Para modelar o comportamento macroscópico da freqüência fundamental, a nível de frase, recorreu-se ao estabelecimento de estruturas limitantes das variações de F0 a nível dos constituintes prosódicos. Tais estruturas nada mais são do que retas. Para cada constituinte prosódico são criadas duas retas. Toda a curva de freqüência fundamental dentro de um constituinte prosódico deve ficar entre os gráficos destas retas.

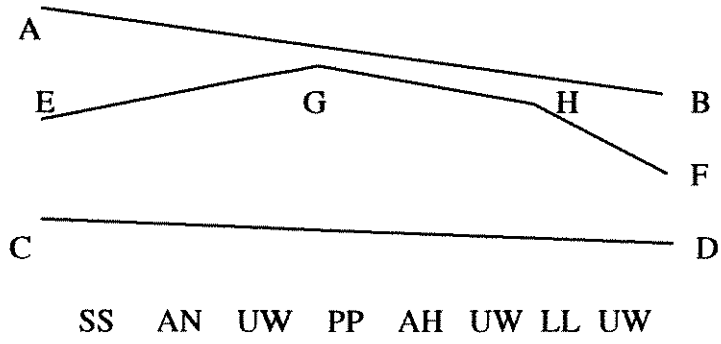


Fig. 7.2 - Esquema do modelamento entoacional para o enunciado de: "_São Paulo".

Cada uma destas retas, superior e inferior, é definida por meio da especificação de dois valores: um valor de F0 para o início do constituinte prosódico e outro valor de F0 para o final do constituinte. Assim, é necessário estabelecer inicialmente quatro valores para cada constituinte prosódico, que são os valores iniciais e finais das retas inferior e superior. Estes valores (correspondentes aos pontos A, B, C e D da figura 7.2) são determinados por meio de funções lineares que obedecem à fórmula F = b + a * n, onde "F" é o valor no extremo de uma linha, "n" representa o número de sílabas do constituinte prosódico, "b" e "a" são parâmetros, obtidos através da análise de contornos naturais de F0. Desse modo, existem quatro funções que conjuntamente determinam as retas entre as quais deve ficar toda a curva de frequência fundamental dentro de um constituinte prosódico. A figura 7.2 ilustra as retas (**AB** e **CD**) geradas para um enunciado composto de um único constituinte prosódico: "_São Paulo".

Também existem duas funções lineares que determinam, a partir do número de sílabas do constituinte prosódico, os valores que a freqüência fundamental deve ter ao início e ao final do constituinte. Na figura 7.2, **E** corresponde ao valor inicial e **F** ao valor final da freqüência fundamental.

No caso da figura 7.2, os pontos A,B,C, D,E e F são dados pelas fórmulas A = 145 +0 * n, B = 106 +0 * n, C = 125 +0 * n, D = 90 +0 * n, E = 135 +0 * n e F = 95 +0 * n, onde n é o número de sílabas, que é igual a 3.

Para o caso de um enunciado com mais de um constituinte prosódico, o processo de determinação dos valores correspondentes aos pontos A, B, C, D, E e F, para cada constituinte, é semelhante. Porém, para cada natureza de constituinte prosódico, as constantes "a" e "b" (que definem as funções que determinam os pontos A, B, C, D, E e F) são distintas, sendo que a natureza de um constituinte prosódico é dada pelo tipo das marcas prosódicas entre as quais ele está. Os tipos de marcas prosódicas são : início de frase, final de frase, início de predicado, início de oração e início de complemento com preposição. Por exemplo, na frase "_Caminhávamos devagar com a certeza dos indecisos", antes da preposição "com" deve haver uma marca prosódica indicando início de complemento com preposição.

Ponto	início de frase		final de frase		início de predicado		início de oração		início de complemento	
	a	b	a	b	a	b	a	b	a	b
A	0	145	--	--	2.3	128	1.8	130	1.2	130
B	--	--	0	106	-1.3	142	-2.1	138	-2.8	140
C	0	125	--	--	2	106	1.5	105	1.2	105
D	--	--	0	90	-1.1	125	-0.8	130	-2.1	110
E	0	135	--	--	2.1	120	1.5	125	1.3	122
F	--	--	0	95	-1	130	-2.5	136	-2	132

Tabela 7.1

A cada marca prosódica são associados seis pares de "a" e "b", sendo que três são usados para determinar os pontos B, F e D do constituinte que está à esquerda da marca e os outros três são usados para determinar os pontos A, E e C do constituinte que está à direita da marca prosódica. Evidentemente, as marcas que indicam início e final de frase possuem apenas três pares "a" e "b" associados a cada uma. Ao todo, portanto, existem 24 pares de "a" e "b".

