**UNIVERSIDADE ESTADUAL DE CAMPINAS**

**FACULDADE DE ENGENHARIA ELÉTRICA E DE COMPUTAÇÃO**

# Implementação de um Sistema de Conversão Texto-Fala para o Português do Brasil

**Flávio Olmos Simões**

**Orientador: Prof. Dr. Fábio Violaro**

**Banca Examinadora:**

**Prof. Dr. Fábio Violaro (DECOM/FEEC/UNICAMP)**

**Prof. Dr. Jaime Portugheis (DECOM/FEEC/UNICAMP)**

**Prof. Dr. Plínio A. Barbosa (DL/IEL/UNICAMP)**

**Dissertação para a obtenção do título de Mestre em Engenharia Elétrica.**

**Campinas, maio de 1999.**

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

Si51i

Simões, Flávio Olmos
Implementação de um sistema de conversão texto-fala para o português do Brasil. / Flávio Olmos Simões.--Campinas, SP: [s.n.], 1999.

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

1. Síntese da voz. 2. Sistemas de processamento da fala. 3. Interação homem-máquina. I. Violaro, Fábio. II. Universidade Estadual de Campinas. Faculdade de Engenharia Elétrica e de Computação. III. Título.

## Resumo

A síntese de fala a partir de texto é o principal objeto de estudo deste trabalho. As dificuldades principais do processo de conversão texto-fala são colocadas em questão e uma estratégia de implementação de um sistema de conversão texto-fala para o português do Brasil é apresentada. Esse sistema, baseado no método de síntese concatenativa, utiliza um inventário de 2.450 segmentos de fala pré-gravados e é capaz de empregar duas técnicas de síntese distintas: TD-PSOLA e síntese híbrida.

A adoção de critérios lingüísticos cuidadosos, principalmente na etapa de transcrição fonética e na elaboração do inventário de unidades constitui o ponto chave deste trabalho. A notação fonética utilizada diferencia dois tipos de segmentos fonéticos (plenos e reduzidos), que se distinguem no grau pelo qual estão sujeitos a fenômenos de coarticulação. O inventário de unidades foi construído de forma a preservar segmentos reduzidos e encontros vocálicos. No intuito de reduzir o tamanho do inventário, alguns cortes no interior de segmentos reduzidos foram efetuados. Mais uma vez, nesse caso, utilizaram-se critérios lingüísticos cuidadosos, a fim de minimizar descontinuidades espectrais após a concatenação.



## Abstract

Text-to-speech synthesis is the main subject treated in this work. Most of the difficulties related to this task are discussed, and an implementation of a Brazilian Portuguese text-to-speech concatenative synthesis system is presented. The system uses an inventory of 2,450 pre-recorded speech segments, and is able to employ two distinct synthesis techniques: TD-PSOLA and hybrid synthesis.

The use of carefully chosen linguistic criteria, mainly during phonetic transcription and also during the creation of the speech segments inventory, is the main contribution of this work. The phonetic notation employed here distinguishes two kinds of phonetic segments (full and reduced), on the basis of the extension of coarticulation phenomena. The main criterion underlying the building of the speech segments inventory was to preserve reduced segments and vowel clusters. Nevertheless, some of the reduced segments were split, aiming at reducing the size of the inventory. Once again, in this case, specific linguistic criteria were employed, in order to minimize spectral discontinuities after concatenation.

Esse trabalho foi financiado pela Fundação de Amparo à Pesquisa do Estado de São Paulo, através da bolsa de Mestrado referente ao processo nº 97/00686-5.

*Não percas nunca pelo vão saber,*
*A fonte viva da sabedoria.*
*Por mais que estudes, que te adiantaria,*
*Se a teu amigo tu não sabes ler?*

(Mário Quintana)

# Agradecimentos

Em primeiro lugar, agradeço ao Prof. Dr. Fábio Violaro pela oportunidade de realizar este trabalho e pela orientação dada ao longo desses dois anos.

Agradeço à Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), pelo apoio financeiro e pela confiança no cumprimento do cronograma estabelecido.

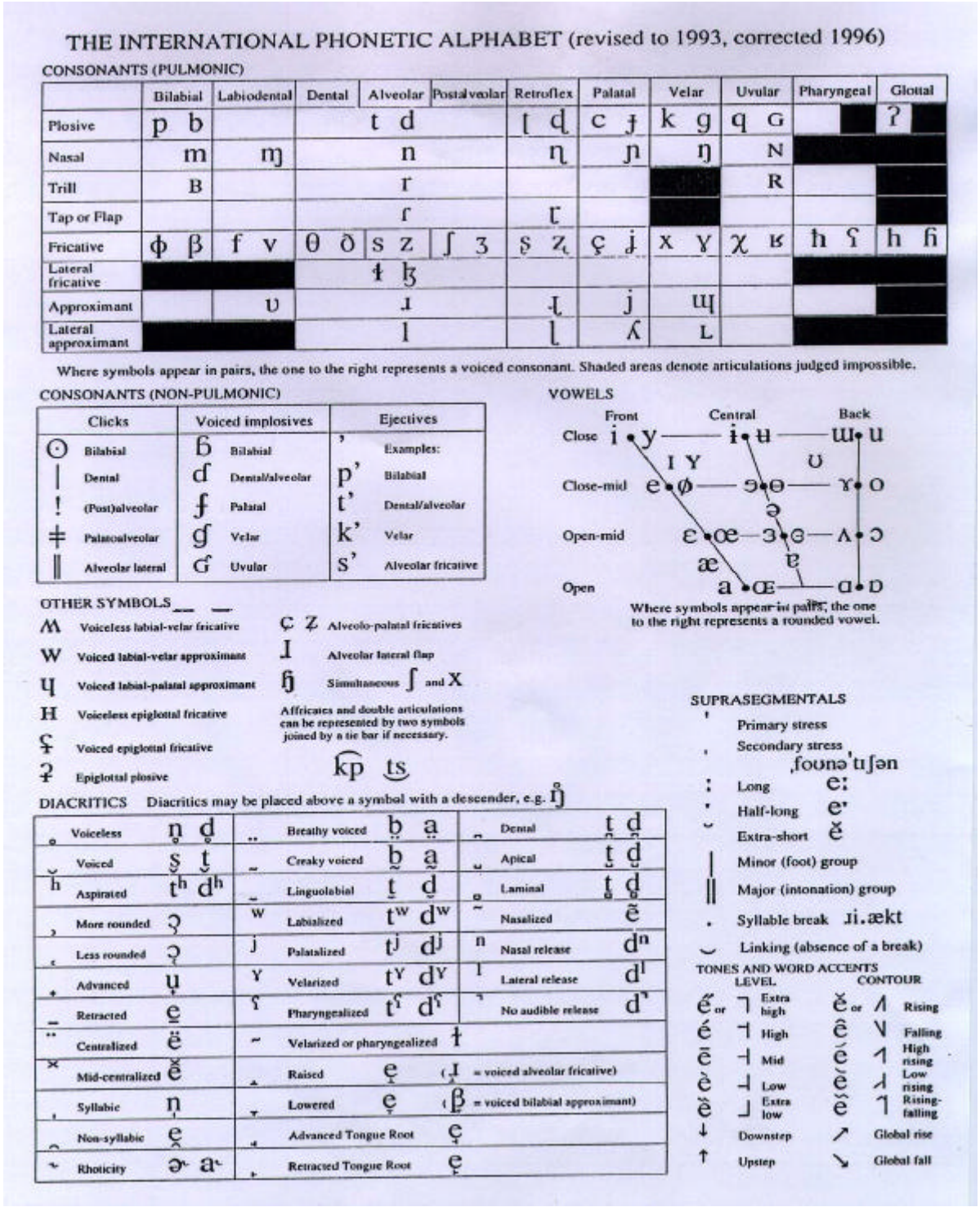
Agradeço imensamente ao Prof. Dr. Plínio Barbosa pelas informações valiosas que me foram passadas durante toda a realização deste Mestrado. Sou especialmente grato pela paciência e espírito de cooperação que sempre demonstrou, aos quais credito grande parte do êxito alcançado neste trabalho.

Agradeço à equipe de trabalho do Laboratório de Fonética Acústica e Psicolingüística Experimental (LAFAPE), em especial à Prof ª. Drª. Eleonora Albano, pela formação básica na área de Lingüística, e a Patrícia Aquino, cujo esforço quase sobre-humano na tarefa de segmentação das unidades não poderia jamais deixar de ser aqui mencionado.

Agradeço aos amigos Luís, Fabrício, Leonardo, Wilson, Raquel, Antônio Marcos e Edmilson, de cuja amizade me orgulho, e a quem desejo a mesma boa sorte que sempre me trouxeram.

Agradeço à minha família, em especial ao meu irmão André, por ser meu irmão e meu amigo, e aos meus pais, que sempre acreditaram em mim, e a quem devo tudo o que tenho hoje, sem exageros. Ao afeto que sempre demonstraram não agradeço com palavras. Que o meu futuro seja um espelho dos seus ensinamentos e do seu amor.

# IPA – Alfabeto Fonético Internacional



THE INTERNATIONAL PHONETIC ALPHABET (revised to 1993, corrected 1996)

CONSONANTS (PULMONIC)

| | Bilabial | Labiodental | Dental | Alveolar | Postalveolar | Retroflex | Palatal | Velar | Uvular | Pharyngeal | Glottal |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Plosive | p b | | | t d | | ʈ ɖ | c ɟ | k g | q ɢ | | ʔ |
| Nasal | m | ɱ | | n | | ɳ | ɲ | ŋ | ɴ | | |
| Trill | ʙ | | | r | | | | | ʀ | | |
| Tap or Flap | | | | ɾ | | ɽ | | | | | |
| Fricative | ɸ β | f v | θ ð | s z | ʃ ʒ | ʂ ʐ | ç ʝ | x ɣ | χ ʁ | ħ ʕ | h ɦ |
| Lateral fricative | | | | ɬ ɮ | | | | | | | |
| Approximant | | ʋ | | ɹ | | ɻ | j | ɰ | | | |
| Lateral approximant | | | | l | | ɭ | ʎ | ʟ | | | |

Where symbols appear in pairs, the one to the right represents a voiced consonant. Shaded areas denote articulations judged impossible.

CONSONANTS (NON-PULMONIC)

| Clicks | | Voiced implosives | | Ejectives | |
|---|---|---|---|---|---|
| ʘ | Bilabial | ɓ | Bilabial | ’ | Examples: |
| \| | Dental | ɗ | Dental/alveolar | pʼ | Bilabial |
| ! | (Post)alveolar | ʄ | Palatal | tʼ | Dental/alveolar |
| ǂ | Palatoalveolar | ɠ | Velar | kʼ | Velar |
| ‖ | Alveolar lateral | ʛ | Uvular | sʼ | Alveolar fricative |

VOWELS




Where symbols appear in pairs, the one to the right represents a rounded vowel.

OTHER SYMBOLS

| | | | |
|---|---|---|---|
| ʍ | Voiceless labial-velar fricative | ɕ ʑ | Alveolo-palatal fricatives |
| w | Voiced labial-velar approximant | ɺ | Alveolar lateral flap |
| ɥ | Voiced labial-palatal approximant | ʧ | Simultaneous ʃ and x |
| ʜ | Voiceless epiglottal fricative | | |
| ʢ | Voiced epiglottal fricative | | |
| ʡ | Epiglottal plosive | | |

Affricates and double articulations can be represented by two symbols joined by a tie bar if necessary.

k͡p t͡s

SUPRASEGMENTALS

| | | |
|---|---|---|
| ˈ | Primary stress | |
| ˌ | Secondary stress | ˌfoʊnəˈtɪʃən |
| ː | Long | eː |
| ˑ | Half-long | eˑ |
| ˘ | Extra-short | ĕ |
| \| | Minor (foot) group | |
| ‖ | Major (intonation) group | |
| . | Syllable break | ɹi.ækt |
| ‿ | Linking (absence of a break) | |

DIACRITICS Diacritics may be placed above a symbol with a descender, e.g. ŋ̊

| | | | | | |
|---|---|---|---|---|---|
| ◌̥ | Voiceless | n̥ d̥ | ◌̤ | Breathy voiced | b̤ a̤ |
| ◌̬ | Voiced | s̬ t̬ | ◌̰ | Creaky voiced | b̰ a̰ |
| ◌ʰ | Aspirated | tʰ dʰ | ◌̼ | Linguolabial | t̼ d̼ |
| ◌̹ | More rounded | ɔ̹ | ◌ʷ | Labialized | tʷ dʷ |
| ◌̜ | Less rounded | ɔ̜ | ◌ʲ | Palatalized | tʲ dʲ |
| ◌̟ | Advanced | u̟ | ◌ˠ | Velarized | tˠ dˠ |
| ◌̠ | Retracted | e̠ | ◌ˤ | Pharyngealized | tˤ dˤ |
| ◌̈ | Centralized | ë | ◌̴ | Velarized or pharyngealized | ɫ |
| ◌̽ | Mid-centralized | e̽ | ◌̝ | Raised | e̝ ( ɹ̝ = voiced alveolar fricative) |
| ◌̩ | Syllabic | n̩ | ◌̞ | Lowered | e̞ ( β̞ = voiced bilabial approximant) |
| ◌̯ | Non-syllabic | e̯ | ◌̘ | Advanced Tongue Root | ɕ |
| ◌˞ | Rhoticity | ɔ˞ a˞ | ◌̙ | Retracted Tongue Root | e̙ |

TONES AND WORD ACCENTS

| LEVEL | | CONTOUR | |
|---|---|---|---|
| é͡ or ˥ | Extra high | ě or ˩˥ | Rising |
| é ˦ | High | ê ˥˩ | Falling |
| ē ˧ | Mid | e᷄ ˧˥ | High rising |
| è ˨ | Low | e᷅ ˩˧ | Low rising |
| ȅ ˩ | Extra low | e᷈ ˥˧˩ | Rising-falling |
| ↓ | Downstep | ↗ | Global rise |
| ↑ | Upstep | ↘ | Global fall |

# Índice

# Índice de figuras

# 1 Introdução

## 1.1 Considerações Iniciais

Os primeiros computadores, surgidos por volta da metade deste século, representaram um grande avanço em relação à tecnologia existente até então. Apesar de facilitar muitas tarefas, o computador era, todavia, um equipamento de uso bastante restrito. O número de pessoas que sabiam operá-lo era pequeno e altamente especializado; não existia ainda uma preocupação com a interface de comunicação entre a máquina e o operador. Com o passar do tempo, o computador foi se tornando um equipamento de uso geral e cotidiano; atualmente existem milhões de computadores sendo utilizados para os mais diversos propósitos e por todo tipo de pessoas.

Essa massificação do uso do computador fez com que o modo de comunicação do usuário com a máquina fosse se tornando, com o passar do tempo, cada vez mais importante. Se, no princípio, um operador especializado se comunicava com o computador por meio de cartões perfurados e códigos binários de difícil interpretação, atualmente há todo um aparato tecnológico visando a tornar mais natural a interação entre o homem e a máquina, fazendo com que teclados, *mouses*, monitores coloridos e equipamentos de multimídia façam parte da vida cotidiana de muita gente.

Um dos veículos mais empregados para a troca de informação entre o homem e o computador é o texto escrito. Sua utilização se mostra bastante adequada em diversas situações, pois a geração e a interpretação de informação escrita é bastante simples, econômica e confiável. Teclados, monitores, impressoras, todos eles são dispositivos de interface que se utilizam da linguagem textual.

Em certas circunstâncias, no entanto, a troca de informação através de texto pode se mostrar bastante inconveniente. De certa maneira, a utilização da comunicação textual "amarra" o operador, pois este deve estar sempre próximo ao terminal. Além disso, a

utilização dos olhos ou das mãos em outras tarefas se torna mais difícil, uma vez que eles estarão comprometidos nas tarefas de leitura e de digitação, respectivamente. Em aplicações que requeiram uma certa agilidade, a troca de informação por meio de texto certamente não é a mais indicada.

A utilização da voz como forma de comunicação com o computador ainda não é tão difundida como a utilização do texto. Isso se deve essencialmente à dificuldade que os sistema computacionais atuais encontram em lidar com a linguagem falada, muito mais complexa que a linguagem escrita. Essa complexidade, no entanto, é conseqüência de uma riqueza maior de informações inerente à própria linguagem falada. Através da fala pode-se transmitir emoção (ironia, firmeza, raiva, incerteza, etc.), bem como informações sobre a pessoa que fala (cada indivíduo possui um timbre de voz característico e uma maneira própria de elocução); podemos inclusive deduzir o sexo, a faixa etária e a procedência de uma pessoa a partir de sua voz [28].

A interação com o computador através da fala dá mais liberdade ao operador. A geração de mensagens faladas, por exemplo, permite que os olhos estejam livres para a realização de outras tarefas. O reconhecimento de comandos de voz, por sua vez, elimina a necessidade de digitação, tornando a execução de tarefas mais ágil, sem exigir, obrigatoriamente, a presença do operador junto ao terminal.

A utilização da fala nos sistemas computacionais segue uma tendência natural que visa a tornar a interação homem-máquina mais direta e efetiva. A fala está presente em todas as culturas, por isso sua utilização permite a comunicação com o usuário de forma mais natural e eficiente.

## 1.2 Objetivos do trabalho

Este trabalho tem por objetivo principal apresentar uma estratégia de implementação de um sistema de conversão texto-fala para o português do Brasil. Todas as etapas do processo de conversão serão cobertas, desde a normalização do texto de entrada até a geração do sinal acústico correspondente à fala sintetizada.

Uma das diretrizes deste trabalho foi a de preocupar-se não apenas com a inteligibilidade do sinal de fala gerado, visto tratar-se esse de um requisito básico a qualquer sistema de conversão texto-fala. Procurou-se, adicionalmente, planejar cuidadosamente cada uma das etapas do processo, tendo-se como meta final a geração de um sinal de fala de alta qualidade, tão próximo quanto possível de um sinal de fala natural.

O método de síntese empregado foi o de síntese concatenativa, na qual segmentos de fala pré-gravados são concatenados e submetidos a um processamento de sinal cujo objetivo é o de fornecer contornos prosódicos apropriados ao sinal de fala gerado. A opção por essa estratégia deveu-se à maior simplicidade de implementação, mas também ao potencial que ela apresentava de gerar fala com qualidade, conforme demonstrava a experiência prévia de trabalhos realizados tanto pelos pesquisadores da UNICAMP como pelo resto da comunidade científica.

Quanto às técnicas de síntese adotadas, optou-se pela utilização de duas: TD-PSOLA e síntese híbrida. A escolha do TD-PSOLA levou em conta o fato de ser esta a técnica de síntese mais comumente utilizada, devido tanto à sua simplicidade quanto aos resultados altamente satisfatórios que ela permite produzir. Já a síntese híbrida foi escolhida por também apresentar potencial para produzir um sinal sintetizado de alta qualidade, além de resolver algumas das limitações inerentes à técnica TD-PSOLA.

Um aspecto importante a ser aqui ressaltado é a incorporação de modelos lingüísticos realistas nas diversas etapas desse trabalho, principalmente com relação ao módulo de transcrição ortográfico-fonética e aos critérios de elaboração do inventário de unidades para

concatenação. Ao reconhecermos que o processo de produção da fala não consiste apenas na geração de um sinal acústico, mas envolve toda uma interação com a língua e com o conteúdo da mensagem a ser transmitida, devemos admitir que o conhecimento lingüístico deverá, necessariamente, estar presente e atuando no processo da geração de fala artificial a partir de texto escrito.

Vale lembrar, por fim, que o objetivo deste trabalho não esteve centrado unicamente em levar a cabo a implementação de um sistema de conversão operacionalmente viável. Uma segunda tarefa igualmente importante, e que consumiu um intervalo de tempo considerável na elaboração desta tese, foi a realização de um levantamento bibliográfico bastante extenso e detalhado a respeito dos diversos assuntos relacionados à síntese de fala e particularmente à conversão texto-fala. Essa tarefa, que acabou por refletir-se na própria estrutura da tese, baseou-se numa crença particular do autor dessa dissertação de que o texto final deveria funcionar não apenas como um relato das atividades desenvolvidas ao longo do Mestrado, mas também como uma referência geral aos assuntos tratados, visto haver pouco material publicado em português sobre o tema. Muito mais do que apresentar soluções definitivas, este trabalho procurou abordar, sob todos os ângulos, cada um dos aspectos do problema da síntese de fala a partir de texto, tentando apontar alguns caminhos a serem seguidos e estimulando a busca por novas soluções.

## 1.3 Estrutura da tese

O presente trabalho está estruturado em nove capítulos, divididos da maneira descrita a seguir.

O Capítulo 1 corresponde à *apresentação geral* do trabalho.

O Capítulo 2 introduz os aspectos básicos dos principais ramos que constituem as *ciências da fala* (codificação de fala, síntese de fala, reconhecimento de locutor,

reconhecimento de fala, reconhecimento de língua, tradução automática.). A seguir, analisa com mais detalhes o ramo da *síntese de fala*, apresentando os diversos problemas a ela relacionados, fazendo considerações a respeito de suas aplicações práticas e apresentando um histórico da evolução dessa tecnologia ao longo do tempo.

O capítulo 3 apresenta a *Teoria Acústica de Produção da Fala*, que constitui o modelo matemático normalmente adotado para explicar o processo de produção da fala. A seguir, tece algumas *considerações de natureza lingüística*, introduzindo algumas noções básicas sobre Fonética e Fonologia.

O capítulo 4 introduz a *conversão texto-fala* como uma das áreas de estudo relacionadas à síntese de fala. Descreve os diversos problemas envolvidos na síntese de fala a partir de texto e apresenta, de forma resumida, as várias etapas percorridas ao longo do processo de transformação de um texto genérico em fala sintetizada.

O Capítulo 5 apresenta a primeira etapa a ser executada por um sistema de conversão texto-fala, que corresponde à fase de *processamento lingüístico*, e entra em detalhes a respeito dos passos que constituem esse processamento (pré-processamento do texto e transcrição ortográfico-fonética).

O Capítulo 6 trata do *módulo de processamento prosódico*, que num sistema de conversão texto-fala é responsável por conferir padrões de entonação e de ritmo apropriados às sentenças a serem sintetizadas. Apresenta alguns conceitos básicos sobre prosódia, descreve os principais parâmetros prosódicos, e discute alguns modelos para a determinação automática desses parâmetros.

O Capítulo 7 discute os mecanismos de geração de fala artificial. Apresenta os três métodos de *síntese* mais utilizados (síntese por regras, síntese articulatória e síntese concatenativa). Dentro da síntese concatenativa, entra em detalhes a respeito de duas técnicas de síntese específicas: a síntese PSOLA e a síntese híbrida.

O Capítulo 8 apresenta um *sistema de conversão texto-fala concatenativo para o português do Brasil*, construído no Laboratório de Processamento Digital de Fala da Faculdade de Engenharia Elétrica e de Computação da UNICAMP, com a colaboração do Laboratório de Fonética Acústica e Psicolingüística Experimental (LAFAPE) do Instituto de Estudos da Linguagem, também da UNICAMP. O capítulo faz a descrição dos detalhes de implementação do sistema e da operação de cada um de seus módulos constituintes, ressaltando a importância dos modelos lingüísticos incorporados ao sistema.

Por fim, o Capítulo 9 constitui a *conclusão* da tese, procurando mostrar as contribuições oferecidas pelo trabalho e apresentando sugestões para trabalhos futuros que visem a dar prosseguimento às linhas de pesquisa seguidas até o momento.

# 2 Ciências da fala e síntese de fala

## 2.1 Ciências da fala

Por ser a mais simples, natural e universal forma de comunicação do ser humano, a fala sempre despertou um grande interesse científico. No entanto, o grande desenvolvimento no estudo da fala deu-se apenas no século XX, com o surgimento de uma teoria bastante consistente e de técnicas poderosas de processamento de fala. Esse *nascimento tardio* tem uma razão de ser: foi apenas com o surgimento dos computadores e da tecnologia digital que se tornou possível armazenar, manipular e extrair informações do sinal de fala de maneira eficiente. Esse apoio tecnológico foi essencial para a interação entre as chamadas *ciências da fala*, bem como o desenvolvimento de algumas delas. Além disso, o grande desenvolvimento dos meios de comunicação, que desempenham um papel cada vez mais fundamental dentro da nossa sociedade, fizeram com que o processamento de fala se tornasse uma exigência natural desse novo mundo *interligado* [55].

Por serem tão dependentes dos avanços tecnológicos, as *ciências da fala* estão em fase de pleno crescimento, o que significa que ainda existe muita coisa "por fazer". À medida que aumentam a velocidade de processamento dos computadores e a capacidade de armazenamento dos dispositivos de memória, novas técnicas de manipulação do sinal de fala tornam-se possíveis

Uma característica marcante das *ciências da fala* é a sua interdisciplinaridade. Muito mais do que um novo ramo da ciência, podemos dizer que as *ciências da fala* representam a fusão de ramos aparentemente distintos, e que no entanto caminham juntos na construção de novas teorias e na busca de soluções. Desta forma a Engenharia, a Física, a Lingüística, a Psicologia Experimental e Cognitiva, a Fisiologia da Fala e a Informática dão, cada uma, a sua contribuição para alcançar um objetivo que, sozinhas, não poderiam alcançar: conhecer o

objeto multifacetado da fala em sua função comunicativa, esta constituída por cinco pólos (língua, produção, percepção, cognição e acústica).

No que se refere ao aspecto da Engenharia, descreveremos brevemente aquelas que constituem as principais áreas de estudo do processamento de fala nos dias de hoje:

### 2.1.1 Codificação de fala

A *codificação de fala* estuda técnicas de compressão da informação contida no sinal de fala com o intuito de:

1. armazenar o sinal de fala de maneira econômica;
2. transmitir o sinal de fala através de canais cuja largura de banda é mais estreita do que a que seria necessária para transmitir o sinal original;
3. extrair parâmetros do sinal de fala que permitam a sua manipulação, com objetivos diversos.

Os codificadores de fala podem ser caracterizados de acordo com certos atributos: a *taxa de bits*, por exemplo, indica a capacidade de compressão do codificador; o *atraso* está relacionado com o tempo de codificação/decodificação; a *degradação* do sinal indica o quanto o sinal original foi alterado pelo processo de codificação/decodificação; por fim, a *complexidade* relaciona-se com o custo de implementação em hardware do codificador. Todos esses atributos são importantes mas, de acordo com a aplicação a que se destinam, os codificadores podem ser projetados tendo em vista a minimização de alguns deles. Em sistemas onde se precisa transmitir o sinal através de canais de largura de banda limitada, por exemplo, é interessante minimizar a taxa de bits do codificador; em situações onde a qualidade do sinal é importante, minimiza-se a degradação; já em aplicações de tempo real, é importante reduzir ao máximo o atraso do sinal.

Em princípio, um mecanismo de codificação pode ser aplicado a qualquer tipo de sinal. No entanto, existem técnicas específicas que foram desenvolvidas especialmente para a codificação eficiente de sinais de fala. Dentre os modelos mais utilizados, destacam-se os modelos de *predição linear* [7][54], como o vocoder LPC, o modelo multipulso (MPLPC-multipulse linear predictive coder) e o modelo estocástico (CELP - code excited linear prediction).

### 2.1.2 Síntese de fala

A síntese de fala pode ser definida como a utilização de mecanismos artificiais para a produção de um sinal de fala. Um sinal de fala sintetizado é um sinal gerado artificialmente, no qual procura-se reproduzir ao máximo as características da voz humana. Os sistemas de fala sintética diferenciam-se entre si em certos atributos. O vocabulário com o qual eles trabalham funciona como um indicador da *flexibilidade* do sistema. Aplicações específicas podem trabalhar com um vocabulário bastante limitado; já os sistemas de conversão texto-fala, por exemplo, devem ser capazes, pelo menos em princípio, de gerar qualquer tipo de sentença em uma determinada língua, e por isso trabalham com vocabulários bastante extensos. Outras características, como a inteligibilidade e a naturalidade do sinal de fala gerado, bem como o custo computacional, diferenciam os sistemas de fala sintética entre si.

Os sistemas de síntese de fala serão discutidos em detalhes nas seções subseqüentes deste trabalho.

### 2.1.3 Reconhecimento de fala

Pode-se entender o reconhecimento de fala como sendo um mecanismo que, ao analisar um sinal de fala, procura determinar a seqüência de palavras correspondente àquela

realização sonora. As aplicações de um sistema desse tipo são inúmeras: ele oferece, por exemplo, a possibilidade de interagir com uma máquina através de comandos de voz, ou de obter acesso automático a informações através da linha telefônica.

Dentro da área de reconhecimento de fala encontramos classes de problemas bastante distintos a serem resolvidos. Primeiramente, os sistemas de reconhecimento podem ser classificados de acordo com o vocabulário com o qual trabalham: podemos ter sistemas que trabalham com algumas dezenas ou centenas de palavras, ou então sistemas mais gerais, chamados sistemas de ditado, onde o vocabulário é formado por algumas dezenas de milhares de palavras. Também podemos classificar os sistemas de reconhecimento pela forma de elocução: *no reconhecimento de palavras isoladas*, as palavras devem ocorrer sempre separadas por pausas; *no reconhecimento de palavras conectadas*, as palavras podem ser pronunciadas sem pausas entre si, mas o vocabulário e o conjunto de combinações de palavras são limitados (por exemplo, o reconhecimento de números de telefone); já no caso do *reconhecimento de fala contínua*, não há restrições quanto ao vocabulário e as palavras podem ser combinadas livremente, sem a necessidade de pronúncia pausada. Por fim, podemos classificar os sistemas de reconhecimento como *dependentes ou independentes de locutor*; no primeiro caso, ele é capaz de reconhecer palavras pronunciadas por uma única pessoa, ao passo que, no segundo, o sistema pode lidar com a voz de vários indivíduos. Devido a essa diversidade de problemas, diferentes estratégias de solução podem ser aplicadas, adequadas a cada tipo de problema [69].

Os sistemas de reconhecimento de fala são extremamente custosos do ponto de vista computacional, e o seu desempenho é altamente dependente do tipo de reconhecimento envolvido (o reconhecimento de fala contínua, por exemplo, requer muito mais capacidade de processamento e apresenta uma taxa de acerto bastante inferior quando comparado ao reconhecimento de palavras isoladas). Os modelos mais utilizados na implementação dos sistemas de reconhecimento são os modelos ocultos de Markov (HMM - Hidden Markov

Models), as redes neurais (ANN - artificial neural networks), e os modelos híbridos, que utilizam uma combinação dos dois modelos anteriormente citados [41][53].

O grau de desenvolvimento dos sistemas de reconhecimento de fala ainda é bastante pequeno quando comparado a outros campos das ciências da fala, como por exemplo a codificação e a síntese. Isso se deve em parte à extrema complexidade do problema de reconhecimento e ao fato de este ser um campo de estudos recente, mas também à falta de interação com outros campos do conhecimento, como a Lingüística, visto que o reconhecimento de fala é visto, hoje em dia, como um problema quase que exclusivo da área da Engenharia.

### 2.1.4 Reconhecimento de locutor

Diferentemente do reconhecimento de fala, que se preocupa em extrair o conteúdo da mensagem falada, no *reconhecimento de locutor* procura-se identificar informações presentes no sinal de fala que permitam associá-lo ao indivíduo que produziu esse sinal. O problema do reconhecimento de locutor pode ser dividido em duas classes distintas: *verificação* e *identificação* de locutor.

Um sistema de verificação de locutor [56] recebe como entrada um sinal de voz, e deve ser capaz de decidir se esse sinal corresponde ou não a uma determinada pessoa. O sistema possui um banco de dados onde estão armazenados padrões de voz de diferentes indivíduos. O usuário deve, de alguma maneira, identificar-se ao sistema, que por sua vez selecionará dentre os padrões armazenados aquele que corresponde à identificação que lhe foi fornecida. Efetua-se então o processo de verificação, no qual o sistema deve comparar o padrão armazenado com o sinal de voz do usuário, e decidir se eles correspondem ou não à mesma pessoa.

Um sistema de identificação de locutor [8], por sua vez, não exige que o usuário anuncie sua identidade. Ele é capaz de identificá-lo apenas pela análise de seu sinal de voz.

Para isso, esse sinal deve ser comparado com todos os padrões de voz armazenados no sistema, que então decidirá qual deles se aproxima mais do sinal de voz emitido pelo usuário. Obviamente, este é um problema mais difícil de ser resolvido do que a simples verificação do locutor, e o desempenho do sistema depende fortemente do número de locutores com o qual é capaz de trabalhar.

### 2.1.5 Outras áreas de estudo

Como outras áreas de estudo do processamento de fala podemos ainda citar o *reconhecimento de língua* [48], onde o sistema deve ser capaz de identificar, a partir do sinal de voz, a língua utilizada pelo locutor (inglês, português, francês, etc.). Há ainda os sistemas de *tradução automática*. O problema da tradução automática consiste em receber como entrada uma mensagem falada, numa determinada língua, e gerar como saída outra mensagem falada, com significado equivalente mas em uma língua diferente. Para alcançar esse objetivo, o sistema de tradução deve conter um módulo de reconhecimento (para interpretar a mensagem de entrada) e um módulo de síntese (para gerar a mensagem de saída). Esses módulos devem ser bilíngües (num sistema de tradução que envolva as línguas inglesa e portuguesa, por exemplo, o sistema deve ser capaz de fazer a tradução do inglês para o português, bem como do português para o inglês). Além disso, a mensagem interpretada pelo reconhecedor deve ser mapeada para uma *representação abstrata*, representação esta dependente, normalmente, apenas do significado da mensagem, e não da língua de entrada ou de saída. A partir dessa representação é que o módulo de síntese irá trabalhar para gerar o sinal de voz de saída.

Como vimos, são muitas as áreas de estudo dentro do processamento da fala. Algumas dessas áreas já foram razoavelmente bem estudadas, enquanto que outras ainda se encontram "engatinhando", sob o ponto de vista da produção científica. Sistemas de reconhecimento de fala contínua trabalhando em tempo real com vocabulário extenso e baixas taxas de erro, ou

sistemas de tradução automática [34] eficientes que funcionem em tempo real, por exemplo, ainda estão longe de ser implementados com sucesso.

Este trabalho pretende apresentar um sistema de conversão texto-fala para o português do Brasil. Dentre as diversas áreas de estudo acima apresentadas, portanto, estaremos concentrando nossa atenção na *síntese de fala* e, particularmente, na *conversão texto-fala*, que discutiremos com detalhes ao longo do restante deste trabalho.

## 2.2 Síntese de fala

De maneira geral, pode se definir um sistema de *síntese de fala* como sendo um sistema capaz de produzir sinais de fala de maneira artificial. Existem diferentes estratégias que podem ser utilizadas para a implementação de sistemas desse tipo; a escolha da estratégia adequada depende fundamentalmente das características do sistema a ser implementado.

A qualidade do sinal de fala a ser gerado é um dos fatores importantes na determinação da estratégia de implementação. Certas aplicações exigem apenas que o sinal de fala produzido seja inteligível, ao passo que outras necessitam que o sinal se aproxime o máximo possível da fala natural. O tamanho do vocabulário com o qual o sistema de síntese trabalha também representa um fator de diferenciação. Os sistemas mais simples trabalham com vocabulários fixos e de tamanho reduzido; sistemas de uso mais geral, por sua vez (como no caso dos sistemas de conversão texto-fala), trabalham com vocabulários bastante extensos. Há ainda casos em que o vocabulário, apesar de não ser muito grande, pode apresentar uma característica *dinâmica*, ou seja, as mensagens a serem geradas não são sempre as mesmas (a título de exemplo, podemos citar um sistema de acesso a informações contidas num banco de dados dinâmico).

Outro fator importante na diferenciação dos sistemas de síntese entre si é a velocidade de execução, especialmente crítica no caso dos sistemas que trabalham em tempo real.

Podemos citar, por fim, o custo do sistema como um todo: quanto maior for a capacidade de processamento e de armazenamento, mais alto será o custo do hardware necessário para implementar o sistema.

A Figura 2-1 ilustra as idéias que serão expostas a seguir, a respeito das diferentes classes de problemas envolvidos na síntese de fala.

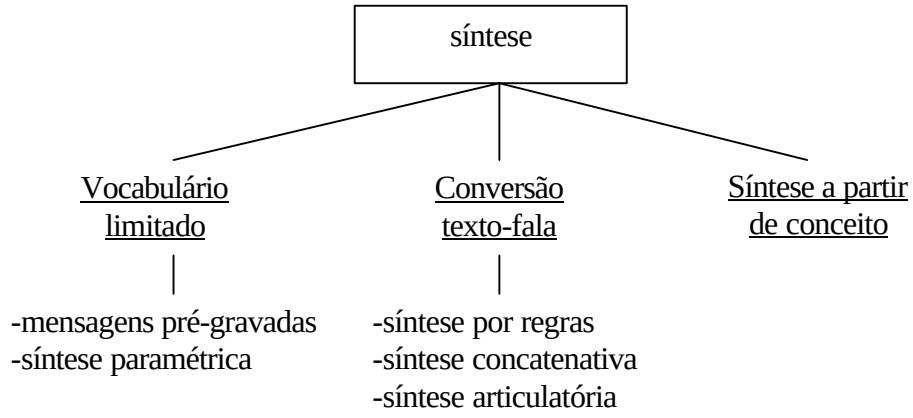



**Figura 2-1 Síntese de fala e seus diversos aspectos**

A maneira mais elementar de produzir um sinal de fala consiste em simplesmente reproduzir trechos de mensagem pré-gravados. Nesse caso, para gerar uma sentença, o sistema seleciona e reproduz uma seqüência de uma ou mais mensagens armazenadas previamente. Esse tipo de estratégia possui a vantagem de ser extremamente simples de implementar: a única tarefa do algoritmo de síntese é a de selecionar a seqüência de mensagens a ser reproduzida. Além disso, a qualidade do sinal de voz gerado é muito boa, pois o que se tem na verdade é um sinal de fala natural. Outra vantagem é que o sistema apresenta um tempo de resposta bastante curto, pois não existe quase nenhum tipo de processamento a ser executado: toda a tarefa consiste em selecionar a seqüência adequada de mensagens.

No entanto, esse tipo de estratégia peca pela sua falta de flexibilidade. O número de sentenças que podem ser geradas é pequeno, consistindo basicamente da combinação das

mensagens pré-gravadas entre si. Além disso, não é possível efetuar nenhum tipo de alteração prosódica na sentença gerada (alterações prosódicas são modificações nos parâmetros de duração, freqüência fundamental (F0) e amplitude ao longo da sentença, essenciais para garantir a naturalidade das frases sintetizadas). Por fim, o custo de armazenamento necessário para implementar um sistema desse tipo é alto. Num sistema computacional, por exemplo, é preciso armazenar cada uma das mensagens sob forma digital.

Esse tipo de estratégia se mostra suficiente para algumas aplicações mais simples, como por exemplo um sistema de acesso a saldos bancários por telefone. Nesse caso o vocabulário seria composto por algumas frases introdutórias, como *"Bom dia"*, *"Digite sua senha"*, *"Obrigado"*, etc., bem como por um conjunto de palavras a partir das quais seriam formados os valores dos saldos (*"um"*, *"dois"*, *"vinte"*, *"milhões"*, *"centavos"*, etc.). Muito embora o resultado da leitura seja artificial, pois a concatenação das mensagens é feita sem alteração prosódica, o resultado da síntese é perfeitamente aceitável. Para sistemas mais complexos e com vocabulários maiores, no entanto, a estratégia acima descrita se torna inviável.

Uma outra estratégia utilizada na geração de sinais de fala sintetizada é conhecida como *síntese paramétrica*. Diferentemente do caso anterior, tem-se agora uma biblioteca de palavras armazenadas sob forma parametrizada. Essa parametrização consiste em extrair, a um intervalo de tempo fixo, valores dos parâmetros do trato vocal, bem como o valor de *pitch* (taxa de vibração das pregas vocais) do sinal de fala. Existem diferentes maneiras a partir das quais pode-se obter essa parametrização (por exemplo, através de coeficientes LPC ou parâmetros cepstrais).

Nesse caso, portanto, o sintetizador é constituído por um filtro digital, cuja função de transferência, variante no tempo, é construída a partir dos parâmetros relativos às freqüências de ressonância do trato vocal. A entrada desse filtro pode ser um ruído branco no caso dos sons surdos ou, no caso dos sinais sonoros, um trem de impulsos com espaçamento

equivalente ao período de *pitch* do sinal de fala [29]. O diagrama de blocos da Figura 2-2 ilustra o sistema de síntese acima descrito.

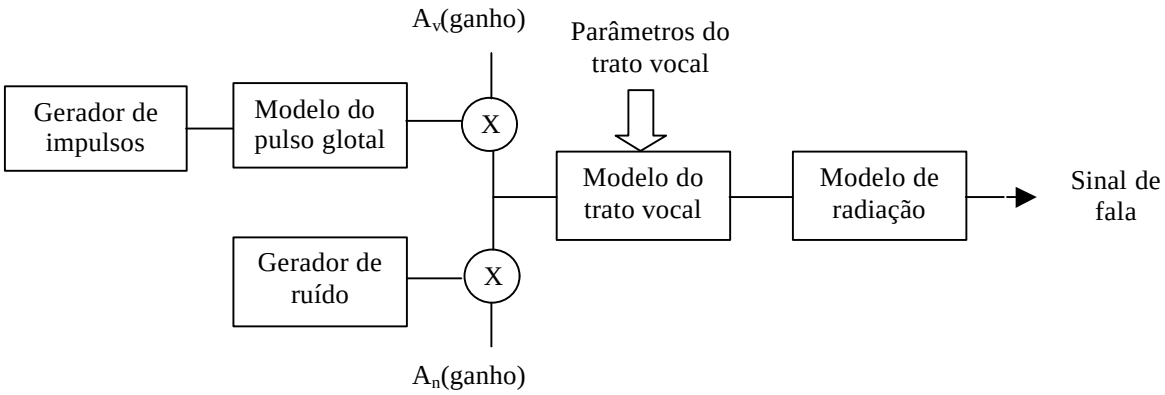



**Figura 2-2 Modelo do sintetizador paramétrico**

Uma das vantagens de parametrizar as mensagens pré-gravadas é a redução do custo de armazenamento. No caso do armazenamento direto da forma de onda, são necessários cerca de 50000 bits para cada segundo de fala digitalizada, supondo-se uma limitação em faixa de 4kHz e que a amostragem seja feita à taxa de Nyquist, com quantização a 7 bits em escala logarítmica. No caso do sinal parametrizado, a economia de armazenamento depende da precisão com que os parâmetros estão especificados. Estudos de caráter perceptual mostram que os parâmetros podem ser amostrados a uma taxa de aproximadamente 35 Hz sem que haja degradação considerável na qualidade do sinal; além disso, o *pitch* pode ser quantizado com cerca de 6 bits e os demais parâmetros com cerca de 4 bits [29]. Isso leva a uma taxa da ordem de 1000 bits/s, o que representa uma redução de 50:1 em relação ao armazenamento direto da forma de onda.

Outra vantagem obtida por meio da parametrização é a possibilidade de efetuar alterações prosódicas nas mensagens armazenadas a partir da manipulação adequada de seus

parâmetros. Para alterar o *pitch* e a intensidade do sinal, basta variar o espaçamento do trem de impulsos (no caso dos sinais sonoros) e o ganho do sinal de entrada, respectivamente. Para alterar a duração, os parâmetros do filtro devem ser atualizados não a intervalos de tempo regulares, mas de acordo com o ritmo que se queira imprimir. Por fim, é possível garantir uma transição suave nas fronteiras de palavras, através do ajuste gradual dos parâmetros na transição de uma palavra para outra. Dessa forma, é possível melhorar de forma considerável a naturalidade do sinal de fala sintetizado.

Obviamente a estratégia acima descrita é mais flexível do que a reprodução pura e simples de mensagens pré-gravadas. De qualquer forma, o número de sentenças que podem ser geradas a partir de mensagens parametrizadas ainda é bastante limitado. Se por um lado economiza-se muito em custo de armazenamento, por outro ocorre uma degradação da qualidade do sinal de fala sintetizado, devida à perda de informação inerente ao processo de codificação e decodificação. A complexidade do algoritmo de síntese e a maior quantidade de processamento causam também uma degradação do tempo de resposta do sistema.

Em muitas situações, as soluções apresentadas anteriormente se mostram totalmente inadequadas. No caso de grandes vocabulários ou de vocabulários irrestritos, por exemplo, é inviável armazenar todas as mensagens, mesmo sob forma parametrizada. Mesmo no caso de vocabulários menores, mas que sejam constantemente modificados, seria preciso atualizar constantemente o inventário de mensagens disponíveis, o que sem dúvida seria muito pouco prático.

Os sistemas de conversão texto-fala são capazes de gerar fala sintetizada a partir de uma mensagem escrita. A utilização desse tipo de sistema é extremamente abrangente pois, em princípio, qualquer tipo de mensagem pode ser representada através de texto e, portanto qualquer tipo de mensagem pode ser sintetizada. Além disso, o custo de armazenamento do sistema é infinitamente menor: um segundo de fala armazenada em formato textual requer aproximadamente 75 bits, em oposição aos cerca de 1000 bits necessários para armazenar fala parametrizada, ou aos 50000 bits utilizados para guardar a forma de onda digitalizada. Alguns

sistemas de conversão texto-fala utilizam-se de segmentos de fala pré-gravados menores do que palavras como base para a geração do sinal de fala sintética. No entanto o número de segmentos que compõem essa base não é grande, e o tamanho de cada segmento é reduzido, o que faz com que o custo de armazenamento da base não seja crítico.

Anteriormente à etapa de processamento de sinal, que inclui a síntese do sinal propriamente dita em conjunto com as modificações prosódicas apropriadas, é necessário realizar também uma etapa de *processamento lingüístico*. O processamento lingüístico é responsável pela geração da *representação fonológica* do texto a ser sintetizado; além disso, é preciso também calcular os *parâmetros prosódicos* (F0, duração e intensidade) relativos a cada um dos segmentos fonéticos da representação. A necessidade de efetuar cada uma dessas etapas faz com que o custo de processamento de um sistema de conversão texto-fala seja bastante elevado; portanto, a otimização do tempo de execução de cada um dos passos do algoritmo é essencial para a implementação de sistemas que trabalhem em tempo real.

A qualidade do sinal de voz sintetizado por um sistema de conversão texto-fala geralmente é inferior àquela gerada por meio das estratégias anteriormente citadas. Uma das razões que levam a isso é o fato de que nem sempre o módulo de processamento lingüístico é capaz de fornecer a transcrição fonética correta de todas as palavras do texto. O cálculo dos parâmetros prosódicos também é crítico: no caso da conversão texto-fala, toda a informação prosódica deve ser calculada a partir do texto, ao passo que nas estratégias de síntese anteriormente descritas uma grande parte da prosódia já se encontra incorporada ao sinal de fala previamente armazenado. Por fim, o cálculo do sinal de fala propriamente dito é inerentemente mais complexo, causando assim uma degradação da qualidade do sinal de fala gerado. Nesse caso, o nível de degradação depende do mecanismo de síntese utilizado. Os métodos de síntese mais utilizados são a *síntese por regras*, a *síntese concatenativa* e a *síntese articulatória*. Cada um deles será discutido com mais detalhes em seções posteriores deste trabalho.

Os sistemas de conversão texto-fala são extremamente complexos, por isso as suas diversas particularidades estarão sendo discutidas ao longo do restante deste trabalho.