A tabela 7.1 mostra os valores de "a" e "b" para os pontos A, B, C, D, E e F de cada tipo de marca prosódica.

Para que não haja descontinuidade na curva de F0, o ponto F de um constituinte prosódico e o ponto E do constituinte seguinte devem sempre coincidir. Não havendo coincidência direta (isto é, pelo simples uso das funções lineares), ela é forçada fazendo-se com que ambos sejam iguais à média aritmética de seus valores.

Dentro de um constituinte prosódico são determinados os valores de F0 entre palavras consecutivas, ou seja, é determinado o valor de freqüência fundamental com que é finalizada uma palavra e iniciada a seguinte. O ponto G da figura 7.2 ilustra o valor de F0 determinado para a fronteira entre as duas palavras que constituem o enunciado.

Como toda a curva de frequência fundamental dentro de um constituinte prosódico deve ficar entre as retas, fica evidente que a curva de freqüência fundamental de cada palavra deve ficar entre as retas estabelecidas para o constituinte prosódico a que ela pertence.

Entre uma palavra e a palavra seguinte, o valor da freqüência fundamental é determinado por meio de um percentual da distância que separa a linha superior da inferior. Esta percentual é estabelecido levando-se em conta a tonicidade das duas palavras, isto é, examina-se se a última sílaba da primeira palavra é tônica ou átona. O mesmo é feito para a primeira sílaba da segunda palavra. Este exame resulta em quatro combinações possíveis. Para cada uma delas foi determinado um percentual conforme ilustra a tabela 7.2.

Tonicidade da última sílaba da palavra i	Tonicidade da primeira sílaba da palavra i + 1	Percentual entre as funções lineares
átona	átona	20
átona	tônica	40
tônica	átona	70
tônica	tônica	90

Tabela 7.2

No caso do exemplo da figura 7.2, "SS AN UW" e "PP AH UW" são tônicas, e portanto o ponto G está a 90% da distância entre as duas retas.

Dentro de uma palavra são determinados os valores de freqüência fundamental entre suas sílabas, isto é, determina-se o valor de F0 ao final e ao início de cada sílaba que constitui a palavra, com exceção dos valores de F0 para o início da primeira sílaba e para o final da última sílaba, porque foram determinados anteriormente a nível de constituinte prosódico. Os valores de freqüência fundamental entre sílabas da mesma palavra são determinados de maneira análoga à determinação de F0 entre palavras, ou seja, por meio de percentuais da distância entre as linhas superior e inferior do constituinte prosódico. Novamente leva-se em conta o caráter átono/tônico das sílabas, mas também é levado em conta o número de sílabas da palavra. Assim, por exemplo, se a estrutura de uma palavra consiste de uma sílaba átona seguida de uma tônica que por sua vez é seguida por outra sílaba átona, tem-se: A(20)T(75)A, onde "A"representa uma sílaba tônica, "T" representa uma sílaba átona e os números entre parênteses são os valores dos percentuais entre as sílabas. Foram determinados percentuais entre sílabas para palavras de até quatro sílabas, com as suas possíveis estruturas de tonicidade:

A(10)T, T(80)A, A(40)A(55)T, A(20)T(75)A, T(85)A(25)A, A(15)A(45)A(10)T, A(30)A(50)T(95)A, A(65)T(95)A(30)A e T(80)A(55)A(15)A.

Para palavras com mais de quatro sílabas generalizou-se os percentuais. Esta generalização deve-se à pequena percentagem de palavras longas no corpus analisado e à predominância de palavras curtas na Língua Portuguesa. Evidentemente, perde-se qualidade por fazer-se esta generalização. No exemplo da figura 7.2, "PP AH UW" é tônica e "LL UW" é átona, e portanto o ponto H está a 80% da distância entre as retas.