Uma alternativa interessante aos sistemas de conversão texto-fala são os *sistemas de síntese a partir de conceito* [73]. A idéia principal atrás desse tipo de síntese é a de reunir as principais vantagens da conversão texto-fala, como a alta flexibilidade no que diz respeito ao conjunto de sentenças sintetizáveis e o baixo custo de armazenamento, sem que haja a necessidade de que a informação a ser transformada em sinal de fala esteja representada em forma de texto. Em muitos aplicações, por exemplo, a informação a ser manipulada não é necessariamente textual, e um sistema de conversão texto-fala precisaria transformar essa informação em texto antes de poder gerar a saída em forma de voz. Essa transformação, além de desperdiçar tempo, poderia causar perda de informação e, conseqüentemente, degradação na qualidade da saída produzida.

Num sistema de síntese a partir de conceito a informação a ser manipulada é mapeada para um *conceito de entrada.* O *conceito de entrada* é uma estrutura de formato pré-definido, cujo objetivo principal é padronizar o formato de representação da informação, para que esta possa ser manipulada adequadamente pelos estágios subseqüentes do processo de síntese.

A principal vantagem dessa tática é que o sistema de síntese consegue gerar, a partir do conceito de entrada, sentenças com estrutura sintática definida, o que facilita o cálculo de fronteiras prosódicas e a determinação de padrões de ritmo e entonação para essas sentenças. Conforme veremos adiante, a análise sintática constitui uma das tarefas mais complexas do módulo de processamento lingüístico de um sistema de conversão texto-fala.

Mais do que isso, as sentenças geradas são formadas por palavras pertencentes ao vocabulário do sistema, cuja transcrição fonética já é conhecida, o que evita erros de transcrição. Esse tipo de erro costuma ocorrer em sistemas de conversão texto-fala quando são usados algoritmos para a determinação da transcrição fonética de forma automática.

Obviamente, a utilização de sistemas de síntese a partir de conceito se limita a aplicações onde a informação esteja representada de forma não textual. Além disso, a flexibilidade de um sistema de síntese a partir de conceito não é a mesma de um sistema de conversão texto-fala, pois não é possível gerar qualquer tipo de sentença, mas tão somente aquelas que possam ser mapeadas para um conceito de entrada.

## 2.3 Aplicações da síntese de fala

O papel básico das aplicações baseadas em síntese de fala é o de facilitar a interação do ser humano com a máquina. Em certas situações, a utilização da voz torna o processo de comunicação mais ágil e mais natural, justificando-se assim o interesse por esse tipo de tecnologia.

A síntese de fala torna viável, por exemplo, a utilização do computador em situações nas quais os olhos estejam ocupados com o monitoramento de outras tarefas. Um automóvel equipado com um computador de bordo falante, por exemplo, pode emitir mensagens de alerta, confirmar um evento ou fornecer informações de forma instantânea (velocidade, nível de combustível, etc.) sem que o motorista precise desviar sua atenção do volante. Outro exemplo: numa linha de produção, um trabalhador pode seguir instruções fornecidas através de fala sintetizada, sem precisar desviar sua atenção da tarefa que está executando.

Outra vantagem importante dos sistemas que se utilizam da síntese de fala é a possibilidade de acessar informações através da linha telefônica. Nesse caso, o sistema de síntese fornece a informação requisitada na forma de uma mensagem falada. As aplicações que seguem esta linha podem ser as mais variadas possíveis: consulta de saldos bancários ou de horários de vôos, pesquisa de itens catalogados, ou mesmo acesso a informações de natureza geral, como previsão do tempo, horóscopo, calendário de eventos culturais e esportivos, etc. Muito mais do que acessar informações, o usuário pode também interagir com o sistema, sem a necessidade de estar junto a um terminal de computador [72]. No caso de transações bancárias, por exemplo, o telefone pode funcionar como um autêntico *caixa*

*automático*, permitindo não só consultas de saldos, mas também operações como depósitos, transferências e aplicações.

A interação com sistemas do tipo acima descrito poderia ser feita através do teclado do telefone. Neste e em muitos outros casos, no entanto, é mais interessante que, aliado ao módulo de síntese, o sistema contenha também um módulo de reconhecimento de fala, que permita ao usuário fornecer informações ao sistema através de comandos de voz, o que sem dúvida tornaria a interação mais natural. Um sistema como o descrito em [30], por exemplo, permite fazer reservas de passagens de avião, baseando a escolha nas melhores opções de preços , horários e trajetos. Ainda a título de exemplo, o computador de bordo anteriormente descrito poderia aceitar comandos de voz, o que permitiria que tarefas como ligar o rádio ou verificar o nível de combustível fossem acionadas a pedido do usuário.

A síntese de fala pode ainda ser utilizada como uma ferramenta importante no auxílio a deficientes. No caso de deficientes visuais, por exemplo, ela permite que o computador se comunique com o usuário por meio de voz, gerando relatórios de eventos, confirmações de comandos, mensagens de erro e outros tipos de resposta na forma de fala sintetizada. Outra aplicação importante nessa linha é a máquina de leitura para cegos. Um exemplo de sistema desse tipo é a máquina de Kurzweil [39], que utiliza texto impresso como entrada e gera como saída o sinal de fala correspondente. Além do módulo de conversão texto-fala, uma máquina desse tipo necessita de um módulo de reconhecimento de caracteres, que faça a transformação do texto impresso em uma representação manipulável pelo módulo de conversão, como por exemplo o formato ASCII.

A síntese de fala pode ser utilizada também no auxílio a deficientes vocais. Um sistema de conversão texto-fala adequado a esse propósito deve utilizar um mecanismo bastante eficiente para a entrada de texto, que permita ao usuário "conversar" numa velocidade adequada. Uma alternativa para acelerar a entrada de texto pode ser um sistema de *entrada preditiva*, que mostre a palavra mais provável deduzida a partir do fragmento de texto até então digitado. Se a palavra prevista for correta, o usuário pode selecioná-la sem precisar

terminar a digitação. Outra alternativa é a utilização de teclados adaptados, que possuam teclas específicas correspondentes às palavras e/ou conceitos mais freqüentes [13]. Para que um sistema desse tipo seja verdadeiramente útil, é necessário que o módulo de conversão texto-fala seja rápido o suficiente para trabalhar em tempo real. Ainda na linha de auxílio a deficientes vocais, pode-se utilizar um sistema de conversão texto-fala no auxílio ao aprendizado de pessoas vítimas de dislexia (incapacidade, devido a lesão central, para ler compreensivelmente).

No campo didático e de pesquisa, os sistemas de síntese de fala podem ser utilizados como ferramentas para o aprendizado da língua, ou ainda como instrumentos importantes na avaliação de teorias acerca do processo de produção da fala. Por fim, podemos pensar em outras aplicações de uso geral, como a leitura de correio eletrônico, leitura de páginas da Web ou de *chats*, revisão de textos, jogos de computador, etc.

## 2.4 Síntese de fala através da história

O interesse do homem acerca dos mecanismos de produção da fala é bastante antigo. Essa curiosidade permitiu ao homem adquirir conhecimento suficiente para que, há pouco mais de duzentos anos, começassem a surgir as primeiras tentativas de produzir sinais de fala por meios artificiais.

Um dos primeiros passos nesse sentido foi dado por von Kempelen em 1769 [17]. Nesse ano ele iniciou a construção de um dispositivo que mais tarde ficaria conhecido como a "máquina falante de von Kempelen"(Figura 2-3). O princípio de funcionamento dessa máquina era inteiramente mecânico: ela era dotada de um fole que funcionava como fonte de ar para uma caixa de ressonância; esse ressoador, por sua vez, podia ser controlado manualmente, de forma a simular o som das diversas vogais. Para produzir os sons das

consoantes, existiam quatros constrições ao longo da passagem de ar, as quais eram controladas pelos dedos da outra mão do operador.

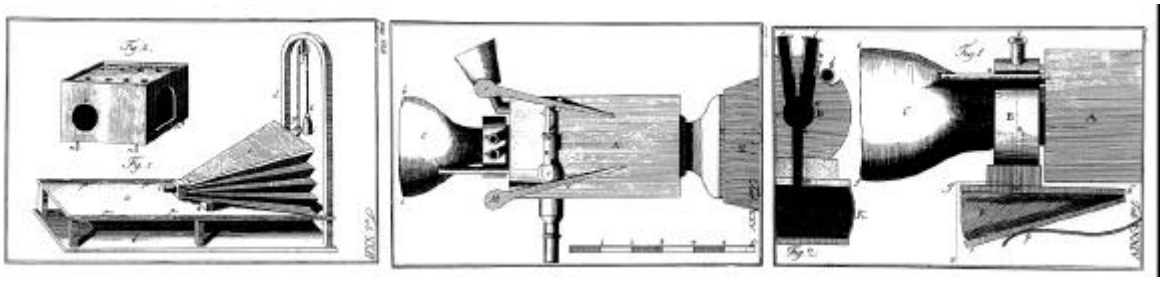


**Figura 2-3 - Máquina falante de von Kempelen**

Dez anos mais tarde, o alemão Kratzenstein construiu um aparelho capaz de reproduzir os sons das vogais "a, e, i, o, u". Esse aparelho foi construído em função de um concurso instituído pela Academia Imperial de São Petersburgo, e valeu o primeiro prêmio ao alemão. O dispositivo era constituído por cinco cavidades ressonantes excitadas por uma palheta vibrante. O formato das cavidades determinava a vogal produzida.

Já no século 20, mais precisamente em 1922, Stewart foi o responsável pelo surgimento do primeiro dispositivo elétrico capaz de gerar alguns sons de fala sintética [68]. Esse dispositivo consistia de dois circuitos ressoadores excitados por um sinal sonoro de entrada: ajustando-se as freqüências de ressonância dos dois circuitos, podia-se simular o som de cada uma das vogais, desde que as freqüências de ressonância se aproximassem das freqüências dos dois primeiros formantes da vogal correspondente.

No entanto, o primeiro grande marco na história da síntese de fala ocorreu sem dúvida no ano de 1939. Nessa ocasião Dudley, engenheiro da Bell Laboratories, apresentou à comunidade científica o sintetizador de fala por ele desenvolvido, denominado *Voder* [18]. O

*Voder*, ilustrado pela Figura 2-4, foi inspirado a partir de um sistema de análise do sinal de fala, também desenvolvido por Dudley, denominado de *Vocoder* [19]. O *Vocoder* decompunha o sinal de fala em diversos parâmetros acústicos que variavam lentamente ao longo do tempo. A idéia por trás do *Voder* consistia em utilizar esses parâmetros como variáveis de controle do sintetizador de fala. Esse sintetizador consistia de uma chave de controle que permitia selecionar o sinal de entrada (sonoro ou ruidoso); um pedal para controle da freqüência fundamental; um teclado a partir do qual o operador controlava a amplitude de dez filtros passa-banda pelos quais passava o sinal de entrada, e um amplificador na saída do sinal sintetizado. Para fazer com que o *Voder* gerasse uma sentença, era necessário que o operador tivesse bastante treino e habilidade no manejo do equipamento (a operadora responsável pela demonstração do equipamento na feira de 1939 em Nova York precisou de um ano de treinamento!). Muito embora a inteligibilidade do sinal de fala gerado fosse pequena, as idéias ali apresentadas serviram como base para muitos dos sistemas de síntese desenvolvidos posteriormente.

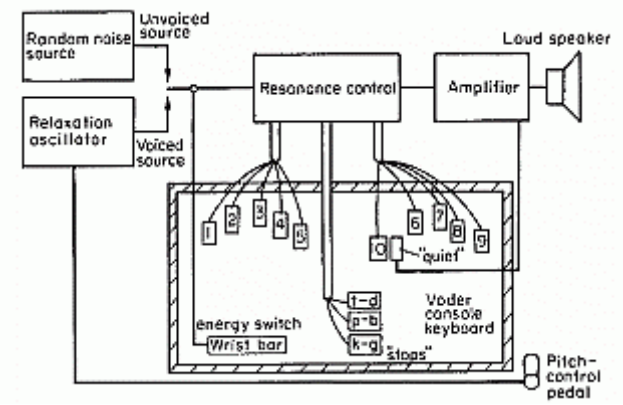



**Figura 2-4 - O *Voder***

O surgimento do espectrógrafo em 1946 permitiu os primeiros estudos a respeito das características acústicas do sinal de fala. Algum tempo depois (1950), surgiu um aparelho, desenvolvido pelo laboratório Haskins, denominado de *Pattern Playback* [16]. Ilustrado na Figura 2-5, o *Pattern Playback* efetuava a leitura ótica dos padrões desenhados em uma correia transparente, correspondentes a um espectrograma de banda larga, transformando-os em sinal sonoro. A partir da análise de espectrogramas verdadeiros ou de padrões estilizados desenhados manualmente era possível encontrar pistas a respeito do papel desempenhado pelos vários parâmetros atuantes no processo de produção e de percepção do sinal de fala.

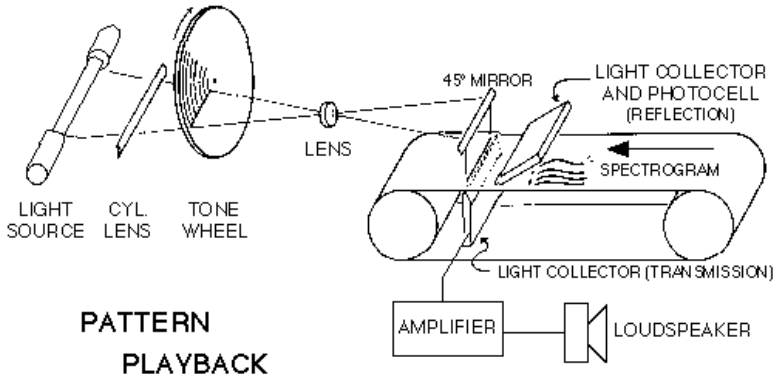



**Figura 2-5 - O *Pattern Playback***

A segunda metade do século XX pode ser considerada como sendo o período onde ocorreram os maiores avanços na área de síntese de fala. Uma das razões para esse fenômeno foi o surgimento da Teoria Acústica da Produção da Fala, formalizada por Fant em 1960 [24]. Essa teoria, que discutiremos em maiores detalhes no capítulo seguinte, serviu como base para a maioria dos sintetizadores que surgiram em seguida, particularmente aqueles baseados na síntese por formantes e na síntese articulatória.

O segundo motivo para o crescimento rápido na área de síntese da fala foi o aparecimento dos computadores e da tecnologia digital. Além de "tornar possível o estudo de aspectos lingüísticos e semânticos do sinal de fala" [67], o surgimento dos computadores

permitiu a utilização de técnicas que anteriormente seriam impensáveis do ponto de vista prático, bem como o surgimento de novas soluções, baseadas em técnicas de processamento digital do sinal de fala.

No ano de 1953 surgiram os primeiros sintetizadores por formantes, baseados na teoria acústica de Fant. O *PAT* (Parametric Artificial Talker), desenvolvido por Lawrence, consistia de uma associação de filtros passa-banda em paralelo, alimentados por um sinal que podia ser sonoro ou ruidoso [40].A associação de filtros em paralelo correspondia à utilização de um modelo do trato vocal contendo pólos e zeros. O *OVE I* (Orator Verbis Eletricis), desenvolvido por Fant , seguia o mesmo princípio do *PAT*, mas utilizava uma associação de filtros em série [23] (modelo do trato vocal contendo apenas pólos). No ano de 1973, Holmes utilizou um sintetizador por formantes em paralelo para mostrar que um sinal de fala sintetizado pode ter qualidade suficiente para se tornar indistinguível de um sinal de fala original [33]. Para isso ele ajustou cuidadosamente seu sintetizador para produzir uma frase simples (demorou um verão para fazer isso), e comparou o sinal obtido com o sinal original. Ao final, confirmou que um ouvinte humano não conseguia perceber a diferença entre os dois.

Vale aqui ressaltar a semelhança entre a experiência feita por Holmes e o famoso teste de Turing, que consistiu em fazer com que um operador, fechado em uma sala, tentasse descobrir se quem respondia suas perguntas, fornecidas a partir de um teclado, era outro homem ou uma máquina. A questão levantada por tal teste, que ainda constitui um dos pontos centrais de discussão entre os estudiosos da Inteligência Artificial, consiste em saber se podemos atribuir à máquina que passa no teste alguma noção de inteligência.

Mais recentemente, surgiram alguns novos métodos de síntese de fala além da síntese por formantes. A *síntese concatenativa* vem sendo utilizada com sucesso em diversos sistemas de síntese, produzindo sinais de fala de alta qualidade. A *síntese articulatória*, por sua vez, ainda se encontra em fase de estudos, e apesar de ainda não estar sendo utilizada na prática, parece apontar para um caminho bastante promissor.

# 3 Alguns aspectos da produção da fala

## 3.1 Teoria acústica de produção da fala

A teoria acústica de produção da fala procura modelar matematicamente o processo de geração do sinal de fala pelo aparelho fonador humano. As bases dessa teoria foram brilhantemente apresentadas por Fant em 1960, através da publicação de *"Acoustic Theory of Speech Production"* [24].

Podemos dividir o aparelho fonador humano (Figura 3-1) em três componentes principais:

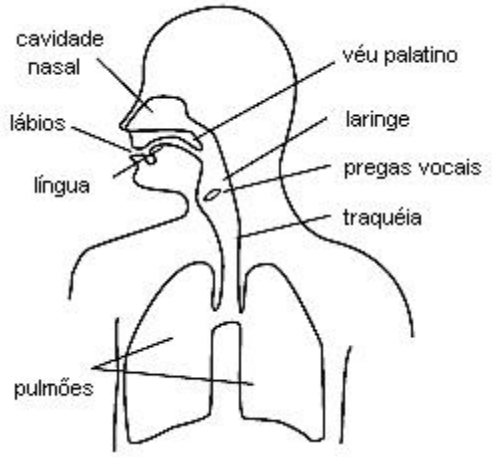



**Figura 3-1 Aparelho fonador humano.**

**pulmões:** localizados no interior da caixa torácica, os pulmões controlam a intensidade do fluxo de ar que passa pela laringe.

**laringe:** localiza-se logo acima da traquéia, e é formada por cartilagens e tecido muscular. Dentre os músculos mais importantes podemos destacar as *pregas vocais,* que representam um papel fundamental no processo de produção da fala. As pregas vocais são formadas por dois pares de músculos. Durante a respiração normal as pregas estão relaxadas e abertas; no processo de produção de voz, no entanto, as pregas se tensionam e vibram com a passagem do ar. A taxa de vibração das pregas vocais está diretamente relacionada com a freqüência fundamental (grave/agudo) do sinal de voz: nos sons mais agudos, as pregas estão mais contraídas e portanto vibram mais depressa. O comprimento das pregas também influi na taxa de vibração; é por isso que as mulheres, cujas pregas vocais são mais curtas que as dos homens, possuem um tom de voz normalmente mais agudo.

**trato vocal:** porção do aparelho fonador humano que se estende desde a glote até os lábios (Figura 3-2). Os diversos elementos formadores do trato vocal são denominados de *articuladores.* O trato vocal funciona como uma *caixa de ressonância,* que atenua ou amplifica certas freqüências do pulso produzido na glote (laringe). O movimento dos articuladores determina o formato do trato vocal e, por conseguinte, as suas características de ressonância.

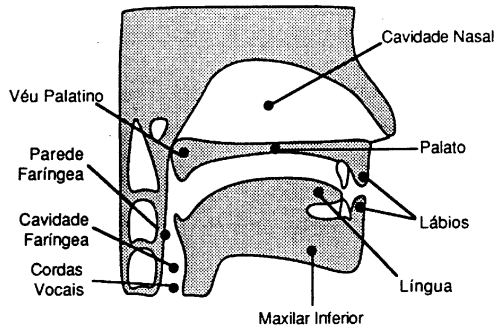



**Figura 3-2 Trato vocal**

A maneira mais simples de modelar o trato vocal consiste em considerá-lo como sendo um tubo cilíndrico com seção transversal de área uniforme, com uma extremidade aberta correspondente aos lábios e uma fonte de excitação sonora na outra extremidade. Um sistema como o descrito acima funciona como uma caixa de ressonância, onde certas freqüências do sinal sonoro gerado na entrada do tubo são amplificadas, ao passo que outras são atenuadas. As freqüências em que ocorre ressonância são dependentes do comprimento do tubo: no caso de um tubo de comprimento $L$, as ressonâncias ocorrem para os comprimentos de onda:

$$\lambda = 4L,\ 4L/3,\ 4L/5,\ 4L/7, \text{ etc.},$$

os quais correspondem às freqüências:

$$f = c/4L,\ 3c/4L,\ 5c/4L,\ 7c/4L, \text{ etc.},$$

onde $c$ é igual à velocidade de propagação do som no meio em questão. Considerando-se um valor de $L$ igual a 17cm, que é um valor típico para o comprimento do trato vocal, e fazendo-se $c$ igual a 340m/s (velocidade do som no vácuo), encontramos valores de ressonância em 500Hz, 1500HZ, 2500HZ, etc..

Essas freqüências de ressonância correspondem às freqüências onde ocorre a máxima amplificação do sinal de entrada, e são normalmente denominadas de *formantes*. Note que os valores dos formantes são independentes da fonte de excitação: seus valores dependem única e exclusivamente da configuração do trato vocal.

O espectro típico do sinal produzido na laringe é mostrado na Figura 3-3(a). Ele corresponde a uma seqüência discreta de harmônicas, onde o espaçamento entre as harmônicas é igual à freqüência fundamental. A energia dessas harmônicas tem uma queda da ordem de 12dB/oitava, por isso a maior parte da energia do sinal concentra-se nas baixas freqüências (até 10 kHz).

De acordo com o modelo fonte-filtro o sinal de fala pode ser considerado como sendo o produto do espectro em freqüência do trem de pulsos produzido na laringe pela função de transferência do trato vocal [35]. Assume-se, nesse caso, que a laringe e o trato vocal funcionam como entidades independentes. Essa é, na verdade, uma simplificação do modelo, pois na verdade existe um certo acoplamento entre a laringe e o trato vocal, o que significa que a função de transferência do filtro não é totalmente independente da fonte.

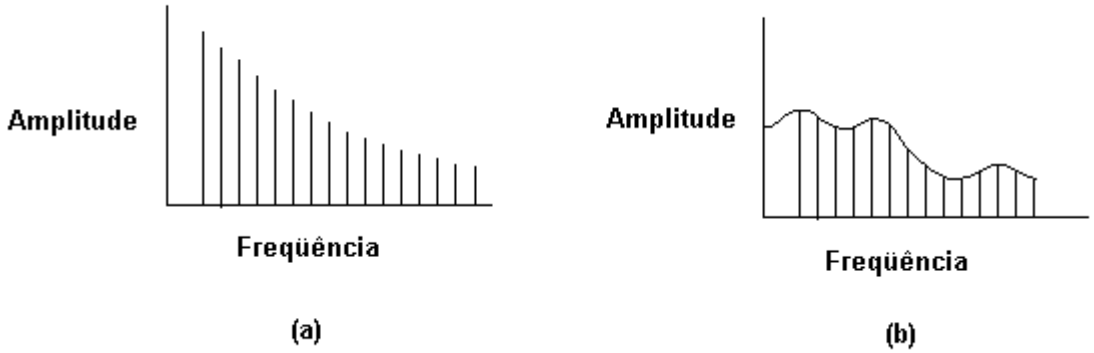



**Figura 3-3 (a) Espectro do trem de pulsos glotal (b) Espectro do trem de pulsos glotal filtrado pela função de transferência do trato vocal.**

Ao passar pelo trato vocal, portanto, o pulso produzido na laringe sofre um processo de "filtragem", conforme ilustra a Figura 3-3(b). A curva sobre o espectro representa a função de transferência do trato vocal convoluída com o espectro do sinal glotal, e os picos dessa curva correspondem às freqüências de ressonância (formantes).

Além do efeito de filtragem do trato vocal, devemos ainda levar em conta o efeito da *radiação.* Este é um fenômeno que ocorre quando o som escapa dos lábios em direção ao ambiente. O efeito de radiação é equivalente ao de um filtro passa-altas, com amplificação da ordem de 6dB/oitava; para modelá-lo, basta acrescentar um zero à função de transferência do trato vocal.

Levados em conta os aspectos acima discutidos, podemos descrever o processo de produção de fala através da seguinte equação:

$$V(f) = T(f) \cdot U(f) \cdot R(f)$$

onde *V(f)* é o espectro do sinal de fala, *U(f)* o espectro do pulso glotal, *T(f)* a função de transferência do trato vocal e *R(f)* o efeito de radiação.

O modelo do tubo uniforme é suficiente para descrever o processo de geração da vogal neutra conhecida por "*schwa"* (, cujo padrão de formantes é equivalente àquele obtido por meio do modelo. No entanto, o trato vocal humano não é rígido nem tampouco possui seção transversal de área uniforme. A movimentação dos articuladores (língua, lábios, mandíbula, etc.) durante o processo de produção da fala determina alterações na área da seção transversal ao longo do tubo. O efeito dessa alteração é a modificação do padrão de ressonância do trato: cada configuração do trato corresponde a um padrão de formantes diferente, cada um desses padrões correspondendo a uma vogal em particular.

Existem outros tipos de sons, além das vogais, que podem ser produzidos pelo aparelho fonador humano. A produção desses sons também pode ser explicada por meio do modelo fonte-filtro.

O processo de produção das *vogais nasalizadas*, por exemplo, é semelhantes ao das vogais orais; nesse caso, no entanto, ocorre uma abertura do véu palatino, permitindo a passagem de ar também pela cavidade nasal. Podemos modelar o trato vocal, nesse caso, não mais como sendo um tubo uniforme, mas sim como dois tubos em paralelo. O efeito do acoplamento desse segundo tubo é a introdução de *pólos e zeros nasais* à função de transferência do trato vocal.

As *consoantes fricativas (/f/, /v/, /s/, /ʃ/, etc.),* por sua vez, são produzidas quando ocorre uma turbulência no fluxo de ar devido à existência de uma constrição ao longo do trato vocal. As fricativas podem ser *vozeadas* ou *não-vozeadas.* No caso das não-vozeadas (ex*: /f/,*

*/s/, /ʃ/)* existe apenas o ruído de turbulência, ao passo que nas vozeadas (ex: */v/, /z/, /ʒ/)* ocorre ainda a vibração das pregas vocais. As *consoantes plosivas(/p/, /b/, /t/, /k/, etc.),* por sua vez, são produzidas por meio de uma obstrução total à passagem do ar ao longo do trato vocal, seguida de uma liberação abrupta do ar retido. Elas também podem ser classificadas como *vozeadas (*ex: */b/, /d/, /g/)* ou *não-vozeadas (*ex.: */p/, /t/, /k/)* Os fonemas aqui indicados seguem a notação do Alfabeto Fonético Internacional (IPA).

## 3.2 Considerações de natureza lingüística

Pode-se definir a Lingüística com sendo a ciência que estuda a linguagem. Este estudo é centrado normalmente nos aspectos funcionais dessa capacidade do homem mas envolve também os pontos de vista físicos como no caso da Fonética

Para melhor compreender os diversos aspectos da produção da fala, é comum utilizar alguma forma de *classificação* que permita identificar quais são as unidades básicas da fala. Essa não é uma tarefa trivial, pois a fala é constituída por sinais acústicos de natureza inerentemente contínua, e não por segmentos discretos, conforme faz supor a idéia de classificação.

A Fonética e a Fonologia são dois ramos da Lingüística que se preocupam com os aspectos acima mencionados. Elas diferem entre si no enfoque sob o qual analisam o processo de produção da fala, bem como quanto ao modo de classificação das unidades básicas da fala. Levando-se em conta essas diferenças, pode-se dizer que a Fonética e a Fonologia são ciências complementares, que desempenham um papel fundamental no estudo da fala como mecanismo de comunicação.

### 3.2.1 Fonologia

A Fonologia é o ramo da Lingüística que estuda os sons constituintes da fala segundo seu aspecto funcional. Isso significa que a Fonologia trata do papel exercido por esses constituintes dentro do sistema de organização da fala, sem se preocupar com as suas propriedades acústicas ou articulatórias.

A Fonologia estuda uma unidade básica da língua: o *fonema.* Isto significa que uma sentença qualquer numa determinada língua pode ser descrita como sendo uma seqüência de fonemas. O fonema é uma unidade abstrata, que restringe sua área de atuação a um domínio psicológico e não físico, e que portanto não apresenta características acústicas. O que diferencia um fonema de outro é o seu papel distintivo dentro da língua. Podemos dizer, por exemplo, que no português, ***/p/*** e ***/b/*** representam fonemas diferentes, pois é a partir deles que podemos distinguir palavras como ***/pasta/*** e ***/basta/.***

Por serem unidades abstratas, os fonemas representam *classes de sons.* Isso significa que dois sons diferentes entre si, mas que não representem papel distintivo dentro da língua, pertencem a uma mesma classe. Considere por exemplo a palavra ***porta.*** Dependendo da região do Brasil em que nos encontremos, veremos o ***"r"*** de ***porta*** ser pronunciado de várias maneiras diferentes. Cada uma dessas variantes do fonema ***/r/*** constitui um *alofone* de ***/r/.*** Diferentemente dos fonemas, os alofones não possuem papel distintivo na língua (no exemplo acima, não importa qual dos alofones de ***/r/*** seja utilizado, a palavra ***porta*** terá sempre o mesmo significado). Se um fonema representa uma classe de sons, um alofone representa, por sua vez, uma sub-classe da classe fonema.

Uma outra forma de fazer essa diferenciação é dizer que um alofone é um fonema para o qual foram especificadas características além daquelas necessárias para diferenciá-lo de um outro fonema, ou seja, um alofone é um fonema *superespecificado.* A notação normalmente utilizada representa os fonemas "entre barras" ( // ), enquanto que os alofones são representados "entre colchetes" ( [] ).

A realização física de um fonema, por sua vez, é denominada de *fone*. O fone não é uma unidade abstrata como o fonema, mas sim uma unidade física real (trecho de sinal acústico). A cada fonema corresponde um número infinito de fones, todos eles com um grau de semelhança suficiente que permita classificá-los como sendo realizações acústicas pertencentes à mesma classe.

### 3.2.2 Fonética

A Fonética é o ramo da Lingüística que estuda os sons constituintes da fala de acordo com suas características acústicas e articulatórias. Nesse caso, as unidades da fala são estudadas como sinais sonoros e não como entidades abstratas, como ocorre na Fonologia.

#### *3.2.2.1. Fonética Articulatória*

A fonética articulatória procura descrever os diversos sons da língua de acordo com a dinâmica (posição e movimentação) dos articuladores que constituem o trato vocal humano (lábios, língua, mandíbula, etc.). Podemos levar em conta tanto o *modo de articulação* como o *ponto de articulação* na classificação dos sons constituintes da fala.

**Modo de articulação:** Ao classificar os sons da língua segundo o modo de articulação, estamos levando em conta o caminho percorrido pelo fluxo de ar ao longo do trato vocal, bem como o grau de obstrução por ele encontrado durante esse trajeto. De acordo com esse critério, podemos dividir os sons da fala em:

- vogais e ditongos: na produção das vogais, a passagem do ar pelo trato vogal é livre, e a configuração dos articuladores é estacionária. Ex.: */a/*, */e/*, */i/*, */o/*, */u/*. Já os ditongos são encontros vocálicos. Eles são semelhantes às vogais, no entanto diferem destas no que diz respeito à dinâmica dos articuladores do trato vocal: se no caso das vogais os articuladores se encontram em uma configuração

estacionária, os ditongos apresentam uma característica dinâmica, pois a posição dos articuladores varia à medida em que ocorre a transição de uma vogal à outra. Ex: */aɪ/, /eɪ/*.

- nasais: na produção das nasais a passagem de ar pelo trato também é livre; nesse caso o fluxo de ar passa também pela cavidade nasal, o que ocasiona um acoplamento desta com a cavidade oral. Ex.: /m/, /n/.

- líquidas: são semelhantes às vogais, com uma obstrução no eixo central do trato vocal.. Ex.: /l/, /r/. Algumas, como o /l/, deixam escapar o ar lateralmente: são as laterais. Outras, como o /r/, obstruem essa passagem rapidamente ou intermitentemente: são as vibrantes.

- fricativas: nas consoantes fricativas ocorre uma obstrução parcial à passagem do fluxo de ar em algum ponto do trato vocal, o que ocasiona a geração de um ruído de turbulência. As fricativas podem ser sonoras (ex.: /v/, /z/, /ʒ/) ou não sonoras (ex. /f/, /s/, /ʃ/), dependendo da ocorrência ou não de vibração das pregas vocais.

- plosivas (oclusivas): no processo de produção das consoantes plosivas ocorre uma obstrução total à passagem do fluxo de ar, seguida de uma liberação abrupta (burst) desse fluxo retido com ruído. As plosivas também podem ser classificadas como sonoras (ex.: /b/, /d/, /g/) ou não sonoras (ex.: /p/, /t/, /k/).

**Ponto de articulação:** Os sons da fala também podem ser classificados de acordo com o ponto de articulação, ou seja, o ponto do trato vocal onde ocorre a obstrução máxima à passagem do fluxo de ar. Esse tipo de classificação é válido para as consoantes, uma vez que, no caso das vogais, não se pode falar exatamente em "obstrução" à passagem do fluxo.

De acordo com o ponto de articulação as consoantes podem se classificadas como:

- bilabiais: as consoantes bilabiais são formadas a partir da constrição dos lábios superior e inferior. Ex.: /p/, /b/.

- labiodentais: são produzidas com a constrição entre o lábio inferior e os dentes superiores. Ex.: /f/, /v/.

- dentais: nesse caso, a ponta da língua toca os incisivos superiores. Ex.: /t/, /d/.

- alveolares: a constrição ocorre entre a ponta da língua e os alvéolos (porção do trato situada entre o palato e os incisivos superiores). Ex.: /l/, /ɾ/.

- palatais: são produzidas com o dorso da língua próximo ao palato duro. Ex.: /ʎ/, /ɲ/.

- velares: nesse caso, a constrição ocorre entre o dorso da língua e o palato mole. Ex.: /k/, /g/.

No caso das vogais, muito embora não se possa falar exatamente em obstrução à passagem do fluxo de ar, podemos nos referir à localização do ponto de constrição máxima. Nesse caso, as vogais podem ser divididas em *anteriores*, *centrais* e *posteriores.* Elas podem ainda ser classificadas de acordo com o formato dos lábios (arredondados ou não arredondados), bem como pelo grau de abertura dos maxilares (aberta ou fechada).

#### *3.2.2.2. Fonética Acústica*

Procura analisar os sons da fala como sinais acústicos, por isso leva em conta suas características espectrais e da forma de onda. De acordo com a Fonética Acústica, podemos classificar os sons da fala em:

- vogais: as vogais apresentam forma de onda com características periódicas. Cada vogal é caracterizada espectralmente pelo seu padrão formântico. São sinais de alta energia, concentrada principalmente nas regiões de baixa freqüência.

- ditongos: são semelhantes às vogais, mas apresentam um padrão formântico dinâmico.

- fricativas: são sinais de forma de onda aperiódica e com baixa intensidade. Espectralmente, a energia se concentra nas altas freqüências. As fricativas sonoras apresentam alguma periodicidade: em seus espectrogramas podemos identificar uma quantidade considerável de energia nas regiões de baixas freqüências (barra de vozeamento), devida à vibração das pregas vocais.

- plosivas: são caracterizadas por um silêncio seguido de uma explosão (período curto de alta energia), devida à liberação do ar retido durante a obstrução do trato vocal. No caso das plosivas sonoras também ocorre uma barra de vozeamento devida à vibração das pregas vocais.

- líquidas: são consoantes sonoras semelhantes às vogais, pois possuem padrão formântico definido. No entanto, são sinais mais curtos e de menor energia. No caso das laterais, ocorre a presença de zeros, devido à existência de uma cavidade sobre a língua.

- nasais: são semelhantes às vogais, mas apresentam menos energia, devido à atenuação introduzida pela presença dos zeros nasais, e também pelo fato do ar escapar somente pela cavidade nasal.

# 4 Síntese de fala a partir de texto

## 4.1 Dificuldades

A meta principal que um sistema de conversão texto-fala se propõe a alcançar é a de produzir um sinal de fala artificial a partir de um texto de entrada qualquer escrito numa determinada língua.

Em princípio, um sistema desse tipo deve ser o mais abrangente possível, o que significa dizer que não deve haver restrições quanto ao tipo de texto a ser sintetizado, desde que, obviamente, esse texto esteja escrito na língua com a qual o sistema se propõe a trabalhar (vale aqui lembrar, no entanto, que os sistema também deve ser capaz de lidar com as palavras estrangeiras incorporadas à lpingua pelo uso).

Para que pudesse funcionar de maneira *ideal*, um sistema desse tipo deveria ser capaz de realizar, de maneira automática, um processo similar ao que ocorre durante a leitura oral humana. Trata-se, entretanto, de uma tarefa bastante complexa. Existem vários fatores que impedem que esta seja simulada exatamente da maneira como ocorre no processo de produção da fala natural.

"A tarefa de leitura não se limita apenas à conversão de cada palavra na sua representação fonológica, mas envolve toda a competência lingüística do leitor. Em conseqüência disso, um texto pode ter uma diversidade de enunciados conforme o seu contexto, o seu leitor ou o efeito pretendido" [50]. Cada indivíduo percorre um caminho único no processo de transformação da mensagem textual em mensagem falada. Obtemos a representação fonológica de uma palavra não através da aplicação de regras de transcrição ortográfico-fonética, mas sim por meio de uma associação direta à representação desta palavra

que temos armazenada em nosso léxico. O conteúdo desse léxico é individual, fruto da competência lingüística e da experiência prévia do indivíduo com a língua.

Muitas vezes nos utilizamos de mecanismos perceptivos que nos permitem “ajustar” a nossa fala à medida em que esta vai sendo gerada. Isto significa que somos capazes de controlar os elementos de nosso aparelho fonador a partir da própria percepção do sinal de fala que estamos produzindo, num mecanismo típico de auto-ajuste. Além disso, o leitor é capaz de utilizar a informação semântica do texto, ou seja, o seu significado, para construir o sinal de fala da maneira que julgar mais adequada. A falta de controle de um sistema automático de síntese de fala sobre fatores como os acima descritos é que torna difícil a simulação exata do processo de leitura oral utilizada pelo ser humano na produção da fala natural.

Muito embora a simulação exata do processo de leitura não possa ser realizado de maneira inteiramente automática, não se pode abrir mão de que um sistema de conversão texto-fala incorpore modelos lingüísticos realistas, pois é exatamente a presença desses modelos que irá garantir um bom desempenho do sistema de conversão.

## 4.2 Aspectos principais da conversão texto-fala

De maneira geral, a tarefa da síntese de fala a partir de texto pode ser dividida em duas etapas distintas realizadas em seqüência: a primeira etapa, correspondente à *análise* do texto, consiste em obter a representação fonológica da mensagem a partir de sua forma ortográfica; a etapa de *síntese*, por sua vez, é responsável pela geração do sinal acústico associado à representação fonológica obtida na etapa anterior.

Podemos detalhar um pouco mais a estrutura acima descrita a fim de melhor compreender a seqüência de passos que um sistema de conversão deve realizar desde a interpretação do texto de entrada até a geração do sinal de fala. Observe o diagrama de blocos a seguir:

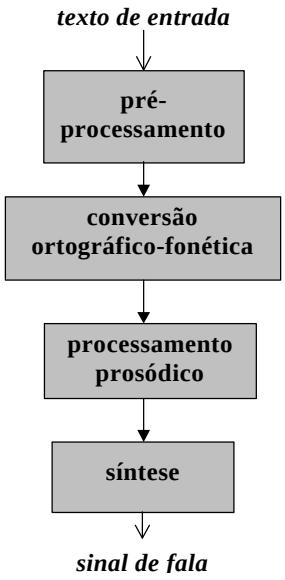



**Figura 4-1 Estrutura geral de um sistema de conversão texto-fala**

Os dois primeiros blocos (pré-processamento e transcrição ortográfico-fonética) fazem parte da etapa de *processamento lingüístico,* o qual corresponde à fase de análise que mencionamos anteriormente. Já a etapa seguinte corresponde aos dois últimos blocos do diagrama (processamento prosódico e síntese do sinal).

### 4.2.1 Pré-processamento

O *pré-processamento* é a primeira etapa  a ser realizada no processo de conversão texto-fala. Sua função principal é a de efetuar a *normalização* do texto de entrada, a fim de que este possa ser manipulado adequadamente pelos blocos seguintes do conversor. O processo de normalização consiste em substituir certos elementos do texto por seus equivalentes “por extenso”, de forma que a entrada do bloco subseqüente (no caso o bloco de

transcrição ortográfico-fonética) seja uma seqüência de sílabas da língua. Alguns elementos tipicamente manipulados na etapa de pré-processamento são as siglas, abreviaturas, dígitos e outros símbolos especiais (“+”, “%”, “@”, etc.). Muito embora a tarefa de normalização aparente ser bastante simples, ela esconde certas particularidades que, conforme veremos adiante, tornam o tratamento de certos elementos do texto bastante complexo.

### 4.2.2 Transcrição ortográfico-fonética

Após o pré-processamento, o texto deve passar pela etapa de *transcrição ortográfico-fonética*. Essa etapa consiste em encontrar a seqüência correta de fonemas que representa cada uma da palavras contidas no texto. Por ser um dos pontos chave do processo de conversão texto-fala, o módulo de transcrição exige a presença de modelos lingüísticos adequados que garantam que a transcrição seja feita de forma correta.

No entanto, por mais que os modelos sejam consistentes, o mapeamento de uma seqüência de letras em seqüência de fones é extremamente complexo, e nem sempre é possível realizar a transformação adequada. Certas regras de transcrição são bastante intrincadas: veja por exemplo o caso da letra *“x”* na língua portuguesa: as palavras *xuxu, exame, tórax* e *próximo* contêm a letra *“x”*, mas em cada caso ele corresponde a um (ou mais) fonemas diferentes. Notório também é o caso das letras *“e”* e *“o”*, que podem ter pronúncia fechada (*pêra, boca)* ou aberta (*bela, bola*), ou podem inclusive ser pronunciadas como “i” e “u”, respectivamente (*vale, pato*).

Certos casos são ainda mais complicados de serem resolvidos. Considere por exemplo as seguintes frases: *“O piloto morreu”* e *“Eu piloto bem”*. Para saber que na primeira sentença o substantivo deve ser pronunciado como “pilôto” e que na segunda sentença o verbo deve ser pronunciado como “piló to”, é necessário realizar a análise morfológica do texto, pois só assim pode-se identificar que a palavra *piloto* corresponde a um substantivo na primeira

sentença  e a um verbo na segunda. Por fim, há casos em que a ambigüidade não pode ser resolvida nem mesmo com a análise gramatical. Na sentença *"A sede da torcida é grande"*, qual é a pronúncia correta da palavra sede (*sêde ou séde)*? Somente a partir da análise do contexto podemos depreender a forma adequada, pois nos dois casos a palavra *sede* é um substantivo.

### 4.2.3 Processamento prosódico

Terminada a fase de análise do texto, tem-se a seqüência de fonemas correspondente à mensagem a ser sintetizada. Para efetuar a síntese do sinal, no entanto, essa informação não é suficiente; é preciso ainda calcular os parâmetros prosódicos correspondentes a cada um dos segmentos fonéticos. Esse é o papel do módulo de *processamento prosódico:* calcular padrões de intensidade, duração e F0 para cada segmento fonético da sentença. A prosódia representa um papel importante no que diz respeito à naturalidade do sinal de voz, pois carrega informação acerca do acento lexical das palavras, além de garantir ritmo e entonação adequados à sentença.

O cálculo desses parâmetros prosódicos também não é trivial, pelo simples fato de que não existe uma seqüência única de parâmetros prosódicos que possa ser associada a uma dada sentença. Obviamente existem características que constituem padrões da língua, como a presença de pausas em fronteiras prosódicas, curvas de entonação típicas para sentenças declarativas, imperativas e interrogativas, etc. Mesmo assim, a prosódia como um todo não possui uma estrutura fixa; cabe ao módulo de análise prosódica, portanto, encontrar valores para os parâmetros prosódicos da sentença que a tornem o mais próxima possível de um enunciado de fala natural.

Apesar de não estar diretamente representado no diagrama de blocos anterior, há ainda um elemento que deve estar presente no sistema de conversão texto-fala atuando em conjunto

com os módulos de transcrição fonética e de processamento prosódico. Trata-se do módulo de *análise gramatical* (denominado *parser*). O papel principal do *parser* é realizar a análise morfo-sintática do texto, fornecendo elementos importantes para os módulos de processamento prosódico e de transcrição fonética. É através da atuação do *parser* que se pode resolver problemas como o da palavra “piloto” descrito anteriormente, pois a transcrição correta só pode ser determinada a partir da identificação da classe gramatical da palavra. No caso do módulo de processamento prosódico, o *parser* também representa um papel importante pois, ao fazer a análise da estrutura sintática da sentença, permite a determinação, por exemplo, das fronteiras sintáticas onde podem ser inseridas as pausas.

### 4.2.4 Síntese do sinal

A última etapa do processo de conversão texto-fala é a síntese do sinal propriamente dita. O papel do módulo de síntese consiste em obter o sinal acústico a partir da representação fonético-prosódica calculada nas etapas anteriores. Existem várias estratégias utilizadas para a obtenção do sinal de fala, dentre as quais podemos destacar a síntese por regras, a síntese concatenativa e a síntese articulatória.

Ao longo dos próximos capítulos deste trabalho estaremos tratando com detalhes cada um das etapas acima descritas.

# 5 Processamento lingüístico

## 5.1 Introdução

Conforme vimos no capítulo anterior, o processo de síntese de fala a partir de texto pode ser dividido em etapas sucessivas. Neste capítulo estaremos discutindo a primeira dessas etapas, que aqui denominaremos de *etapa de processamento lingüístico.*

O objetivo principal da etapa de processamento lingüístico é obter a representação fonológica do texto de entrada, ou seja, transformar a mensagem textual em uma representação simbólica que indique a seqüência de unidades básicas de fala correspondentes à mensagem a ser sintetizada. Veremos neste capítulo que esse procedimento pode ser dividido em duas etapas distintas: o *pré-processamento* do texto e a *transcrição ortográfico-fonética.* Estaremos estudando, a seguir, cada uma dessas etapas.

## 5.2 Pré-processamento

Um sistema de síntese de fala a partir de texto pode ter aplicações de natureza bastante diversa, por isso é importante que o texto de entrada sobre o qual ele opera seja o mais genérico possível. Esse caráter genérico do texto de entrada pode trazer algumas dificuldades para a etapa de transcrição ortográfico-fonética, pois um texto, em geral, é constituído por muito mais do que uma simples seqüência de palavras.

Para que o texto de entrada possa ser manipulado adequadamente pelo módulo de transcrição ortográfico-fonética é necessário que ele sofra um processo de *normalização.* Essa é a função do pré-processamento.

O pré-processamento é, portanto, uma etapa preliminar à etapa de processamento lingüístico propriamente dita. Sua função é a de transformar um texto irrestrito em uma seqüência de palavras e de sinais de pontuação. Para chegar a isso ele deve substituir certos elementos não-lexicais do texto por seus equivalentes "por extenso", avaliar elementos não pronunciados (que podem ser úteis para outros módulos, como o de processamento prosódico, mas não são tratados durante a etapa de transcrição fonética) e expandir outros elementos que aparecem no texto sob forma sintética.

Numa primeira análise, o pré-processamento pode parecer uma tarefa trivial, pois trata-se simplesmente de substituir, suprimir e expandir símbolos. No entanto, essa tarefa esconde particularidades que a tornam extremamente complexa. Muitas das operações a serem efetuadas são extremamente difíceis de se determinar: alguns resultados são dependentes do texto com o qual estamos trabalhando (livro, jornal, texto científico, texto literário, etc.), e em muitos casos não é nem mesmo possível determinar a transformação mais adequada dentre as diversas opções possíveis.