De maneira análoga ao que ocorre nos níveis acima, dentro de uma sílaba são determinados os valores de F0 entre os seus fones constituintes. Isto é feito traçando-se uma reta que vai do início da sílaba até o final dela. Neste estágio, cada fone já tem definidos os valores de freqüência fundamental para o início e para o final de sua curva. Além de ter as funções lineares para orientar o traçado da curva de F0 ao longo do fone. Na figura 7.2, os segmentos de reta **EG**, **GH** e **HF** ilustram as retas que ligam o valor de F0 no início ao valor de F0 no final de cada sílaba.

Denomina-se microprosódia o formato que a curva de F0 assume internamente a cada fone. Sendo este um trabalho inicial na área de prosódia, optou-se por deixar para trabalhos posteriores um tratamento mais sofisticado de microprosódia. No presente modelo, os valores de F0 inicial e terminal de cada fone são simplesmente interpolados de maneira linear. Daí resulta que, no modelo implementado, a curva de F0 é uma curva contínua composta por uma sucessão de segmentos de reta, onde cada segmento corresponde a uma sílaba. Foi verificado que esta limitação representa, do ponto de vista do ouvinte, uma degradação bastante pequena.

Foram realizados testes em que contornos de freqüência fundamental obtidos pela análise de enunciados naturais, isto é, ditos por pessoas, foram atribuídos a enunciados sintetizados com o emprego da técnica PSOLA. Ou seja, emulou-se um modelamento prosódico "perfeito" no sentido de ser igual ao de uma pessoa. Nestes testes, as curvas de freqüência fundamental para cada fone foram gradativamente estilizadas, por meio de composições de segmentos de reta, até que cada fone recebesse uma curva entoacional na forma de apenas um segmento de reta. Os enunciados sintetisados com as curvas em diferentes graus de estilização foram submetidos à apreciação de ouvintes que julgaram pouco relevantes as diferenças.

Outros testes foram feitos utilizando-se a ferramenta de software descrita no capítulo 3. Nestes testes, procurou-se estilizar gradativamente a curva de freqüência fundamental, interna a cada fone, extraída de um enunciado natural. Para cada grau de estilização foi realizada uma reprodução do sinal através de vocoder LPC. Novamente, os ouvintes julgaram pouco relevantes as diferenças. Mas como a qualidade do vocoder LPC não é muito boa, talvez tenha sido difícil para os ouvintes apreciar as diferenças.

Conforme foi dito no capítulo 3, analisou-se enunciados de dois falantes, sendo que um deles é um locutor profissional enquanto que o outro é um falante comum. Obteve-se desta forma um único conjunto de regras que, juntas, constituem o modelo entoacional. Entretanto, as curvas de freqüência fundamental geradas pelo modelo, quando se está realizando síntese concatenativa a partir do dicionário de unidades acústicas de um falante, não podem coincidir com aquelas geradas para o outro falante. Isto ocorre porque um dos falantes possui a voz bem mais grave do que a voz do outro. Caso se tentasse fazer com que a mesma curva de freqüência fundamental servisse para ambos, o resultado seria uma péssima qualidade de fala para aquele falante cuja média de valores de F0 ficasse distante da média dos valores

estabelecidos pelo modelo. Isto ocorre devido a limitações da técnica PSOLA ao lidar com grandes variações de freqüência fundamental. Para que os mesmos contornos entoacionais pudessem ser aplicados para dicionários de ambos os falantes, recorreu-se ao artifício de multiplicar, por uma constante, os contornos gerados pelo modelo para que correspondam melhor à média da freqüência fundamental do falante cujo dicionário esteja sendo utilizado no momento. Por multiplicar um contorno entoacional entenda-se multiplicar toda a curva por um fator apropriado que pode ser, por exemplo, 1.3.

O artifício de escalonamento das curvas de freqüência fundamental geradas mostrou que o modelo entoacional apresenta uma relativa independência com relação ao locutor utilizado para a criação do dicionário de unidades acústicas.

No capítulo seguinte é feita uma comparação entre um contorno de freqüência fundamental, obtido a partir da análise de um enunciado natural, e os valores gerados pelo modelo entoacional.