A seguir são descritos alguns dos elementos passíveis de tratamento durante a etapa de normalização do texto:

**Números:**

Os números são elementos cuja ocorrência é bastante comum dentro dos mais diversos tipos de texto. A tarefa do pré-processamento consiste em substituir as ocorrências de números ao longo do texto de entrada por sua forma extensa, ou seja, sua transcrição ortográfica. Considere por exemplo a seguinte sentença:

*"No dia 4 de dezembro de 1.999 eu farei 25 anos".*

Após a normalização essa sentença deve ser transformada em:

*"No dia quatro de dezembro de mil novecentos e noventa e nove eu farei vinte e cinco anos".*

O pré-processamento de números é bastante complexo pois, dependendo do tipo de informação por eles representada, os números são lidos de maneiras diferentes. A maneira mais usual de leitura é a forma cardinal, na qual os números representam valores de quantidade. No exemplo anterior, os números foram expandidos para a forma cardinal. Em muitos casos, no entanto, essa não é a forma correta de leitura. Há por exemplo o caso dos números ordinais, que são utilizados quando a informação representada pelo número não é de quantidade, mas sim de ordem. Considere o exemplo abaixo:

*"Ela está na 3ª série do 2° grau".*

Após o pré-processamento essa sentença seria transformada em:

*"Ela está na terceira série do segundo grau".*

Números de telefone, por sua vez, são lidos normalmente dígito a dígito, ou às vezes em blocos de dezenas.

Ex.: *269-8550 -> "Dois - seis - nove - oito - cinco - cinco - zero", "dois - meia - nove - oitenta e cinco - cinqüenta", etc.*

Horas e valores em dinheiro também possuem maneiras bem particulares de serem transcritos. Considere os exemplos abaixo:

*"11:45h" -> "onze horas e quarenta e cinco minutos".*
*"R$49,50" -> "quarenta e nove reais e cinqüenta centavos".*

Há ainda alguns casos extremamente difíceis de serem resolvidos. Considere a expressão a seguir:

*"23/12"*

Uma possibilidade é interpretar essa expressão como sendo uma fração; nesse caso ela pode ser transcrita como *"vinte e três sobre doze"*, *"vinte e três dividido por doze"* ou *"vinte e três doze-avos"*. Mas podemos interpretar também que tal expressão se refere a uma data, e nesse caso devemos transcrevê-la como *"vinte e três do doze"* ou *"vinte e três de dezembro"*. A forma correta de transcrição somente pode ser determinada a partir do contexto.

Alguns números carregam também informação de gênero (masculino/feminino). Em uma expressão do tipo *"2.000 pessoas"* a forma correta da transcrição ortográfica é *"duas mil"* e não *"dois mil"*. Nesse caso, o pré-processador precisa identificar que o substantivo *"pessoas"* pertence ao gênero feminino. Este é um exemplo bem ilustrativo que demonstra o grau da dificuldade envolvida no problema do pré-processamento de números.

**Abreviaturas:**

As abreviaturas são elementos do texto que também devem ser tratados durante a etapa de pré-processamento. Muito embora sejam representadas por meio de uma seqüência ortográfica, as abreviaturas não são nunca lidas da maneira como estão escritas, e por esse motivo devem ser expandidas.

Em geral, o tratamento das abreviaturas é mais simples do que o tratamento dos números, pois normalmente existe uma correspondência biunívoca entre a abreviatura e sua expansão ortográfica. Veja os exemplos a seguir:

*Av. -> avenida*
*etc. -> etcetera*
*Sr. -> senhor*
*Sra. -> senhora*
*cm -> centímetro(s)*
*$km^2$ -> quilômetro(s) quadrado(s)*

Existem casos, no entanto, em que a expansão correta da abreviatura depende do contexto em que ela está inserida, o que dificulta bastante a tarefa de pré-processamento. Considere os exemplos a seguir:

*"Cap. 1: Introdução às técnicas de animação do programa do Bozo".*
*"O Cap. Amâncio Pinto foi condecorado com uma medalha de honra ao mérito".*

Na primeira sentença a abreviatura *"Cap."* deve ser expandida como *"capítulo"*, e na segunda como *"capitão"*.

Normalmente as abreviaturas são terminadas por ponto ("."), o que facilita a sua identificação. Podemos perceber, no entanto, a partir da análise de expressões como *"18h 45min"* ou *"1.500 m"*, que essa regra também tem suas exceções, e que certas abreviaturas não são necessariamente terminadas por ponto.

Uma dificuldade adicional no tratamento das abreviaturas é que algumas delas carregam consigo informação de número (singular/plural). Em expressões como *"1 m"* e *"2 m"*, por exemplo, muito embora as abreviaturas utilizadas sejam as mesmas, elas devem ser expandidas de forma diferente (*"metro"* e *"metros"*, respectivamente). Para fazer a transcrição correta o pré-processador deve ser capaz de identificar o valor expresso pelo elemento quantificador que, no texto, antecede a abreviatura.

**Siglas**

As siglas normalmente aparecem no texto como uma seqüência de letras maiúsculas, separadas ou não por pontos (*CIC, R.G., PMDB,* etc.).A principal dificuldade encontrada no tratamento das siglas é saber se elas devem ser lidas tal qual estão escritas ou se devem ser soletradas. As siglas compostas apenas por consoantes costumam ser soletradas, conforme mostra o exemplo a seguir:

*PSDB -> pê - ésse - dê - bê*
*FHC -> éfe - agá - cê*

As siglas cujas letras são separadas por pontos também são normalmente soletradas. Já quando a sigla é composta por consoante e vogais não separadas por pontos, normalmente não existe regra fixa. Há casos em que a sigla deve ser soletrada (*SOS, RA, TRE,* etc.). Outras, por sua vez, não podem jamais ser soletradas, mas devem ser lidas como uma palavra comum (*LAFAPE, USP, PUCC, FAPESP,* etc.).

Há ainda siglas que apresentam um comportamento misto (*FFESP*), ou cuja forma de transcrição é totalmente anômala (*IEEE* – pronuncia-se *"I três E").*

Nem todas as seqüências de letras maiúsculas correspondem a siglas. Considere por exemplo a expressão *"MW"*. Não se trata de uma sigla, mas sim de uma abreviatura, que deve ser expandida como *"megawatt"*.

**Pontuação:**

O pré-processador deve ser capaz de identificar o significado correto dos sinais de pontuação contidos no texto. Normalmente eles são mantidos, pois determinam fronteiras prosódicas ao longo da sentença e são tratados em etapas posteriores do processo de conversão texto-fala. Em alguns casos, no entanto, os sinais de pontuação são tratados já na etapa de pré-processamento.

O ponto final ("."), por exemplo, normalmente funciona como indicador de final de sentença, mas pode ser também utilizado nas abreviaturas e nas siglas. É o mesmo caso da vírgula, que pode ser utilizada na composição de números decimais, e nesse caso deve ser transcrita de forma explícita, como no exemplo a seguir:

*103,7 -> cento e três vírgula sete.*

Um último exemplo é o caso dos dois pontos (":"), que pode ser utilizado nas expressões de hora (*"11:30h"*, por exemplo).

**Símbolos especiais:**

Os textos em geral também são compostos por diversos símbolos não-alfabéticos que devem ser tratados durante a etapa de pré-processamento. Considere os exemplos a seguir:

*@ -> arroba*
*+ -> mais*
*% -> porcento*

Alguns símbolos podem ter mais de uma transcrição possível; a forma correta deve ser determinada a partir do contexto. Os exemplos a seguir ilustram essa ambigüidade:

*X -> vezes / versus.*
*" -> polegadas / segundos / abre aspas / fecha aspas.*

Os símbolos não-alfabéticos também podem carregar consigo informações de número (singular/plural), como no caso das expressões *R$1,00* ("*um real*") e *R$2,00* ("*dois reais*").

O pré-processamento deve, por fim, ser capaz de lidar com os elementos de formatação do texto (caracteres em negrito, itálico, sublinhado, barras de separação de seções, tabulações, etc.), bem como elementos não pronunciáveis do texto, como gráficos e ilustrações, de forma que o texto normalizado seja composto única e exclusivamente por palavras e sinais de pontuação.

## 5.3 Transcrição ortográfico fonética

A etapa seguinte ao pré-processamento do texto é a transcrição ortográfico-fonética, que consiste em fazer a transformação da seqüência ortográfica em uma cadeia de símbolos que represente a seqüência de sons que compõe cada uma das palavras do texto. Além de determinar a seqüência fonética, é preciso também fazer a identificação da sílaba tônica de cada palavra, pois essa informação será utilizada mais tarde durante a etapa de processamento prosódico.

A tarefa de transcrição fonética não é trivial, pois o texto normalizado que dá entrada no módulo de transcrição é uma seqüência discreta de símbolos ortográficos, ao passo que o sinal de fala correspondente varia de forma contínua ao longo do tempo. A transcrição fonética, apesar de estar mais próxima da realização oral do que o texto ortográfico, também é uma forma de representação com característica discreta.

O resultado da transcrição fonética é altamente dependente da língua com a qual se está trabalhando, pois o conjunto de sons existentes, bem como o mapeamento de letras em fonemas, varia de uma língua para outra. O grau de dificuldade enfrentado na tarefa de transcrição também varia de acordo com a língua envolvida. Certas línguas, como o italiano, o russo e o espanhol, possuem uma ortografia bastante fonêmica, ou seja, sua forma escrita é bastante próxima à realização oral, o que facilita bastante o estabelecimento de regras de transcrição. Já no caso de línguas como o inglês e o francês não há muita regularidade no mapeamento de seqüências de letras em seqüências de fonemas.

O português se situa entre as línguas de ortografia razoavelmente fonêmica; mesmo assim, não podemos dizer que a transcrição de um texto em português seja uma tarefa simples. O sistema ortográfico e o sistema fonético não são equivalentes, pois nem sempre é verdadeira a afirmação de que cada grafema da seqüência ortográfica corresponda a um fonema. Há casos em que um mesmo fonema pode ser representado por grafemas diferentes (o fonema */s/*, por exemplo, pode ser representado pelos grafemas **“s”**, **“c”** e **“x”**, como nas palavras *“**s**ela”*,

*"**c**edo"* e *"pró**x**imo"*, respectivamente). O inverso também acontece: fonemas distintos podem ser representados pelo mesmo grafema (considere o exemplo do grafema **"g",** que pode representar o fonema */ʒ/*, como na palavra *"**g**ente"*, ou o fonema */g/*, como na palavra *"**g**ato"*). Há ainda casos em que um fonema simples pode ser representado por uma seqüência não unitária de grafemas, como em *"ca**rr**o"* (*/r/*), *"a**lh**o"* (*/ʎ/*) ou *"fi**ch**a"* (*/ʃ/*). Um único grafema também pode representar uma seqüência de fonemas, como a letra **"x"** da palavra *"tó**x**ico"* (*/k/+/s/*). Por fim há situações em que um dado grafema da seqüência ortográfica não é mapeado para nenhuma unidade fonológica, como por exemplo a letra **"h"** na palavra *"**h**omem"*.

A existência de um dicionário de pronúncias contendo a representação fonética e o padrão de acentuação de cada uma das palavras existentes na língua faria com que a tarefa de transcrição se transformasse num processo simples de substituição. Essa, no entanto, não é a forma de solução mais adequada. O empecilho principal à adoção desse tipo de estratégia é, sem dúvida, o tamanho do dicionário a ser empregado. Tal dicionário demandaria do sistema uma capacidade de armazenamento gigantesca, e a busca por uma palavra dentro do dicionário certamente comprometeria a velocidade de operação do sistema, especialmente no caso de sistemas que se propõem a trabalhar em tempo real.

Além disso, há situações em que a utilização de um dicionário como o acima descrito se mostra insuficiente para que se possa determinar a transcrição correta das palavras do texto. Certas palavras podem ter mais de uma transcrição possível, de acordo com o contexto em que estão inseridas (é muito comum no português, por exemplo, a ambigüidade entre verbo e substantivo ou adjetivo, que se diferenciam unicamente pela forma de pronúncia da vogal tônica, que normalmente é aberta no caso dos verbos e fechada nos substantivos ou adjetivos. Palavras como *molho, seco* e *piloto* ilustram esse fenômeno). Um dicionário de pronúncias também não daria conta de neologismos, novas siglas e palavras estrangeiras que são freqüentemente incorporados à língua, a não ser que o conteúdo do dicionário fosse constantemente atualizado.

Há certas estratégias que podem ser utilizadas visando a evitar a utilização de dicionários de pronúncia extensos. A aplicação de *regras de transcrição*, por exemplo, tem por objetivo tratar das correspondências regulares entre letras e sons. Obviamente a aplicação de regras não dá conta de todas as formas de transcrição, pois nem sempre tal mapeamento segue um padrão definido. No caso de línguas muito irregulares, como o inglês, é comum a utilização de um *dicionário de morfemas*. Um morfema pode ser definido como sendo a unidade mínima de uma palavra, dotada de significado (considere, a título de exemplo, a palavra *atípicos:* fazendo sua decomposição em unidades básicas, veremos que ela é formada pelo radical "*típico*", acompanhado pelo prefixo "*a*" e pela desinência de plural "*s*").

Um dicionário de morfemas contém a pronúncias dos componentes básicos das palavras. Do ponto de vista prático, ele apresenta a vantagem de ser muito mais compacto do que um dicionário de pronúncias completo (no caso do inglês, por exemplo, um dicionário com cerca de 12.000 morfemas pode dar conta da maior parte das palavras da língua, ao passo que um dicionário de pronúncias deveria ser composto por cerca de 100.000 palavras [36]). Para utilizar um dicionário de morfemas é necessário desenvolver regras de decomposição das palavras em seus componentes básicos, o que nem sempre é uma tarefa trivial [5]**.**

No caso da língua portuguesa, em que existe uma regularidade grande entre a representação ortográfica e a transcrição fonética, pode-se determinar a pronúncia correta da maioria das palavras através da aplicação pura e simples de regras de transcrição. A aplicação de tais regras parte do princípio de que uma seqüência de grafemas pertencente a uma dada palavra pode ser convertida em uma seqüência fonética a partir da análise do contexto dos grafemas que lhe são adjacentes. A transcrição do texto é feita, portanto, a partir aplicação de regras de produção do tipo:

***<contexto_esquerdo> contexto_de_análise <contexto_direito> Þ transcrição***

Considere a regra de produção a seguir:

***"c" <"e", "i"> Þ /s/***

Tal regra diz, simplesmente, que a consoante ***"c"***, quando sucedida pelas vogais ***"e"*** ou ***"i"***, deve ser transformada no fonema ***/s/*** (a sintaxe utilizada no exemplo acima não segue nenhuma gramática em particular, e foi utilizada apenas como forma de exemplo).

Há várias seqüências de grafemas no português que são facilmente transcritas a partir da aplicação de regras desse tipo. Em português brasileiro o tratamento das consoantes costuma ser bem mais simples do que o das vogais, exceção feita talvez à consoante ***"x",*** cujo tratamento demanda um número bastante grande de regras, as quais, mesmo assim, não dão conta da transcrição correta em todos os casos. Outra dificuldade típica do português é o tratamento das vogais ***"e"*** e ***"o"***, que podem ser convertidos para vogais abertas ou fechadas, dependendo do contexto. As regras necessárias para dar conta da transcrição correta nesses casos são complexas e em número elevado; além disso, como no caso da letra ***"x***", nem sempre são suficientes para efetuar a transcrição correta.

A determinação da sílaba tônica das palavras também é feita, normalmente, a partir da aplicação de um conjunto de regras. Felizmente a língua portuguesa apresenta algumas características que facilitam a identificação do acento primário das palavras. Primeiramente, muitas das palavras do português são dotadas de acento gráfico (como, por exemplo, todas as proparoxítonas), e nesse caso a vogal tônica será sempre a vogal acentuada, o que torna o processo de identificação trivial. A maioria das palavras não acentuadas do português são paroxítonas. Sendo assim, a dificuldade maior na determinação da sílaba tônica das palavras em português diz respeito à identificação das palavras oxítonas não acentuadas; nesses casos faz-se necessária a existência de regras mais elaboradas.

Como já vimos anteriormente, a aplicação de regras de produção não é suficiente para determinar a pronúncia e a acentuação corretas de todas as palavras encontradas nos textos escritos. Muitas das palavras mais freqüentes da língua são palavras que não seguem os padrões regulares de pronúncia. Por esse motivo, a presença de um *dicionário de exceções* é

fundamental para minimizar a ocorrência de erros no processo de transcrição. O dicionário de exceções é um dicionário de pronúncias, mas é composto apenas pelas palavras para as quais as regras de transcrição ou acentuação falham. Portanto, tais regras devem ser aplicadas *após* a busca no dicionário de exceções, e apenas para aquelas palavras que não foram encontradas no dicionário.

Quanto maior for o dicionário de exceções maior será a taxa de acerto do módulo de transcrição. Essa relação, no entanto, não é linear, pois a partir de um certo limite o acréscimo de mais vocábulos ao dicionário acarreta um aumento cada vez menor na taxa de acerto. Pouco adianta acrescentar uma quantidade grande de palavras cuja ocorrência na língua é rara: o importante é que o dicionário contenha as palavras de uso mais freqüente, e que o seu tamanho não exija uma capacidade de armazenamento exagerada e nem comprometa a velocidade de operação global do sistema.

Como já vimos anteriormente, a determinação da pronúncia correta de algumas palavras não pode ser determinada somente a partir da aplicação de regras de transcrição ou do dicionário de exceções. Muitas palavras no português são homógrafas não homófonas, ou seja, são ortograficamente equivalentes mas têm pronúncias diferentes. A determinação da fonetização correta dessas palavras somente pode ser feita a partir de uma análise lingüística adequada. O caso dos verbos e substantivos (ou adjetivos) homógrafos (*molho, seco, piloto,* etc.), descrito anteriormente, encaixa-se nesse perfil. O módulo de transcrição deve receber dados de um *parser* morfo-sintático que, ao fazer a análise do texto, determinará a classe gramatical da palavra em questão. Somente a partir dessa informação será possível ao módulo de transcrição determinar a pronúncia correta.

Certos casos são ainda mais complicados, pois a análise morfo-sintática não é suficiente para determinar a transcrição fonética adequada. Considere por exemplo a palavra *"sede"*. Dependendo do significado ela pode ser pronunciada com a vogal tônica aberta (ex.: *"A sede da ONU fica em Nova York"*) ou fechada (ex.: *"Obedeça sua sede, beba Sprite"*). Em ambos os exemplos a classe gramatical da palavra é a mesma (no caso, um substantivo).

Somente uma análise semântica poderá depreender, a partir do contexto em que a palavra está inserida, o seu significado, e a partir dele obter a pronúncia correta. O atual estágio de desenvolvimento na área da conversão texto-fala permite-nos supor que ainda há um longo caminho a ser percorrido até que seja possível efetuar esse tipo de análise no texto de entrada com resultados satisfatórios.

Palavras derivadas por composição ou afixação também costumam fugir às regras normais de transcrição. Considere por exemplo palavras como *telecomunicações* e *socioeconômico.* Podemos observar que a primeira vogal é aberta, diferentemente do que se esperaria normalmente. Um algoritmo de identificação de palavras compostas evitaria que elas tivessem que, necessariamente, fazer parte do dicionário de exceções. Mesmo assim, há palavras compostas que fogem ao exemplo acima descrito pois tiveram sua pronúncia modificada pelo uso, como *televisão, telefone* e *fotografia*.

A tarefa de conversão ortográfico-fonética de um texto não consiste apenas em fazer a transcrição de cada uma de suas palavras. As palavras podem se coarticular entre si, por isso é necessário efetuar uma *análise pós-lexical*. O objetivo dessa análise é ajustar o resultado da transcrição das palavras em isolado de forma a levar em conta as alterações produzidas pela coarticulação em fronteira de palavra.

Há muitos casos em que tais fenômenos ocorrem. O fone [***s***]em final de palavra, por exemplo, normalmente é uma fricativa não sonora. Entretanto, quando o primeiro fonema da palavra seguinte é uma vogal (ex.: *"muita**s** **e**moções*) ou uma consoante sonora (ex.: *vário**s** **d**ias*), o fone [***s***] também se sonoriza. Quando a última vogal de uma palavra é idêntica à primeira vogal da palavra seguinte (ex.: *"noss**a** **a**mizade"*), elas também podem se coarticular, transformando-se numa única vogal (fenômeno conhecido como *"sândhi externo"*). Esses são apenas alguns exemplos de situações em que tais fenômenos ocorrem; se tal análise não for efetuada durante a etapa de processamento lingüístico do texto, com certeza a qualidade do sinal de fala sintetizado estará comprometida.

Muito embora a utilização de regras de transcrição e dicionários de exceções seja predominante nos sistemas de síntese atuais, existem algumas outras técnicas que também se propõem a resolver o problema da transcrição fonética. Dentre elas podemos destacar as técnicas de *aprendizagem de padrões*, como por exemplo as redes neurais artificiais. Um exemplo da aplicação de redes neurais à tarefa de transcrição é apresentado por Sejnowski [60], que utiliza uma seqüência de caracteres como janela de entrada para a rede; a saída corresponde à transcrição fonética do grafema central da janela de entrada. Os resultados obtidos mostram que as taxas de acerto obtidas por meio da utilização de redes neurais são expressivas, muito embora sejam sempre menores àquelas obtidas a partir da aplicação de regras de transcrição, mesmo no caso de línguas fonemicamente irregulares como o inglês.

# 6 Processamento prosódico

## 6.1 Prosódia

Uma mensagem falada não pode ser analisada única e exclusivamente sob o ponto de vista dos segmentos fonéticos que a constituem. Num sistema de conversão texto-fala, a etapa de transcrição fonética trata tão somente da determinação da seqüência de sons que irá constituir o sinal de fala correspondente ao texto de entrada; no entanto, existem muitas outras características importantes da fala que estão acima do nível *segmental* tratado durante a etapa de transcrição fonética.

O termo *prosódia* diz respeito às características da fala que atuam a nível de sílabas, palavras, orações, sentenças ou mesmo parágrafos. O processamento prosódico é, portanto, um processamento de natureza predominantemente *suprassegmental* (o que não significa, no entanto, como veremos adiante, que ele não atue também sobre os segmentos).

A prosódia carrega informações adicionais àquelas expressas pela seqüência de segmentos fonéticos. O processamento prosódico é essencial para garantir a inteligibilidade do sinal de fala sintetizado e, principalmente, para assegurar a sua naturalidade. Por isso um sistema de síntese de fala a partir de texto não pode abrir mão de um tratamento prosódico adequado, caso queira produzir um sinal de fala com características minimamente semelhantes às da fala natural.

A prosódia possui inúmeras funções no processo de codificação da informação da mensagem falada. Em primeiro lugar, é através da prosódia que o falante confere estruturação oral à sentença, dividindo-a em blocos lógicos menores; dessa forma ela pode ser quebrada mentalmente pelo ouvinte, facilitando-se assim a sua compreensão. A prosódia funciona também como uma portadora da individualidade do falante, pois cada pessoa tem uma maneira particular de enunciar as sentenças. Muitas características do falante podem ser depreendidas a

partir da análise de seu parâmetros prosódicos (as mulheres, por exemplo, possuem um valor de F0 intrinsecamente mais alto que o dos homens).

Aspectos da personalidade do falante são também expressos pelos parâmetros prosódicos (arrogância, humildade, timidez, por exemplo, podem ser expressos através de diferentes estilos de elocução), bem como suas emoções (alegria, tristeza, surpresa, etc.) e atitudes em relação a si mesmo ou a outrem (ironia, seriedade ou graça).

Outros tipos de informação, essenciais à transmissão correta da mensagem falada, manifestam-se através dos parâmetros prosódicos ao longo da sentença. A estrutura sintática da sentença, por exemplo, pode ser depreendida em parte a partir de sua divisão em constituintes prosódicos, conforme veremos mais adiante. O acento lexical (relacionado à palavra) e o acento frasal também são expressos por meio de parâmetros prosódicos, normalmente através de contrastes de F0, duração e amplitude. Informações semânticas também se refletem na prosódia (através da resolução de ambigüidades, por exemplo, via foco ou acento frasal).