A **Tabela 7.3** ilustra o resultado da aplicação do modelo entoacional sobre a sentença: "[m1] O preço da tarifa telefônica [m2] foi reduzido [m3]", onde [m1], [m2] e [m3] são, respectivamente, marcas prosódicas de início de frase, início de predicado e final de frase. As colunas entituladas "F0 I" e "F0 D" correspondem, respectivamente, aos valores de freqüência fundamental ao início e ao final dos fones constituintes do enunciado.

Fone	F0 I (Hz)	F0 D (Hz)	Fone	F0 I (Hz)	F0 D (Hz)	Fone	F0 I (Hz)	F0 D (Hz)	Fone	F0 I (Hz)	F0 D (Hz)
UW	135	131	AA	117	115	FF	109	111	RR	117	112
PP	131	131	RX	115	116	OO	111	112	EE	112	106
RX	131	131	II	116	121	NN	112	107	DD	106	105
EE	131	132	FF	121	115	IY	107	103	UW	105	104
SS	132	125	AA	115	109	KK	103	106	ZZ	104	105
UW	125	121	TT	109	109	AA	106	124	II	105	105
DD	121	119	EE	109	108	FF	124	121	DD	105	100
AA	119	118	LL	108	109	OO	121	119	UW	100	95
TT	118	117	EE	109	109	IY	119	117			

Tabela 7.3

Capítulo 8
Resultados obtidos

Neste capítulo é apresentada uma avaliação do trabalho realizado.

8.1 Avaliação

Avaliar os resultados do presente trabalho é uma tarefa difícil. Esta dificuldade deve-se ao fato de ser este um trabalho pioneiro no modelamento da prosódia para a conversão texto-fala se for considerado que um trabalho desta natureza é dependente do idioma a que ele se aplica, ou seja, trata-se do primeiro trabalho desta natureza desenvolvido para a Língua Portuguesa falada no Brasil. Assim, torna-se impossível fazer uma avaliação por comparação de resultados com um trabalho anterior.

Em termos absolutos, pode-se dizer que o trabalho realizado, apesar de sua simplicidade e caráter inicial, assegurou uma enorme melhoria na qualidade da fala gerada pelo sistema de conversão texto-fala utilizado para teste e evolução de regras, se comparada à fala sintetizada sem tratamento prosódico.

Recorreu-se ao método de comparação dos parâmetros prosódicos gerados pelos modelos de duração e entonação para uma determinada sentença com os parâmetros obtidos por meio de uma análise desta mesma sentença quando enunciada por uma pessoa.

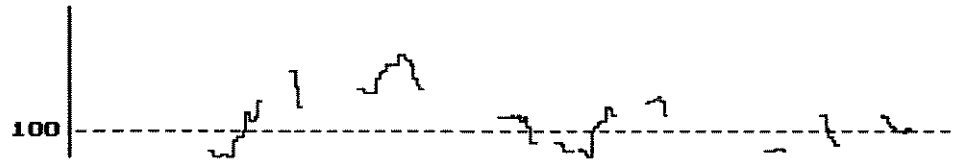
A tabela 8.1 ilustra os valores obtidos para as durações fonéticas pelo modelo de duração desenvolvido (coluna entitulada "Dur Modelada") em comparação com valores obtidos através de análise acústica de um enunciado natural correspondente (coluna entitulada "Dur Natural").

A Fig. 8.1 mostra gráficos de curvas entoacionais correspondentes a uma realização natural (enunciado realizado pelo locutor profissional cuja fala serviu para estudo), a uma síntese com freqüência intrínseca (aquela que possuem as unidades do dicionário na forma como foram gravadas) e à saída do modelo entoacional.

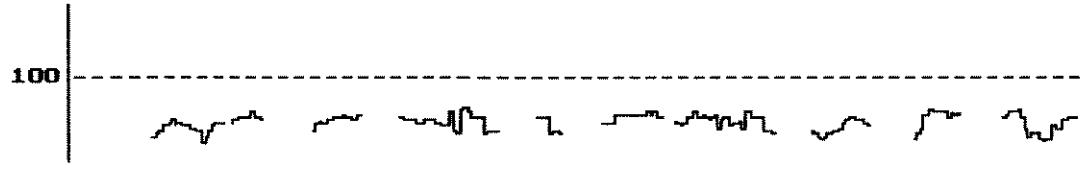
Fone	Dur Modelada	Dur Natural	Fone	Dur Modelada	Dur Natural
EH	238	178	IY	45	28
NN	52	60	UW	44	52
EE	79	54	PP	99	98
SS	120	90	EE	86	70
EE	79	54	RX	40	46
SS	198	146	MM	88	56
AH	170	134	II	174	114
RX	40	36	TT	68	68
IY	45	42	AA	73	52
UW	43	20	UW	52	52
##	100	0	IN	125	98
KK	80	80	TT	70	68
IY	55	0	EE	83	88
UW	52	68	RX	40	44
KK	90	68	KK	66	96
ON	152	112	AN	117	132
VV	116	60	BB	110	104
EE	91	108	II	145	98
NN	44	36	UW	79	90

Tabela 8.1

a)



b)



c)

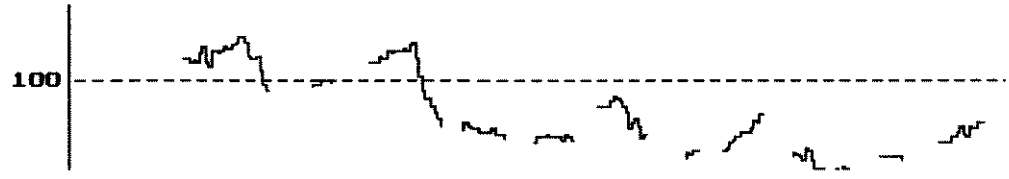


Fig. 8.1 - Gráficos de F0 para enunciados relativos à sentença "_É necessário que o convênio permita o intercâmbio". (a) F0 natural, (b) F0 intrínseca e (c) F0 modelada.

A tabela 8.1 e o gráfico 8.1 mostram que tanto a duração quanto a freqüência fundamental, obtidas através do modelo, acompanham as tendências gerais dos parâmetros prosódicos da realização natural da frase.

Na figura 8.1, os gráficos não apresentam a mesma duração porque os enunciados a, b e c foram produzidos, respectivamente, com duração natural, intrínseca e modelada.

Capítulo 9
Conclusões

Neste capítulo é apresentada uma discussão a respeito do trabalho realizado e são feitas propostas para futuros trabalhos.

9.1 Considerações sobre o trabalho realizado

O presente trabalho resultou de um estudo inicial sobre o complexo problema do modelamento prosódico para a conversão texto-fala. Todo o modelamento apresentado nos capítulos anteriores caracteriza-se principalmente pela simplicidade na abordagem e busca de uma visão global do problema.

O maior empecilho para a realização deste trabalho foi a dificuldade de encontrar estudos lingüísticos sobre a prosódia da Língua Portuguesa falada no Brasil. Não se dispondo, portanto, de um estudo descritivo sobre os fenômenos prosódicos, foi necessário realizar um trabalho de análise sobre realizações de fala natural antes de começar o modelamento prosódico propriamente dito.

Para realizar a análise de realizações de fala natural foi necessário desenvolver um programa de computador voltado para esta finalidade. Na época, a decisão de construir tal programa deveu-se ao fato da não disponibilidade, para uso do autor, de uma ferramenta já pronta. Apesar desta decisão ter sido uma imposição das circunstâncias, ela se revelou bastante acertada devido a três motivos. Primeiro, foi possível construir uma ferramenta totalmente voltada para as necessidades, o que levou a uma economia de tempo quando da utilização desta ferramenta. Segundo, foi auferida experiência na determinação dos requisitos a que uma ferramenta desta natureza deve atender. Terceiro, a construção do programa levou ao exercício da programação orientada ao objeto.

Sendo este um trabalho inicial, seria de se esperar que fosse feita uma opção entre atacar apenas o problema do modelamento da freqüência fundamental ou restringir-se ao modelamento

da duração. Optou-se, entretanto, por abordar ao mesmo tempo os dois tópicos devido ao objetivo de auferir uma melhora significativa na qualidade da fala sintetizada pelo sistema de síntese disponível para teste. Se fosse realizado apenas um dos modelamentos, por mais eficiente que se mostrasse, não seria alcançado o objetivo de melhoria da qualidade da fala produzida.