## 6.2 Parâmetros prosódicos

Os parâmetros prosódicos são as características do sinal de fala cuja manipulação adequada irá refletir a estrutura prosódica do enunciado. Estes parâmetros estão associados a cada um dos segmentos fonéticos da sentença. Existem três parâmetros prosódicos principais (duração, freqüência fundamental e intensidade), os quais serão discutidos a seguir.

### 6.2.1 Duração

O parâmetro duração está associada ao intervalo de tempo entre o início e o final de um segmento fonético. Os segmentos fonéticos possuem durações médias da ordem de dezenas a centenas de milissegundos , mas obviamente o valor médio e a dispersão são características

individuais de cada falante. A duração é um parâmetro prosódico importante, pois varia de acordo com a taxa de elocução do enunciado e reflete também o contexto prosódico em que o segmento fonético está inserido (ambientes prosódicos fortes normalmente ocasionam alongamento dos segmentos fonéticos). Segmentos localizados em fronteiras de constituintes prosódicos também costumam ter sua duração aumentada.

### 6.2.2 Freqüência fundamental (F0)

A freqüência fundamental (F0) de um sinal de fala é um valor instantâneo que está diretamente associado à taxa de vibração das pregas vocais, e que se manifesta através da periodicidade da forma de onda nos sinais sonoros. Obviamente, não faz sentindo falar em freqüência fundamental quando lidamos com segmentos de fala não sonoros, pois nesse caso não ocorre vibração das cordas vocais, e a forma de onda tem características aperiódicas.

O conceito de *pitch* está intimamente associado ao de freqüência fundamental e na literatura sobre síntese e reconhecimento de fala os dois termos costumam ser utilizados de forma equivalente. Na verdade, a freqüência fundamental é um valor numérico real, associado a cada instante do sinal de fala, e corresponde ao inverso do período do sinal sonoro. Normalmente a freqüência fundamental é medida em Hertz (Hz). O *pitch*, por sua vez, é um conceito meramente perceptual, e diz respeito à sensação de altura (grave/agudo): quanto maior for a freqüência fundamental, maior será o *pitch* ou, equivalentemente, mais agudo será o sinal. A relação entre a freqüência fundamental e a sensação de altura do sinal é quase logarítmica, e portanto não linear.

Além de ser uma característica individual do falante (certos falantes possuem registro mais grave, outros mais agudos), a freqüência fundamental constitui um dos parâmetros mais importantes, juntamente com a duração, a ser considerado durante a etapa de tratamento prosódico.

### 6.2.3 Intensidade

A intensidade está associada à amplitude da forma de onda (é proporcional ao quadrado da amplitude). É através do parâmetro de intensidade que podemos distinguir os sons fortes dos sons fracos (quando gritamos, por exemplo, estamos produzindo sinais mais intensos do que quando sussurramos).

Intuitivamente, poder-se-ia acreditar que a amplitude é um parâmetro prosódico tão importante quanto os demais; o que acontece na prática, porém, é que a intensidade do sinal tem uma função de contraste muito menos significativa do que os outros parâmetros prosódicos, como a duração e a freqüência fundamental. Durante algum tempo acreditou-se que os segmentos fonéticos acentuados se destacavam dos demais por meio de um padrão de energia mais alto; no entanto, o que se observa efetivamente é que são as variações de duração e de F0 que determinam, muito mais do que o aumento da energia, a localização do acento nas sentenças, mas a contribuição específica de cada parâmetro varia de língua para língua.

A maioria dos sistemas de síntese de fala a partir de texto não fazem o modelamento de energia durante a etapa de processamento prosódico, atendo-se, em contrapartida, ao tratamento dos padrões de duração e de freqüência fundamental. Acredita-se que, dada a importância relativamente menor da energia como parâmetro prosódico, o seu modelamento não traria ganhos significativos à qualidade do sinal de fala sintetizado.

Isso é verdade, mas apenas em termos. A queda de amplitude das vogais tônicas para as pós-tônicas, por exemplo, tem uma importância significativa para a acentuação lexical. Além disso, um estudo mais detalhado a respeito do fenômeno de ênfase talvez suscitasse uma nove análise a respeito da necessidade de modelar a intensidade como parâmetro prosódico.

## 6.3 Macroprosódia e microprosódia

Muito embora possamos dizer que a atuação da prosódia esteja acima do nível segmental, seria incorreto afirmar que a prosódia e os segmentos representam entidades inteiramente independentes. Em primeiro lugar, porque os parâmetros prosódicos estão diretamente associados aos segmentos da sentença (cada um deles possui um valor de duração, bem como padrões de F0 e amplitude). Além disso, as características articulatórias de cada segmento impõem limites às variações prosódicas que estes podem sofrer. O fenômeno de coarticulação também determina que variações prosódicas em um dado segmento podem se estender ao segmento seguinte.

Tendo em vista as considerações acima desenvolvidas, percebe-se claramente que há dois níveis de atuação da prosódia. A *macroprosódia* influencia os parâmetros prosódicos de forma a estruturar a sentença a nível de sílabas, palavras, frases, etc.. A *microprosódia*, por sua vez, tem sua área de atuação restrita aos segmentos fonéticos adjacentes e procura garantir a continuidade desses segmentos tendo em vista seus limites articulatórios.

## 6.4 Acentuação e ritmo

Ao longo do enunciado existem certas sílabas que são mais salientes em relação às demais. Dizemos que tais sílabas são *acentuadas*, e a esse recurso de colocar a sílaba em proeminência damos o nome de *acentuação.*

Pode-se falar em dois graus de acentuação, dependendo do domínio em que os parâmetros atuam. O *acento lexical* é um componente inerente a cada um dos vocábulos da língua portuguesa: toda palavra de nossa língua possui uma sílaba tônica (guardada a exceção das palavras gramaticais átonas), e muitas vezes essa sílaba tônica é indicada por meio de um acento gráfico. As palavras do português podem ser classificadas, segundo a localização do

acento lexical, em: 1)oxítonas (acento na última sílaba); 2) paroxítonas (acento na penúltima sílaba); 3) proparoxítonas (acento na antepenúltima sílaba).

O *acento frasal*, diferentemente do lexical, não ocorre a nível de palavra, e sim de frase. O acento frasal é utilizado para salientar de forma culminativa certas palavras, de forma a facilitar a compreensão do enunciado por parte do ouvinte.

Além de revelar a intenção do falante, o acento frasal também revela a estrutura sintática da sentença, pois normalmente ele recai sobre os núcleos dos constituintes sintáticos.

Certas sílabas, muito embora não possam ser consideradas sílabas tônicas, também são mais proeminentes que as demais. Esse fenômeno, que pode ocorrer tanto a nível lexical quanto a nível frasal, é denominado de acento secundário. Considere por exemplo a palavra *"polivalente"***.** Trata-se de uma paroxítona, cuja sílaba tônica é *"len"***.** Não obstante, podemos identificar também um acento secundário na primeira sílaba (*"po"*). A presença de acentos primários e secundários ao longo dos enunciados confere à fala um padrão duracional o qual denominamos *ritmo*.

A primeira impressão do ouvinte é que as sílabas acentuadas são mais fortes que as demais. Poder-se-ia supor, portanto, que o fenômeno da acentuação estivesse diretamente relacionado ao parâmetro prosódico de intensidade. O que acontece na prática, no entanto, é que a existência de contrastes nos valores de duração e de *pitch* são mais importantes para a determinação do acento do que a intensidade em si.

## 6.5 Estrutura sintática e estrutura prosódica

Vimos que uma das funções da prosódia é conferir estruturação oral à sentença, de forma que ela possa ser quebrada mentalmente pelo ouvinte. Essa estruturação se manifesta através da divisão da sentença em *constituintes prosódicos.*

Um constituinte prosódico é um conjunto de palavras adjacentes que apresentam um certo grau de coesão dentro da sentença, sendo que o conjunto tem a propriedade de influenciar a evolução dos parâmetros prosódicos ao longo das palavras que o constituem.

A estrutura prosódica de uma sentença está intimamente relacionada com sua estrutura sintática. Toda e qualquer sentença pode ser analisada como uma estrutura hierárquica de constituintes (sintagmas nominais, verbais, preposicionais, etc.), os quais desempenham determinadas funções dentro da frase [51].

Considere, por exemplo, a sentença a seguir:

*"O gato comeu a sua língua."*

Essa é uma sentença bastante simples, na qual podemos identificar: 1) um sintagma nominal na função de sujeito ("*o gato*"); 2) um sintagma verbal na função de predicado ("*comeu a sua língua*"); e 3) um outro sintagma nominal ("*a sua língua*"), desempenhando função de complemento verbal.

Normalmente os constituintes prosódicos estão relacionados com a estrutura sintática da sentença; no entanto, essa relação nem sempre é direta. Cada sentença possui uma única estrutura sintática, ao passo que há várias estruturas prosódicas possíveis. Um dos fatores que determina o afastamento entre a estrutura sintática e a prosódica é o princípio da *isocronia* (ou *eurritmia*). Segundo esse princípio, os constituintes prosódicos tendem a ter durações próximas entre si, o que pode afetar o agrupamento das palavras: dessa forma, evita-se a decomposição da sentença em blocos muito curtos ou muito compridos, que dificultariam a tarefa de compreensão. Obviamente, esse princípio da isocronia não é absoluto, e pode ser quebrado se a estrutura sintática da sentença assim o exigir.

## 6.6 O *parser*

Como vimos anteriormente, a estrutura prosódica de uma sentença está intimamente associada com a sua estrutura sintática. A função do *parser* morfo-sintático dentro de um sistema de conversão texto-fala é justamente a de determinar uma estrutura sintática, mínima que seja, às sentenças do texto de entrada. A informação gerada pelo *parser* será utilizada pelo módulo prosódico (e também por outros módulos, como o de transcrição fonética), nas etapas posteriores do processo de conversão texto-fala.

Através da análise sintática efetuada pelo *parser* é possível fazer a decomposição da sentença em constituintes sintáticos e, a partir dessa decomposição, inferir a localização provável das fronteiras prosódicas ao longo do enunciado.

Outra função do *parser* é a resolução de ambigüidades, como aquelas que ocorrem entre verbo e substantivo e que já foram citadas no capítulo anterior, através de exemplos como os das palavras ***molho, seco e piloto.*** Esse tipo de resolução não está diretamente associado ao processamento prosódico, mas sim à transcrição fonética. Com isso vemos que, na verdade, o *parser* constitui um módulo independente dentro do sistema de síntese de fala a partir de texto, e que o resultado do seu processamento é utilizado em diferentes pontos ao longo do processo de conversão.

Para efetuar a análise morfo-sintática de maneira apropriada é necessário que o *parser* possua um conhecimento extenso acerca da língua com a qual opera. Normalmente o seu funcionamento é baseado na existência de um *léxico*. O léxico contém palavras com as respectivas classes gramaticais, e a partir desse léxico o algoritmo é capaz de fazer a classificação das palavras do texto. Sob o ponto de vista de suas classes, as palavras podem ser divididas entre palavras de *conteúdo* ou *lexicais* (como verbos, substantivos, adjetivos, etc.), que carregam significado próprio mesmo quando ocorrem isoladamente, e palavras *funcionais* ou *gramaticais*(como preposições, artigos, conjunções), que não possuem significado próprio e normalmente funcionam como elementos de ligação dos constituinte sintáticos. A partir da

identificação das classes de palavras, a tarefa do *parser* é efetuar a decomposição das sentenças em seus constituintes sintáticos. Normalmente as palavras de conteúdo atuam como núcleos de constituintes, de forma que o algoritmo procura agrupar palavras em torno dos núcleos a fim de gerar os constituintes sintáticos apropriados.

Existem alguns algoritmos eficientes para efetuar a tarefa de *parsing*, mas é comum que esses algoritmos produzam estruturas sintáticas incorretas. Por isso é usual que os sistemas de síntese de fala a partir de texto aceitem alguma forma de indicar a estrutura sintática das sentenças dentro do próprio texto de entrada, a partir de caracteres especiais.

A derivação da estrutura sintática das sentenças do texto de entrada não é suficiente para a tarefa de conversão texto-fala. É necessário também efetuar a determinação de informações semânticas (significado da palavras) e pragmáticas (intenção do falante), informações estas que são utilizadas em diversos pontos ao longo da tarefa de conversão texto-fala, entre elas a etapa de processamento prosódico. No entanto, são poucos os sistemas, hoje em dia, que efetuam alguma forma de análise semântico-pragmática no texto a ser sintetizado.

## 6.7 Geração automática da prosódia

### 6.7.1 Modelo de duração

Num sistema de síntese de fala a partir de texto, uma das funções do módulo de processamento prosódico é prover uma forma de tratamento capaz de determinar, de maneira *automática*, a duração de cada um dos segmentos e pausas que compõem o enunciado a ser sintetizado. Um *modelo de duração* adequado a tal tarefa deve ser capaz, portanto, de levar em conta os diversos fatores já aqui descritos que afetam a duração dos segmentos constituintes da sentença.

Não se trata de uma tarefa simples, pois não existe uma única solução possível (a prosódia é uma marca da individualidade do falante; por isso, uma mesma sentença pode ser lida com modos de elocução diferentes, mas igualmente aceitáveis). Portanto, a tarefa do módulo prosódico consiste em determinar valores de duração para os constituintes da sentença, de forma a aproximá-la o máximo possível dos padrões de fala natural.

Há várias estratégias que podem ser utilizadas para implementar um modelo de duração. Uma das abordagens bastante utilizadas é a de um modelo duracional baseado em *regras*. O modelo proposto por Klatt [4] para a língua inglesa segue esse princípio e será aqui apresentado como forma de ilustrar os princípios básicos por trás de tal abordagem.

O modelo desenvolvido por Klatt baseia-se nas seguintes suposições:

- cada segmento fonético possui uma duração intrínseca, que corresponde à média das durações que esse segmento pode assumir.
- os segmentos fonéticos podem ser alongados ou encurtados de acordo com o ambiente prosódico em que se encontram. Um conjunto de regras é responsável por determinar, de maneira independente umas das outras, o nível de alongamento ou encurtamento do segmento.
- cada segmento possui um valor de duração mínima a ele associado. Nenhum segmento pode ser encurtado aquém de seu valor mínimo de duração.

O modelo pode ser expresso pela fórmula a seguir:

$$D = D_{Min} + \prod_{j=1}^{N} (k_j)(D_I - D_{Min})$$

onde:

$D$ = duração calculada para o segmento;

$D_{Min}$ = duração mínima do segmento;

$D_I$ = duração intrínseca do segmento;

$k_j$ = fator de ajuste de duração associado à regra *j;*

$N$ = número de regras aplicáveis ao contexto.

Cada uma das regras que se aplica ao contexto fonético-prosódico do segmento cuja duração está sendo calculada contribui com um coeficiente de variação da duração intrínseca. Os diversos coeficientes aplicáveis são multiplicados, e o resultado dessa multiplicação constitui o coeficiente global de variação do segmento. Note que a nova duração nunca será menor que a duração mínima, pois $(D_I - D_{Min})$ é sempre positivo.

O resultado obtido com a aplicação do modelo depende do conjunto de regras que o constitui. Obviamente tais regras são altamente dependentes da língua com a qual se está trabalhando. As regras podem ter vários níveis de atuação, dependendo da extensão do contexto fonético de análise. As regras podem atuar a nível de fonema, sílaba, palavra, constituinte prosódico ou mesmo sentenças inteiras. *"Reduzir por um fator de 0,8 a duração de um segmento pertencente a uma sílaba pré-tônica"* ou *"o primeiro fonema de um constituinte prosódico deve ser aumentado de um fator de 1,3"* são exemplos típicos de regras de um modelo como o aqui exposto, atuando a nível de sílaba e de constituinte prosódico, respectivamente. Quanto mais completo for esse conjunto de regras, melhor será o resultado obtido por meio do modelo.

Existem outras abordagens que também podem ser utilizadas na elaboração de um modelo de duração. Podemos destacar, por exemplo, os modelos *estatísticos*. Tais modelos necessitam da existência de um extenso *corpus* de fala de um determinado locutor adequadamente etiquetado. A informação contida no *corpus* é utilizada pelo modelo para derivar as características prosódicas do locutor a ele associado. A partir dessa informação o modelo consegue calcular o padrão de duração das sentenças a serem sintetizadas. Redes

neurais [12] e árvores de classificação e regressão [58] são exemplos de abordagens baseadas em métodos estatísticos. Tais modelos podem produzir bons resultados, mas possuem a desvantagem de, eventualmente, gerar erros bastante grosseiros, especialmente ao encontrar contextos fonético-prosódicos mais raros, não contemplados no *corpus* de base. Esse fato advém da dificuldade de elaborar um *corpus* que constitua um espaço amostral completo dos fenômenos prosódicos que ocorrem na língua.

No caso da língua portuguesa, vale citar o modelo proposto por Barbosa [9] [10], também baseado em uma abordagem estatística. Tal modelo utiliza duas unidades rítmicas mínimas, a sílaba e o GIPC (grupo *inter-perceptual-center*,do início de uma vogal ao início da próxima). A geração automática de duração divide-se em duas etapas: primeiramente utiliza-se uma rede neural (perceptron multicamada) para calcular fatores de alongamento (*z-scores*) correspondentes a cada uma das unidades rítmicas do enunciado, para em seguida fazer-se a distribuição da duração das unidades entre os segmentos que as constituem.

### 6.7.2 Modelo entoacional

A segunda função importante do módulo de processamento prosódico é a de procurar determinar, de forma também automática, um contorno de freqüência fundamental apropriado a cada uma das sentenças do texto a ser sintetizado. Para isso, é necessária a existência de um *modelo entoacional* .

Muito do que foi discutido em relação ao modelo de duração vale também para o modelo entoacional. Não existe solução única para o problema da determinação do contorno de F0: deve-se procurar, portanto, um contorno que se aproxime o máximo possível dos padrões que ocorrem na fala natural.

Como nem sempre é possível extrair informações semântico-pragmáticas somente a partir da análise do texto escrito, opta-se, normalmente, por atribuir-se às sentenças sintetizadas um contorno entoacional *neutro* (por exemplo, uma linha de declinação simples).

Vale lembrar ainda uma diferença adicional em relação ao modelo de duração. Neste, o problema consiste em determinar um valor absoluto correspondente ao intervalo de tempo entre o início e o final de cada segmento fonético. No caso do modelo entoacional, cada segmento deve ter uma *curva* de F0 a ele associada: a junção dessas curvas determina o contorno de F0 da sentença como um todo. Obviamente, o valor final da curva associada a um segmento fonético deve coincidir com o valor inicial da curva associada ao segmento seguinte, de forma a garantir a continuidade do contorno global ao longo da sentença.

Muito embora o contorno de F0 não possua regras fixas que permitam a sua determinação de maneira única, normalmente ele segue alguns padrões de uniformidade. A incorporação desses padrões de comportamento constitui o primeiro passo na construção de um modelo entoacional.

Uma das tendências gerais observadas em sentenças declarativas é o declínio gradual do valor de freqüência fundamental. Essa tendência mostra-se mais verdadeira quanto mais neutra for a sentença.

Muito embora exista uma tendência de declínio gradual, o contorno de F0 de uma sentença apresenta, a nível de sílabas e fones, tanto trechos em que a derivada primeira é positiva como trechos em que a derivada primeira é negativa. Essa variação evidencia a estrutura prosódica mais fina da sentença. Normalmente, as regiões onde a inclinação da curva é maior estão associadas à localização de uma sílaba acentuada em relação às suas vizinhas. Além disso, os vales e picos da curva de F0 assumem, à medida que se percorre o enunciado, valores cada vez mais próximos entre si, ou seja, a variação de F0 tende a diminuir ao longo do enunciado [62].

Uma outra tendência freqüentemente observada a nível de constituinte prosódico é denominada de *padrão em chapéu.* Nesse caso, a curva de F0 mantém-se em um valor baixo até a primeira sílaba tônica do constituinte, quando se observa, então, uma elevação significativa. O nível de F0 mantém-se elevado até a última sílaba tônica, quando então torna a cair [36].

Assim como no caso da duração, há fatores que influenciam a determinação do contorno de F0 em níveis hierárquicos distintos. Normalmente, procura-se determinar inicialmente um contorno geral associado à sentença como um todo (declarativa, imperativa ou interrogativa); em seguida procura-se refinar esse contorno geral, levando-se em conta os fatores que atuam num nível mais baixo: constituinte prosódico, palavra, sílaba e, por fim, fone.

Para o português do Brasil, podemos citar os estudos sobre modelo entoacional desenvolvidos por Madureira [42][43].

# 7 Síntese do sinal de fala

## 7.1 Introdução

A última etapa a ser executada durante o processo de conversão texto-fala é a síntese do sinal propriamente dita. O processo de síntese consiste na geração de um sinal acústico correspondente ao texto de entrada, a partir das informações obtidas nas etapas anteriores do processo de conversão.

Existem diferentes estratégias que podem ser utilizadas para efetuar a geração do sinal de fala sintética. O objetivo final de cada uma dessas estratégias, no entanto, é sempre o mesmo: gerar um sinal acústico que corresponda à seqüência de fonemas determinada pelo módulo de transcrição ortográfico-fonética e aplicar à seqüência de fones correspondente os parâmetros prosódicos calculados na etapa de processamento prosódico. O mecanismo de síntese também deve procurar evitar descontinuidades ao longo do sinal gerado, de forma que a fala produzida ao final do processo seja inteligível e tão natural quanto possível.

Podemos dividir as diferentes estratégias utilizadas na geração da fala sintética em três métodos básicos: a *síntese por regras*, a *síntese concatenativa* e a *síntese articulatória*. Os três métodos se distinguem entre si quanto à forma pela qual manipulam a informação gerada nas etapas anteriores do processo de conversão. Muito embora a finalidade seja a mesma, existem diferenças quanto à qualidade do sinal gerado por meio de cada uma das estratégias.

Ao longo deste capítulo estaremos analisando as particularidades dos três métodos de síntese acima mencionados, levando em conta as vantagens e desvantagens de cada uma em relação aos demais.

## 7.2 Síntese por regras

O princípio de funcionamento do sintetizador por regras é baseado no modelo fonte-filtro da teoria acústica de produção da fala. Segundo esse modelo, o sinal de fala produzido pelo aparelho fonador humano corresponde ao resultado da passagem de uma fonte de excitação (que pode ser sonora, não sonora ou mista) por um filtro, cuja função de transferência é determinada pela configuração instantânea do trato vocal. Ao fornecermos um modelo adequado da fonte de excitação ao sintetizador, podemos supor que ele é capaz de produzir sinal de fala na sua saída, desde que o modelo seja capaz de simular a função de transferência do trato vocal humano. Em outras palavras, a qualidade do sinal sintético gerado depende do modelamento correto do processo de filtragem e também da fonte de excitação.

Um sintetizador adequado à síntese por regras deve possuir parâmetros de controle que permitam atuar sobre todas as características acústicas do sinal de fala, tanto aquelas relacionadas à fonte de excitação (como período de *pitch*, amplitude, presença ou ausência de ruído de aspiração, etc.), como também aquelas ligadas à configuração do trato vocal (freqüência, amplitude e largura de banda dos formantes, presença de pólos e zeros nasais, etc.).

A primeira versão do sintetizador de formantes de Klatt [38], apresentada em 1980, funcionou como alavanca inicial para os sistemas baseados em síntese por regras. O sintetizador proposto é composto por 39 parâmetros de controle, sendo 20 deles variáveis, que permitem o controle sobre as principais características acústicas do sinal de fala. Várias atualizações desse modelo foram propostas: a versão apresentada por Klatt & Klatt em 1990 [37] inclui diversos aperfeiçoamentos no modelo de fontes de voz, a fim de melhorar a qualidade da voz sintetizada, principalmente com relação a vozes femininas.