Todos os esforços foram direcionados no sentido da obtenção rápida de resultados. Assim, não foi possível, infelizmente, abordar com grande profundidade nehum tópico, seja relacionado à duração ou ao pitch.

Ambos os modelos, de pitch e de duração, foram construídos com o propósito bem definido de operarem requerendo uma quantidade mínima de informações extraídas a partir do texto submetido à síntese. De fato, as únicas informações necessárias para o funcionamento dos modelos são as marcas de fronteiras prosódicas. A decisão de construí-los assim deveu-se à não viabilidade de realizar um processamento mais elaborado do texto, dada a limitação de tempo para desenvolvimento. Por operarem com uma quantidade mínima de informações, os modelos apresentam evidentemente um desempenho aquém do que seria alcançado se fossem usadas mais informações.

Devido ao fato do corpus de frases ser bastante restrito, os dados extraídos a partir de sua análise foram em quantidade bastante limitada. Essa escassez de dados levou à adoção de uma metodologia de desenvolvimento em que as regras não são derivadas em sua forma final a partir apenas da análise dos dados, mas adquirem forma final após sucessivos testes e correções. Isto significou que a construção do modelo prosódico somente foi possível graças à disponibilidade de um conversor texto-fala.

O desenvolvimento do presente trabalho não foi orientado por preocupações descritivas, isto é, não foi feita uma descrição formal de fenômenos prosódicos do Português falado no Brasil, mesmo porquê o estudo de enunciados de apenas dois locutores não permitiria tirar conclusões a respeito da Língua.

Grande parte da complexidade encontrada no desenvolvimento deste trabalho deveu-se a seu caráter interdisciplinar. Foi necessário aliar conhecimentos de computação, processamento digital de sinais e fonética.

A pouca disponibilidade de tempo e o propósito de alcançar uma melhoria na qualidade da fala sintetizada, impossibilitaram a realização de testes de várias outras abordagens possíveis para o modelamento prosódico. Poderia-se, por exemplo, testar o uso de sistemas de aprendizado.

9.2 Propostas para futuros trabalhos

Talvez a maior limitação do modelo apresentado aqui seja o fato de somente operar sobre sentenças declarativas neutras. É necessário, portanto, ampliar o modelo para tratar sentenças interrogativas, exclamativas e imperativas. Outra grande limitação é a marcação manual de fronteiras prosódicas. É mister, portanto, automatizar o processo.

Além disso, devem ser realizados estudos voltados para as abordagens baseadas em bancos de dados e sistemas de aprendizado.

Apêndice

Corpus de 164 frases declarativas usadas para análise prosódica

A cotação do dólar aumentou, e as bolsas fecharam em baixa.

A cotação do dólar aumentou, mas as bolsas fecharam em baixa.

A bolsa ficará estável ou sofrerá uma pequena queda.

Não haverá ajustes, nem modificações radicais no plano.

Foi detectado um problema em seu cartão; ele deve ser substituído.

É necessário que o convênio permita o intercâmbio.

Posso afirmar-lhes que o convênio permite o intercâmbio.

O convênio que foi assinado recentemente permite o intercâmbio.

O convênio permite o intercâmbio porque visa a integração entre alunos de culturas diferentes.

O convênio permite intercâmbio quando se trata de universidades vinculadas ao projeto de integração.

O convênio, que foi assinado recentemente, permite o intercâmbio.

O convênio assinado na última reunião é mais interessante do que o anterior.

À medida que o tempo passa, mais nos convencemos da eficácia do convênio.

Se o convênio permite intercâmbio, devemos aproveitar a oportunidade para desenvolver o projeto de integração.

Localizado a uma quadra do centro da cidade, o condomínio permite que você una trabalho e conforto.

É suficiente.

Isto é suficiente.

O saldo é suficiente.

O saldo de sua conta é suficiente.

O saldo disponível é insuficiente.

O saldo disponível em sua conta é insuficiente.

Isto parece insuficiente.

O saldo parece ser insuficiente.

O saldo sempre está disponível.

O saldo está sempre disponível no início do mês.

No início do mês, o saldo está disponível.

Esta é a última chamada para o vôo 737 da Rio-Sul.

Isto é uma pesquisa de opinião pública.

O valor de sua conta telefônica é baixo.