A idéia básica por trás do modelamento da função de transferência do trato vocal é fazer a associação de filtros de segunda ordem (pares de pólos complexos conjugados), cada um deles associado a um formante. Dentro dessa linha, há duas abordagens diferentes que

podem ser utilizadas. A associação de filtros em *cascata* é a mais intuitiva: nela as funções de transferência de cada filtro se multiplicam; por esse motivo, não é possível efetuar o controle individual sobre a amplitude dos formantes, pois as funções de transferência associadas a cada formante se sobrepõem. A associação de filtros em paralelo, por sua vez, permite efetuar o controle direto sobre a amplitude de cada formante, pois todos os filtros recebem a mesma entrada e podem ter ganhos individuais a eles associados.

O sintetizador de formantes de Klatt utiliza as duas configurações. A associação em cascata é própria para simular a produção de segmentos sonoros, como as vogais, em que a fonte sonora está localizada na laringe e o trato vocal inteiro funciona como um único ressoador. A associação em paralelo, por sua vez, é mais apropriada para modelar a produção de segmentos não sonoros, como no caso das fricativas, em que a fonte de ruído está localizada em algum ponto no interior da cavidade oral. Nesse caso, somente a porção do trato vocal posterior à localização do ruído funciona como ressoador.

A Figura 7-1 apresenta um diagrama de blocos do sintetizador de formantes de Klatt (modelo de 1980). Podemos perceber que a associação em cascata é composta por 8 filtros (linha horizontal superior), correspondentes a seis formantes e mais um par de pólo e zero nasais (RNP e RNZ, respectivamente). A associação em paralelo, por sua vez, possui um pólo nasal, mas nenhum zero (linha vertical à direita).

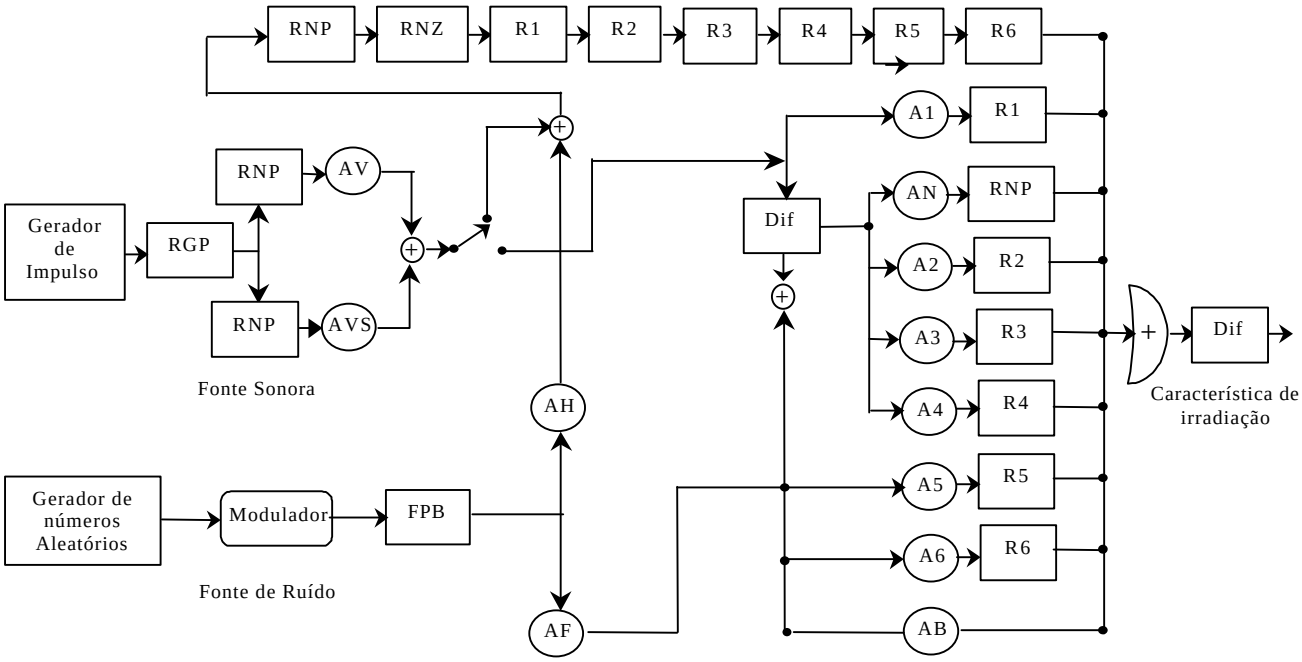



**Figura 7-1 Diagrama de blocos do sintetizador de formantes de Klatt (1980)**

Os parâmetros do sintetizador de formantes de Klatt são atualizados a cada 5ms; no caso dos ressoadores, esses parâmetros dizem respeito às freqüências de ressonância e larguras de banda dos formantes, bem como às amplitudes dos formantes da associação em paralelo.

Para que a qualidade do sinal sintetizado seja a mais natural possível não basta apenas modelar a função de transferência do trato vocal. É preciso efetuar também o modelamento correto do sinal proveniente da fonte de excitação que funciona como entrada para os ressoadores.

Esse sinal de entrada pode ser de dois tipos: sonoro e não sonoro. O modelamento da fonte sonora é feito a partir de um gerador de impulsos separados por um intervalo de tempo igual ao período de *pitch.* Esse trem de impulsos passa por alguns filtros cuja função é a de

suavizar o espectro, aproximando-o do espectro do pulso glotal (esses filtros são representados na Figura 7-1 por RGP, RGZ e RGS. AV e AVS, por sua vez, são parâmetros de controle de amplitude do sinal sonoro).

O sinal gerado pela fonte de ruído pode ser dividido em duas categorias: ruído de aspiração e ruído de fricação. A base para a criação dos dois tipos de ruído é um gerador de números aleatórios. O sinal obtido dessa maneira pode ser misturado à fonte sonora por meio de um modulador, de forma a simular a produção de sons de característica mista, como as fricativas vozeadas. O ruído de aspiração funciona como entrada para o ressoador em cascata, pois é produzido na laringe e, nesse caso, o trato vocal inteiro funciona como um único ressoador. O ruído de fricação, por sua vez, funciona como entrada do ressoador em paralelo. Na Figura 7-1, AH e AF funcionam como parâmetros de controle da amplitude dos ruídos de aspiração e de fricação, respectivamente.

Através do fornecimento correto de parâmetros de controle das fontes e dos ressoadores é possível simular a produção dos diversos tipos de sons gerados pelo aparelho fonador humano. Cada segmento fonético possui valores alvo para os parâmetros de controle do sintetizador. Para sintetizar uma seqüência de segmentos fonéticos é preciso adotar uma estratégia de atualização dos parâmetros de forma a gerar a transição adequada entre um segmento fonético e o seguinte, levando em conta os fenômenos de coarticulação. O controle da prosódia também é feito por meio da manipulação adequada dos parâmetros: a freqüência fundamental está diretamente associada ao período do gerador de impulsos; a duração, por sua vez, pode ser controlada a partir da taxa de atualização dos parâmetros do sintetizador.

Um sintetizador por regras bem projetado é capaz de gerar um sinal de fala de alta qualidade. Esse potencial foi demonstrado por Holmes [33] que, conforme já foi visto no capítulo 2, conseguiu gerar através da síntese por regras um sinal de fala indistinguível de um sinal de fala natural. Além disso, trata-se de uma técnica de síntese bastante flexível, pois permite a geração de diferentes qualidades de voz a partir do ajuste de parâmetros do modelo da fonte.

O principal impedimento prático à utilização da síntese por regras em um sistema de conversão texto-fala deve-se à dificuldade na determinação dos parâmetros de controle do sintetizador. Num sistema de síntese a partir de texto, esses parâmetros devem ser determinados de maneira automática. Para tanto, é preciso definir um número extenso de regras que dêem conta dessa determinação: a dificuldade maior se encontra em definir as regras de transição entre segmentos fonéticos, a fim de levar em conta os fenômenos de coarticulação. Muito esforço tem sido dispendido na determinação de regras adequadas à língua inglesa, de forma que resultados expressivos têm sido conseguidos para esse idioma a partir da utilização da síntese por regras. No caso dos outros idiomas, no entanto, o nível de sucesso até agora não tem sido o mesmo.

## 7.3 Síntese concatenativa

A idéia por trás da síntese concatenativa é a de gerar um sinal de fala artificial a partir da concatenação de segmentos pré-gravados de fala natural. Tais segmentos devem ser selecionados a partir de um inventário de unidades previamente construído, e o conteúdo desse inventário deve ser tal que seja possível sintetizar todas as seqüências fonéticas possíveis de serem realizadas dentro de uma determinada língua.

A principal decisão a ser tomada ao se projetar um sistema de síntese concatenativa diz respeito ao tamanho das unidades básicas de fala que irão constituir o inventário para concatenação. Como já foi visto anteriormente, a utilização de palavras inteiras não é conveniente no caso da síntese de fala a partir de texto irrestrito. Isso porque o número de palavras que podem ocorrer dentro de um texto genérico é imenso, e para efetuar a síntese concatenativa seria necessário que cada uma dessas palavras fosse previamente gravada e que estivesse armazenada em um banco de dados à disposição do sistema de síntese. Uma estratégia desse tipo demandaria um enorme custo de armazenamento, mesmo que essas palavras fossem guardadas sob forma paramétrica. E mesmo que isso fosse possível, haveria

ainda o problema dos neologismos e das novas siglas, o qual demandaria que o inventário de unidades fosse constantemente atualizado

A solução mais intuitiva para minimizar o custo de armazenamento seria a utilização de fones como blocos constituintes básicos do sinal de fala sintetizado. Essa parece ser uma solução coerente, pois a partir da seleção adequada de fones, podemos gerar qualquer seqüência possível de sons em uma dada língua. Além disso, o número de fonemas necessário para compor o inventário, muito embora varie de língua para língua, normalmente é inferior a uma centena, o que em termos de custo de memória é um valor extremamente pequeno.

O que se observa na prática, no entanto, é que a concatenação pura e simples de fones não é capaz de gerar um sinal de fala com qualidade satisfatória. Na maior parte das vezes o sinal de fala gerado por meio dessa estratégia não chega nem mesmo a ser inteligível, e muito menos natural. A principal razão disso é o já aqui discutido fenômeno da *coarticulação*. As características espectrais de um segmento fonético são fortemente influenciadas pelos segmentos subjacentes; por isso a utilização de um fone dentro de um contexto fonético muito diferente do qual ele foi extraído pode gerar descontinuidades espectrais significativas [32], o que do ponto de vista do ouvinte é tão catastrófico que muitas vezes o fonema original não pode nem ao menos ser identificado.

Deve-se, portanto, encontrar uma solução de compromisso: se por um lado a utilização de unidades extensas leva à necessidade de se utilizar inventários de tamanho impraticável, não se pode ignorar também as interações que ocorrem entre os segmentos fonéticos, e os critérios de definição sobre as unidades mais apropriadas à concatenação devem necessariamente levar em conta esse tipo de fenômeno.

Proposto inicialmente por Wang e Peterson [71], o *difone* constitui uma das alternativas mais comumente utilizadas na elaboração de inventários para síntese concatenativa. Como o próprio nome diz, o difone é uma unidade formada por uma dupla de fones: ele se inicia na metade do primeiro fone e termina na metade do fone seguinte. A

grande vantagem do difone é que a transição entre os fones está inteiramente contida no interior da unidade. A influência exercida por um fone sobre o seu sucessor se manifesta essencialmente nessa zona de transição; por isso o difone é capaz de capturar grande parte do fenômeno coarticulatório que ocorre entre segmentos subjacentes. O processo de concatenação sempre ocorrerá em porções espectralmente estáveis de fones de mesma natureza, minimizando-se com isso a ocorrência de descontinuidades.

Obviamente um dicionário de difones capaz de gerar todas as seqüências de sons possíveis será maior do que um simples dicionário de fones. Em princípio, deve-se levar em conta todas as combinações de fones dois a dois: no entanto, muitas dessas combinações não ocorrem de fato na língua e, portanto, não devem ser levadas em conta na elaboração do inventário de unidades para concatenação. No caso da língua portuguesa, pode-se construir um dicionário de difones  com pouco mais de 1.000 elementos.

Os difones são unidades curtas, e por mais cuidado que se tome no processo de concatenação, a ocorrência de descontinuidades na região de junção entre duas unidades costuma ocorrer e funciona como um fator de degradação da qualidade do sinal sintetizado. A utilização de unidades maiores é uma estratégia que visa exatamente a minimizar o número de junções necessárias no processo de síntese.

Além disso alguns segmentos fonéticos podem  ter uma duração muito curta em determinados contextos, ou então, por possuírem uma característica inerentemente dinâmica, não ter uma região espectralmente estável. O corte de segmentos desse tipo no ato da criação de difones não é aconselhável, pois a concatenação pode produzir resultados insatisfatórios. Por fim, muitas vezes o fenômeno de coarticulação pode se estender muito além do segmento fonético seguinte, atingindo alguns fonemas mais distantes. A utilização de unidades contendo apenas dois fones despreza esse tipo de ocorrência.

O emprego de *sílabas* como unidades básicas para concatenação pode, em princípio, parecer uma boa solução, pois a coarticulação que ocorre entre fonemas pertencentes a sílabas

diferentes normalmente é bem menor do que aquela que ocorre entre segmentos intra-silábicos. No entanto, o número de sílabas existentes na língua é muito grande; por conseguinte, a utilização de um inventário de sílabas completo certamente demandaria um alto custo de armazenamento. Uma alternativa proposta por Fujimura et al. [31] é a utilização de meias-sílabas ou *demissílabas.* No entanto, nem sempre é possível desprezar a interação que ocorre entre segmentos pertencentes a sílabas diferentes.

A tendência atualmente observada no desenvolvimento de sistemas de síntese concatenativa é a utilização de unidades com características mistas entre os difones e as demissílabas [52]. Tais unidades, que aqui denominaremos genericamente de *polifones*, podem ser formadas por dois ou mais fones em seqüência e, portanto, alguns segmentos fonéticos mais sujeitos à coarticulação aparecem inteiramente contidos dentro de uma única unidade. O sistema de síntese desenvolvido neste trabalho de Mestrado segue exatamente essa linha: os detalhes a respeito dos critérios utilizados na elaboração do inventário em questão serão expostos posteriormente.

Uma vez determinado o conjunto das unidades que irão compor o inventário para concatenação é necessário partir para o processo de criação efetiva dessas unidades. O processo consiste em efetuar a gravação de amostras de fala natural, cada uma delas contendo uma das unidades do inventário. Essas amostras de fala natural devem ser segmentadas, de forma a isolar as unidades nela contidas. Uma vez extraídas, as unidades serão armazenadas dentro do inventário e estarão à disposição do sistema de síntese quando necessário.

Uma série de cuidados deve ser tomada no processo de elaboração do dicionário acima descrito. Em primeiro lugar, é preciso selecionar adequadamente as amostras de fala natural que irão conter as unidades a ser segmentadas. Normalmente são utilizados *logatomas* (palavras sem sentido, mas que contêm a seqüência de fonemas que compõe a unidade em questão). Os logatomas são escolhidos de forma que o contexto fonético em torno dos fones da unidade a ser extraída seja o mais neutro possível, a fim de minimizar a coarticulação destes com os segmentos vizinhos que não fazem parte da unidade. Esses logatomas não são lidos

isoladamente, mas são inseridos no interior de *frases veículo.* O objetivo da utilização dessas frases é o de prover um ambiente prosodicamente neutro para as unidades. Essa neutralidade é desejável pois, no ato da síntese, as unidades devem se adequar a ambientes prosódicos os mais diversos possíveis.

Seleciona-se então um locutor que irá "emprestar" a sua voz ao sistema de síntese. Esse locutor deve efetuar a leitura de cada uma das frases-veículo utilizando taxa de elocução e entoação relativamente semelhantes entre as diversas sessões de gravação.

As frases gravadas devem então ser submetidas a um processo de segmentação. Primeiramente deve-se isolar os fones da unidade contida naquela frase. Em seguida deve-se escolher a região de corte adequada no primeiro e no último fone da unidade. Como o número de unidades a serem segmentadas é bastante grande, trata-se de uma tarefa extremamente trabalhosa. Existem alguns algoritmos que visam à realização desse trabalho de forma automática [61]. No entanto, esses algoritmos não são precisos, e qualquer tipo de erro na segmentação pode acarretar resultados ruins no processo de concatenação. O que se faz normalmente é utilizar a segmentação inteiramente manual, num processo trabalhoso porém preciso; ou então utiliza-se a segmentação automática com posterior ajuste manual das fronteiras de fones e das regiões de corte.

As etapas até então descritas correspondem somente à criação do dicionário de unidades: essa é uma tarefa que será executada apenas uma vez, durante o processo de construção do sistema. A tarefa de síntese propriamente dita consiste das seguintes etapas:

- escolher, a partir do dicionário, o conjunto de unidades que ao serem concatenadas corresponderão à seqüência fonética que se pretende sintetizar;
- efetuar a concatenação dessas unidades, procurando minimizar descontinuidades espectrais nas junções;

- promover as alterações prosódicas adequadas em cada um dos segmentos fonéticos constituintes da sentença, de forma que esta venha a ter o contorno prosódico determinado durante a etapa de processamento prosódico (e que corresponde ao conteúdo lingüístico expresso no texto escrito).

A escolha das unidades adequadas depende, obviamente, do conjunto das unidades que compõem o dicionário: supõe-se que o dicionário seja completo, ou seja, que permita a geração de qualquer seqüência fonética possível. Quanto ao problema da ocorrência de descontinuidades nas junções, ele pode ser bastante minimizado ao se tomar os cuidados necessários na gravação e segmentação das unidades. A escolha de ambientes fonética e prosodicamente neutros, com a utilização de logatomas e frases-veículo apropriados, visa exatamente a isso. A escolha do locutor também é importante, pois este deve procurar manter um padrão de *pitch*, taxa de elocução e qualidade de voz regulares ao longo de todo o processo de gravação. Os cuidados necessários com o equipamento de gravação e com o nível de ruído também devem ser tomados, a fim de evitar, por exemplo, descontinuidades no padrão de amplitude das diferentes unidades. Por fim, deve-se ressaltar também a importância da determinação adequada das regiões de corte das unidades. Elas devem ocorrer preferencialmente em regiões espectralmente estáveis do sinal. Além disso, nos segmentos sonoros (vozeados), o corte deve ser efetuado em pontos equivalentes do sinal periódico (a segmentação deve ser síncrona com o período de *pitch*), a fim de minimizar a ocorrência de descontinuidade de fase após a concatenação.

Mesmo com todos esses cuidados, é preciso que a técnica de síntese utilizada no ato da concatenação também admita mecanismos de suavização das descontinuidades nas junções das unidades: ao analisarmos mais adiante algumas técnicas de síntese concatenativa, veremos a maneira pela qual elas lidam com esse problema.

Para gerar o sinal de fala sintetizada, no entanto, não basta apenas efetuar a concatenação das unidades selecionadas. Isso porque os parâmetros prosódicos dos fones contidos nessas unidades dizem respeito ao contexto prosódico das frases-veículo das quais as

unidades foram extraídas, mas não correspondem necessariamente aos parâmetros que foram calculados durante a etapa de processamento prosódico para os fones da sentença a ser sintetizada. Ou seja, o mecanismo de síntese deve prover meios de alterar os parâmetros prosódicos (duração, F0 e amplitude em uma menor medida) dos fones contidos nas unidades, de forma que a sentença final sintetizada possua um padrão prosódico apropriado.

Uma das maiores vantagens da síntese concatenativa encontra-se em sua relativa simplicidade. O tempo necessário para a construção de um sistema desse tipo é bem menor do que o tempo associado à construção de um sistema de síntese por regras, por exemplo, que exige uma fase extremamente meticulosa na determinação das regras necessárias para controlar parâmetros do sintetizador. No caso da síntese concatenativa, o maior trabalho corresponde justamente à elaboração do inventário de unidades. Trata-se de uma tarefa essencialmente mecânica e perfeitamente realizável, desde que se disponha de equipamento adequado e pessoal treinado. A qualidade do sinal sintetizado por um sistema desse tipo mostra-se equivalente à dos melhores sintetizadores por regras; por esse motivo, trata-se de uma das estratégias de síntese mais utilizadas nos sistemas atualmente em desenvolvimento, especialmente pelos sistemas que trabalham com línguas diferentes do inglês [65].

Obviamente, a síntese concatenativa também apresenta algumas desvantagens. A primeira delas diz respeito às distorções introduzidas no sinal sintetizado devido ao processo de concatenação. Esse tipo de efeito pode ser atenuado, mas jamais compensado totalmente. A síntese concatenativa mostra-se ainda menos flexível do que a síntese por regras no que diz respeito à qualidade de voz sintetizada. Um sintetizador por regras possui parâmetros que permitem controlar livremente as características do trato vocal e do pulso glotal de excitação: é possível, pelo menos em princípio, produzir vozes masculinas, femininas ou infantis, com padrões de F0 mais graves ou mais agudos, com diversos níveis de aspiração, e daí por diante. No caso da síntese concatenativa, a qualidade de voz é sempre a mesma, pois está atrelada às características do locutor que "emprestou" a sua voz durante o processo de gravação das unidades do inventário. Alterações prosódicas muito intensas, com o intuito de, por exemplo,

alterar o padrão de F0 do sinal original, costumam introduzir distorções na fala sintetizada. O grau de distorção, nesse caso, depende da técnica de síntese utilizada na manipulação do sinal.

Dentre as técnicas de síntese normalmente utilizados, a técnica LPC [21] destaca-se pela sua simplicidade. As unidades são armazenadas sob forma de parâmetros LPC, o que certamente representa uma economia considerável em termos de espaço de armazenamento. A síntese por técnica LPC provê meios de atenuar os efeitos de descontinuidade espectral, através da interpolação dos coeficientes LPC nas regiões de junção entre as unidades. Além disso, é possível efetuar alterações nos parâmetros prosódicos do sinal original, a partir da manipulação adequada dos parâmetros [49].

No entanto, a técnica LPC apresenta algumas limitações. Em primeiro lugar, o filtro LPC consiste de um função de transferência formada apenas por pólos e, portanto, não é capaz de modelar adequadamente os sons que contenham zeros na função de transferência do trato vocal, como por exemplo as vogais nasais. O modelo LPC simples também considera apenas dois tipos de excitação (sonora ou não sonora): essa abordagem compromete o modelamento de sons de características mistas como as fricativas vozeadas. Por fim, existem alguns erros na estimação espectral que são inerentes ao modelo de análise LPC. Num processo simples de análise e ressíntese esses erros acabam se compensando; no caso de se efetuar alterações prosódicas no sinal original, no entanto, esses erros acabam ocasionando uma degradação considerável no sinal sintetizado, pois os cálculos são realizados em cima de uma envoltória espectral incorreta.

Por esses motivos, a qualidade do sinal gerado por meio da técnica LPC não é plenamente satisfatória. Atualmente existem outras técnicas capazes de gerar um sinal sintetizado de qualidade superior. À medida que os dispositivos de memória se tornam mais baratos, a preocupação em torno do custo de armazenamento se torna menos importante, por isso já se pode pensar em técnicas de síntese que operem diretamente sobre a forma de onda do sinal.

Duas das técnicas de síntese capazes de gerar melhores resultados são *as técnicas PSOLA* e a *síntese híbrida*. A seguir, estaremos analisando cada uma delas com detalhes.

### 7.3.1 Técnicas PSOLA (Pitch-Synchronous Overlap and Add)

Os algoritmos de síntese do tipo PSOLA (*Pitch-Synchronous Overlap and Add*) caracterizam-se por sua extrema simplicidade do ponto de vista conceitual e por serem capazes de gerar, não obstante, um sinal sintetizado de alta qualidade.

Uma característica dos algoritmos PSOLA é o fato de que eles trabalham de maneira síncrona com o período de *pitch* do sinal. Por isso, a qualidade do sinal gerado é intimamente dependente da existência de um algoritmo de marcação de *pitch* eficiente.