É de trinta mil cruzeiros, o valor de sua conta telefônica.

O vencimento de sua prestação será no dia quatro de junho.

O preço aumentou.

O preço do café aumentou.

O preço do café expresso aumentou.

O preço do café aumentou consideravelmente.

O preço do café aumentou consideravelmente na semana passada.

Aumentou o preço do café.

As taxas de juros no mercado interno estão subindo bastante.

As contas chegaram atrasadas.

As contas chegaram muito atrasadas ontem.

As contas telefônicas deste mês chegaram muito atrasadas ao banco.

Ontem, as contas chegaram aqui muito atrasadas.

Chegaram atrasadas.

Chegaram atrasadas todas as contas telefônicas deste mês.

O governo aumentou o imposto no mês passado.

O governo aumentou o imposto sobre importação.

O governo entregou os formulários aos contribuintes.

O governo entregou aos contribuintes os formulários.

O banco colocará a sua disposição o novo cheque.

A conta telefônica em nome de Adelaide Barroso terá vencimento amanhã.

A perda da atratividade das aplicações em caderneta de poupança está provocando um aumento de consumo no país.

O mercado foi considerado inadequado.

O mercado foi considerado inadequado pelos analistas.

O mercado foi considerado inadequado naquele momento.

Naquele momento, o mercado foi considerado inadequado pelos analistas das 55 melhores instituições de pesquisa.

Estação Santa Cruz.

Passageiros com destino a São Paulo, Recife e Fortaleza, embarque imediato, portão sete. Última chamada.

Os bancos atrás de mais eficiência.

Descontos de até 50%.

Número incompleto.

A Telebrás, a empresa de telecomunicações brasileira, está investindo em pesquisa.

A Telebrás, uma empresa estatal, está investindo em pesquisa.

Tivemos recentemente a seguinte notícia: a Telebrás passará a investir mais em pesquisa.

Telesp informa: dezenove horas e trinta minutos.

Empresário, é preciso antecipar o futuro.

Prezado cliente, aguardaremos o seu comparecimento.

O código foi registrado pelo funcionário.

O convênio, um documento de trinta páginas, tem permitido o intercâmbio.

Os caixas eletrônicos não aceitarão depósitos.

Os caixas eletrônicos não aceitarão mais depósitos a partir das 15 horas.

Os caixas eletrônicos não aceitarão mais depósitos a partir das 15 horas do próximo dia vinte e nove.

Os caixas eletrônicos não aceitarão mais depósitos a partir das 15 horas e não estarão disponíveis para saques a partir das 17 horas do dia trinta.

Todos os bancos devem fazer a atualização cadastral até 31 de dezembro

Todos os bancos devem fazer a atualização cadastral de seus clientes até 31 de dezembro.

Todos os bancos devem fazer a atualização cadastral de seus clientes até 31 de dezembro de acordo com a determinação do Banco Central.

Todos os bancos devem fazer a atualização cadastral de seus clientes até 31 de dezembro de acordo com a determinação do Banco Central, em cumprimento à Resolução 2025 do Conselho Monetário Nacional.

Todos os bancos instalados no Brasil devem fazer a atualização cadastral de seus clientes até 31 de dezembro de acordo com a determinação explícita do Banco Central, em cumprimento à Resolução 2025 do Conselho Monetário Nacional.

As operações continuam.

As operações de crédito continuam.

As operações de crédito e financiamento continuam a seguir as regras.

As operações de crédito e financiamento continuam a seguir as regras do Banco Central.

As operações de crédito e financiamento corrigidas por outros indexadores continuam normalmente a seguir as regras do Banco Central ainda em vigor, beneficiando assim a maioria.

O aumento no consumo devido ao Plano Real impulsionou os preços.

Segundo dados do IBGE, o aumento no consumo devido ao Plano Real impulsionou os preços.

Segundo dados do IBGE, o aumento no consumo devido ao Plano Real tem impulsionado consideravelmente os preços nas últimas semanas.

Veio, parou, foi-se.

Ela veio, parou pouco, foi-se depressa.

Alice estudou e fez ótimos exames.

Alice estudou e fez ótimos exames de espanhol.

Ela não veio, nem me telefonou.

Ela não veio, mas me telefonou.

Ela não veio, portanto não haverá tempo.

Soube que aconteceu.

Souberam que algo aconteceu.

Elas bem cedo souberam que algo de errado aconteceu.

A verdade é que existe algo errado.

É que existe algo errado.

O professor deseja que os alunos sejam aprovados.

Deseja que os alunos sejam aprovados.

Ela teve a impressão de que a casa estava cheia.

A menina, que vi, partiu ontem.

A menina, que estava aqui, partiu ontem.

A menina, que estava exatamente aqui, partiu ontem.

Quando não havia mais tempo, veio.

Veio quando não havia mais tempo.

Elas vieram quando não havia mais tempo.

Quando não havia mais tempo, elas vieram.

Continuará a vir se lhe for consentido.

Se lhe for consentido, continuará a vir.

Se puder, continuará a vir.

Continuará a vir, se puder.

Ela é bonita.

A mina é bonita.

A menina é bonita.

A menina alegre é bonita.

A menina inteligente é bonita.

Ela está bonita.

Ela parece bonita.

A menina sempre está bonita.

A menina alegre sempre está bonita.

A menina parece estar feia.

A menina parece estar sempre feia.

Parece bonita.

A menina não é bonita.

Ele cantou.

O sapo cantou.

O sapato cantou.

O sapato feio cantou.

O sapo cantou depressa.

O sapo nunca cantou.

Cantou depressa.

Cantou.

José chegou cansado.

Aquele José chegou cansado.

José chegou aqui cansado.

José chegou aqui muito cansado.

José chegou muito cansado.

Chegou cansado.

O menino quebrou o vidro.

O menino desastrado quebrou o vidro.

O menino quebrou o vidro azul da casa grande.

José foi considerado o responsável.

José sempre foi considerado o responsável.

José sempre foi considerado o principal responsável.

José sempre foi considerado o principal responsável pela desordem administrativa em que se encontra o país.

Diário.

Diariamente.

Curva perigosa.

Dia vinte do sete.

Sim.

Não.

Saldo.

Saldo: vinte e cinco reais.

Poupança.

Poupança: vinte e cinco reais.

Vinte e cinco.

Vinte e cinco reais.

Cento e vinte e cinco.

Cento e vinte e cinco reais.

Quatrocentos e quarenta e nove.

Dois mil, cento e vinte e cinco.

Dezesseis mil e quinhentos.

Oitenta milhões, trezentos e sessenta mil e duzentos e setenta e um.

Referências bibliográficas:

[1] Allen, J., Hunnicutt, S., & Klatt, D. H.; "From text-to-speech: The MITalk System"; Cambridge, UK, 1987.

[2] O'Shaughnessy, D.; "Speech Communication"; Addison-Wesley, New York, 1987.

[3] Egashira, F.; "Síntese de voz a partir de texto para a Língua Portuguesa"; Tese de Mestrado, Faculdade de Engenharia Elétrica da UNICAMP, julho de 1992.

[4] Charpentier, F. & Moulines, E.; "Nouvelles techniques de synthèse de la parole"; L'écho des RECHERCHES, no. 137, pp. 37-46, 1989

[5] Aubergé, V.; "Semi-automatic of a prosodic contour lexicon for text-to-speech system"; Elsevier Science Publishers, no. 39, pp. 274-287, 1992.

[6] Kurt Shäfer - Vincent; "Pitch period detection and chaining: method and evaluation"; Phonética, no. 40, pp. 177-202, 1983.

[7] Egashira, F. & Violaro, F.; "Conversor texto-fala"; I Forum Nacional de Ciência em Saúde, XIII Congresso Brasileiro de Engenharia Biomédica - 20 a 24 de novembro de 1992, Caxambu, MG.

[8] Journal of the International Phonetic Association; vol. 23 (1), June 1993.

[9] Quené, H. & René, K.; "The derivation of prosody for text-to-speech from prosodic sentence structure"; Computer Speech and Language, no. 6, pp. 77-99, 1992

[10] Simões, A. R. M.; "Predicting sound segment duration in connected speech: an acoustical study of Brazilian Portuguese", In the First International Conference on Speech Synthesis, ETRW-GALF90, 1990, september 23-28. Autrans, France: Institut de la communication parlée.