O sinal de voz a ser processado deve ser submetido a um algoritmo de marcação de *pitch*. No caso da síntese concatenativa, essa marcação é efetuada uma única vez, durante a confecção do inventário de unidades. As marcas são posicionadas nos picos do sinal nas porções sonoras e espaçadas de um valor fixo (tipicamente 10ms) nas porções não sonoras. Independentemente do segmento ser ou não sonoro, essas marcas são designadas marcas de *pitch.*

Dentre os algoritmos da classe PSOLA o mais utilizado é sem duvida o TD-PSOLA [47] (T*ime-Domain Pitch-Synchronous Overlap and Add*). Sua grande popularidade provém do fato de que se trata de um algoritmo extremamente simples e de custo computacional bastante baixo, capaz de realizar o processo de síntese praticamente em tempo real, gerando um sinal de alta qualidade. A razão principal por trás de sua simplicidade computacional reside em um fator principal: o algoritmo não exige qualquer tipo de análise espectral do sinal, trabalhando diretamente sobre a forma de onda.

A primeira etapa do algoritmo PSOLA consiste em particionar o sinal a ser modificado (chamado sinal de análise) em uma seqüência de sinais menores, denominados sinais elementares. Esse particionamento é feito de modo que a soma dos sinais elementares

corresponda ao sinal original. Para alcançar esse resultado submete-se o sinal original a uma seqüência de janelamentos de Hanning, onde a freqüência de análise é síncrona com o período de *pitch* do sinal. O janelamento é feito de modo que haja sobreposição de 50% entre janelas adjacentes. A Figura 7-2 ilustra esse processo:

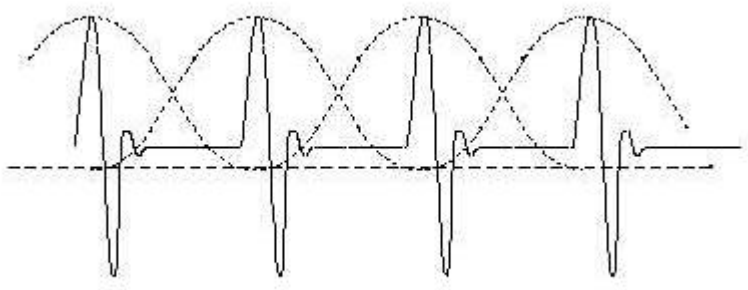


**Figura 7-2 Janelamento do sinal de análise**

Em seguida, uma nova seqüência de sinais, denominada seqüência de síntese, é gerada a partir da seqüência de análise descrita anteriormente. Para isso, os sinais elementares são manipulados, de forma a alterar os parâmetros prosódicos do sinal original. Há dois tipos de manipulações que podem ser efetuadas no sinal original: alteração da duração e da freqüência fundamental.

#### *7.3.1.1. Alteração da duração*

O procedimento básico para alterar a duração do sinal consiste em omitir ou duplicar alguns dos seus sinais elementares. A omissão é utilizada quando se deseja diminuir a duração, ao passo que a duplicação permite aumentar a duração do sinal. Em ambos os casos, o número de sinais elementares omitidos (ou duplicados) determina a nova duração do sinal. O processo de alteração da duração pode ser feito tanto com as marcas sonoras como também com as

marcas não sonoras. As figuras a seguir ilustram o processo acima descrito: na Figura 7-3, o terceiro sinal elementar é omitido, de forma que duração total do sinal é reduzida; já na Figura 7-4 o sinal sintetizado possui uma duração maior que a do sinal original, uma vez que o terceiro sinal elementar foi duplicado no processo de síntese.

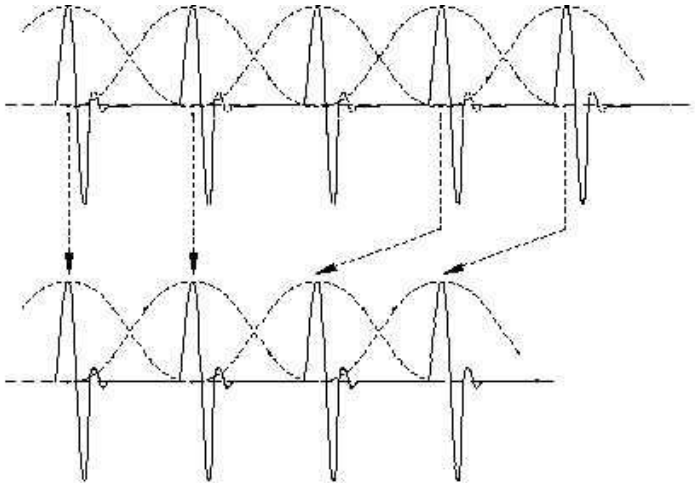


**Figura 7-3 Redução da duração de um sinal de voz por omissão de sinais elementares**

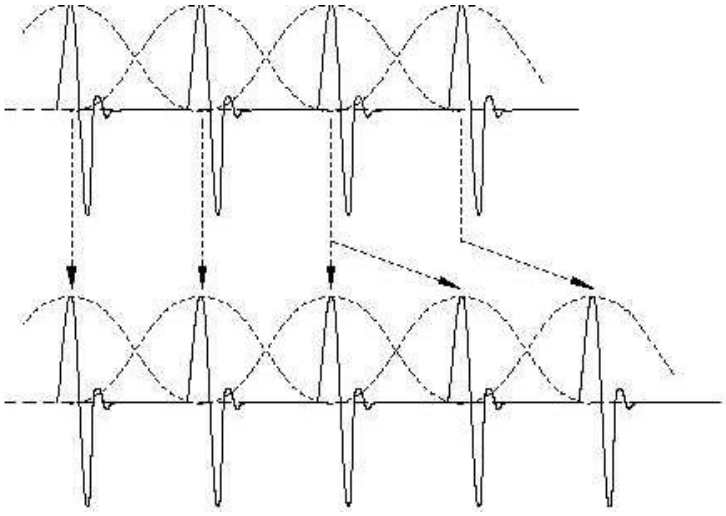


**Figura 7-4 Aumento da duração de um sinal de voz por duplicação de sinais elementares**

### *7.3.1.2. Alteração da freqüência fundamental*

Para alterarmos a freqüência fundamental do sinal original devemos modificar o intervalo de tempo entre 2 sinais elementares. Ao aumentarmos o intervalo de tempo, diminuímos a freqüência, e vice-versa.

Matematicamente temos:

$\Delta t_a$ = intervalo de tempo entre 2 sinais elementares do sinal de análise;

$\Delta t_b$ = intervalo de tempo entre 2 sinais elementares do sinal de síntese;

$\beta$ = fator de alteração da freqüência.

$$\Delta t_b = \Delta t_a / \beta$$

Nesse caso, a freqüência do sinal resultante é $\beta$ vezes a freqüência do sinal original.

Diferentemente do que ocorre no caso da alteração da duração, a alteração da freqüência é feita apenas com as marcas sonoras do sinal de análise.

As figuras a seguir ilustram o processo descrito acima. Na Figura 7-5, a freqüência do sinal foi aumentada, ao passo que a Figura 7-6 ilustra um sinal resultante com uma freqüência menor que a do sinal original.

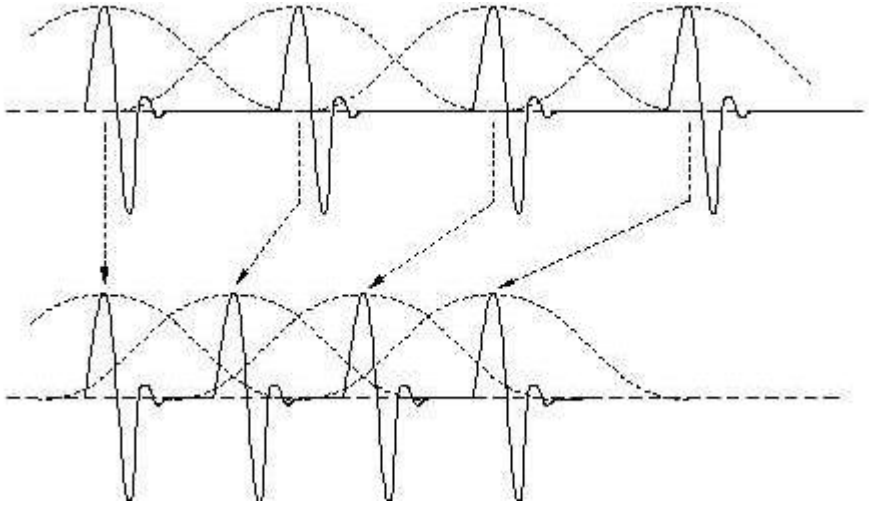


**Figura 7-5 Aumento da freqüência fundamental de um sinal**

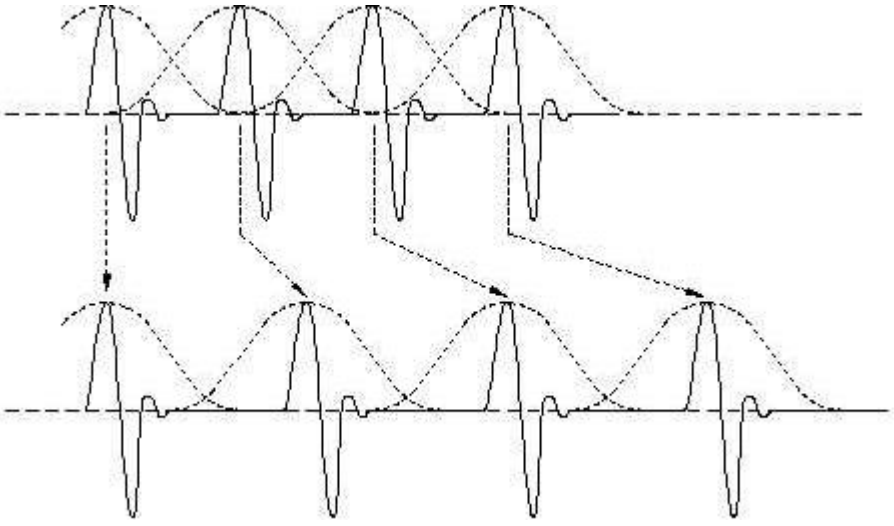


**Figura 7-6 Redução da freqüência fundamental de um sinal**

A última etapa do algoritmo PSOLA consiste em simplesmente somar os sinais elementares que irão compor a seqüência de síntese.

Apesar de todas as vantagens, o algoritmo TD-PSOLA apresenta também algumas limitações, que são apresentadas a seguir

Em primeiro lugar, o algoritmo não permite multiplicar a duração original do sinal por valores de expansão ou compressão quaisquer. Isso porque, para expandir (ou comprimir) o sinal, deve-se replicar (ou suprimir), necessariamente, um número inteiro de janelas de análise; por isso o valor global de duração só pode ser diminuído ou aumentado de um valor múltiplo do período de *pitch*.

Um outro problema típico da técnica TD-PSOLA ocorre ao se tentar efetuar um aumento de duração em segmentos de fala não sonoros. Tais segmentos possuem característica aperiódica; no entanto, ao efetuar-se a replicação de algumas janelas de análise a fim de expandir a duração, introduz-se no sinal uma certa periodicidade indesejada, que se manifesta auditivamente através de um murmúrio metálico de fundo no sinal sintetizado. Uma estratégia proposta por Charpentier [14] para minimizar esse efeito consiste em inverter o eixo do tempo a cada réplica da mesma janela. Com isso, conserva-se o espectro de amplitude do sinal e inverte-se o espectro de fase, diminuindo-se portanto a correlação total do sinal gerado. Essa técnica pode ser utilizada apenas em porções não sonoras do sinal; no entanto não se aplica aos segmentos mistos como as fricativas vozeadas.

Além disso, ao se efetuar uma alteração na freqüência fundamental do sinal, o espaçamento entre as janelas é modificado, e com isso a duração do sinal também é afetada de maneira indesejada. Portanto, modificações da freqüência fundamental devem vir sempre acompanhadas de correções do fator de duração a fim de compensar essa distorção.

As alterações de F0 afetam, ainda, o nível de superposição entre janelas de análise, de forma que o grau de superposição entre janelas de Hanning não será mais de 50%: quanto maior o desvio, maior será a distorção introduzida no sinal sintetizado. Muito embora o efeito dessa superposição alterada seja muito menor do que se poderia supor em princípio (e nisso

reside a simplicidade do TD-PSOLA), variações prosódicas muito intensas podem introduzir distorções consideráveis no sinal.

Por fim, o algoritmo não possui estratégias adequadas para minimizar descontinuidades nas junções entre as unidades. Descontinuidades de *pitch* ocorrem quando o *pitch* final da primeira unidade for diferente do *pitch* inicial da unidade seguinte. A manutenção de um padrão de voz constante ao longo do processo de gravação das unidades pode minimizar esse problema, mas não eliminá-lo. Descontinuidades de fase, por sua vez, ocorrem porque as janelas de análise nem sempre estão posicionadas no mesmo ponto em relação ao período no sinal. O cuidado na confecção do corte das unidades também ajuda a minimizar esse problema, mas não o elimina completamente. Por fim, descontinuidades na envoltória espectral também podem ocorrer, pois o TD-PSOLA não possui nenhuma estratégia de manipulação espectral das unidades: se os espectros não forem idênticos, fatalmente a descontinuidade se manifestará após a concatenação.

O FD-PSOLA (F*requency-Domain Pitch-Synchronous Overlap and Add*) [14] é uma versão do PSOLA que realiza a manipulação espectral do sinal. Trata-se de um algoritmo mais eficiente, pois é capaz de lidar melhor com as questões de descontinuidade espectral acima descritas: no entanto, esse tratamento adicional torna o algoritmo computacionalmente muito mais custoso do que o TD-PSOLA tradicional.

Uma alternativa interessante, baseada também no TD-PSOLA tradicional, é o MBR-PSOLA (*Multi-Band Pitch-Synchronous Overlap and Add*) [20]. A idéia por trás do MBR-PSOLA é a de efetuar um processo de análise/ressíntese no inventário de unidades. Baseado no modelo MBE (*multi-band excited*), esse processo procura efetuar uma normalização, fazendo ajuste de fase e atribuindo um valor de F0 constante para todas as unidades do inventário. O processo de análise/ressíntese causa pouca degradação no sinal original, e elimina os problemas de descontinuidade de *pitch* e de fase que costumam afetar o resultado da concatenação das unidades. Além disso, a ressíntese permite lidar com o problema de descontinuidade espectral de maneira extremamente simples, pois ela faz com que a

interpolação da envoltória espectral equivalha a uma interpolação simples no domínio do tempo, obedecidas as condições de igualdade de fase e de F0. Portanto, ao introduzir um algoritmo de interpolação temporal que atue nos períodos próximos à região de junção, consegue-se uma suavização espectral impossível de se obter pela técnica TD-PSOLA tradicional. Trata-se de uma operação simples, realizada sempre no domínio do tempo.

O custo computacional introduzido é pequeno, uma vez que o processo de análise/ressíntese é efetuado uma única vez durante a criação do inventário de unidades. Recentemente, o algoritmo MBR-PSOLA foi modificado, dando lugar a uma versão mais eficiente denominada MBROLA (*Multi-Band Resynthesis Overlap and Add*).

### 7.3.2 Síntese Híbrida

A técnica de síntese híbrida [70], assim como a técnica PSOLA, trabalha de forma síncrona com o período de *pitch*, e possui essa denominação por ter o seu funcionamento baseado na decomposição do sinal original em duas componentes distintas: uma componente harmônica e uma componente de ruído. A componente harmônica é modelada como uma soma de senóides com freqüências múltiplas da freqüência fundamental, e a componente de ruído como uma excitação aleatória (ruído branco) aplicada a um filtro LPC. No processo de síntese, cada uma das componentes é submetida a um processamento distinto a fim de efetuar as modificações prosódicas, sendo que o sinal sintetizado corresponde à soma das duas componentes após efetuar-se esse processamento.

A técnica híbrida é, portanto, composta por duas etapas: uma etapa de *análise*, que é efetuada uma única vez durante a criação do inventário de unidades, na qual são calculadas as componentes harmônica e de ruído do sinal original; e uma etapa de *síntese*, na qual cada uma das componentes do sinal é submetida às alterações prosódicas necessárias.

#### *7.3.2.1. Análise*

Antes de efetuar a decomposição do sinal, é necessário que este seja submetido a um processo de marcação de *pitch*. Tal processo de marcação é idêntico àquele já descrito para a técnica PSOLA.

Para calcular a componente de ruído nas porções sonoras do sinal é necessário calcular uma componente harmônica e subtrair essa componente do sinal original. Já nas porções não sonoras, assume-se que não há componente harmônica, portanto o sinal original e a componente de ruído são equivalentes.

Para calcular a componente harmônica a ser subtraída do sinal original, utiliza-se um procedimento baseado em análise DFT sobre segmentos superpostos e centrados nas marcas de *pitch*, que funciona como um filtro passa-baixas de banda variável. Para cada marca de *pitch pm(i)* utiliza-se uma janela de análise assimétrica de comprimento *N*, onde *N=pm(i+1)-pm(i-1)*. O procedimento consiste em determinar um valor *fm(i)* para cada marca de *pitch*, correspondente ao valor de freqüência máxima da análise DFT para o segmento associado a essa marca.

Para calcular o valor de *fm(i)*, efetua-se o cálculo da DFT do segmento até 8 kHz, e calcula-se a porcentagem de energia contida nas faixas 0 kHz a 2 kHz, 2 kHz a 3 kHz, 3 kHz a 4 kHz e 4 kHz a 5 kHz. Como o cálculo da DFT é feito sobre um segmento cujo comprimento é dois períodos de *pitch*, somente as harmônicas pares são consideradas para o cálculo da energia, pois somente elas são harmônicas da freqüência fundamental. Se a porcentagem de energia contida entre 4 kHz e 5 kHz foi maior que 1% da energia total, então *fm(i)* = 5 kHz. Caso contrário, se a porcentagem de energia contida entre 3 kHz e 4 kHz foi maior que 1% da energia total, então *fm(i)* = 4 kHz. Esse procedimento continua para as outras faixas de freqüência; por fim, se porcentagem de energia contida entre 2 kHz e 3 kHz for menor que 1%, então *fm(i)* = 2 kHz.

A utilização de um valor distinto de *fm* para cada janela de análise visa a refinar a decomposição do sinal em suas componentes, pois a escolha de um valor baixo de *fm* implica em modelar parte da componente harmônica como componente de ruído, ao passo que a seleção de um valor de *fm* muito elevado faz com que parte da componente de ruído seja modelada como fazendo parte da componente harmônica. Isso introduz distorções na etapa de síntese.

Uma vez selecionados os valores de *fm(i)*, calcula-se a DFT inversa de cada segmento. O resultado é submetido a uma janela de Hanning assimétrica, centrada em *pm(i)* e estendendo-se de *pm(i-1)* a *pm(i+1)*. Os segmentos janelados são somados e o resultado corresponde à componente harmônica, que deve ser subtraída do sinal original para obter-se a componente de ruído. A Figura 7-7 ilustra esse procedimento.

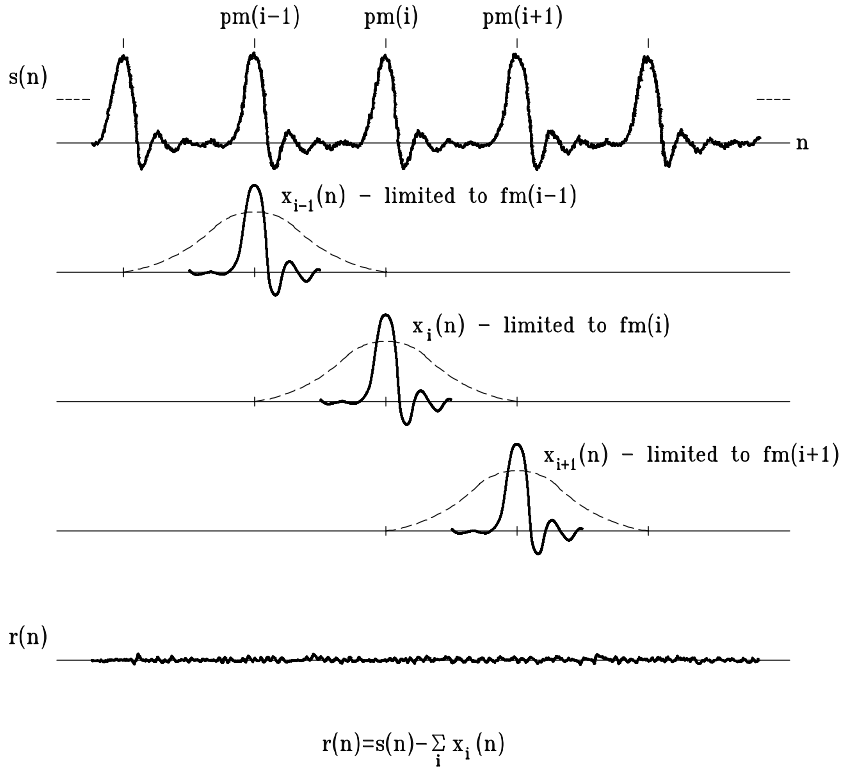



**Figura 7-7 Análise DFT utilizada para o cálculo da componente de ruído.**

Os parâmetros da componente de ruído relativos à marca *pm(i)* correspondem aos coeficientes *parcor* de um filtro LPC de ordem 15, bem como aos ganhos do filtro, calculados a intervalos de 2 ms, entre *pm(i-1)* e *pm(i)* [70].

A componente harmônica já calculada não se presta à tarefa de síntese, pois a análise foi feita sobre segmentos com comprimento de dois períodos de *pitch* e os parâmetros calculados correspondem a harmônicas de F0/2. Um novo conjunto de parâmetros harmônicos é então calculado. Dessa vez, utilizam-se janelas simétricas, centradas nas marcas de *pitch*, estendendo-se de *pm(i-1)* a *pm(i)+[pm(i)-pm(i-1)-1]*. Nesse novo contexto, a utilização de uma análise DFT considerando apenas as harmônicas de F0 pode ocasionar descontinuidades nas regiões de *pitch*, por isso o cálculo dos parâmetros harmônicos é efetuado a partir do critério de minimização do erro quadrático médio, ponderado por uma janela de Hamming, entre o sinal original e a componente harmônica. A minimização desse erro corresponde à solução de um sistema sobredeterminado com 2N equações lineares e 2K+1 variáveis:

$$e^2 = \sum_{n=0}^{2N-1} w_H(n)[s(n) - s_H(n)]^2$$

$$s_H(n) = \sum_{k=0}^{K} A_k(i)\cos(kw_0(i)n + q_k(i))$$

$$kw_0(i) \le 2pfm(i)/F$$

onde:

$N = pm(i) - pm(i-1);$

$w_H(n)$ é a janela de Hamming;

$s(n)$ é o sinal original;

$s_H(n)$ é a componente harmônica;

$w_0(i) = 2\pi/N$;

$A_k(i)$, $q_k(i)$ são os parâmetros harmônicos associados à marca *i;*

*F* é a freqüência de amostragem.

#### 7.3.2.2. Síntese

A etapa de síntese corresponde à introdução de variações prosódicas no sinal original. Nesse caso, as marcas de *pitch pm(i),* às quais estão associados os parâmetros harmônicos e de ruído, devem ser mapeadas para um novo conjunto de marcas *m(i).*

Sejam $a(i)$ e $b(i)$ os fatores de alteração de duração e de F0, respectivamente. Na presença de alterações de duração as novas marcas devem seguir a seguinte relação:

$$m[i] - m[i-1] = a[i](pm[i] - pm[i-1])$$

No caso da componente de ruído faz-se apenas modificações de duração, pois esta não possui valor de F0 a ela associado. Os valores de ganho, calculados a intervalos de 2 ms na etapa de análise, são atualizados a cada $2a(i)$ ms e variados linearmente dentro desse intervalo. Para o cálculo da nova componente de ruído entre duas marcas consecutivas, faz-se uma combinação do ruído produzido pelos dois conjuntos de parâmetros. Os dois sinais calculados são multiplicados por janelas de Hanning e submetidos a um processo de *Overlap and Add* para o cálculo do sinal resultante.

A componente sonora pode ser submetida a alterações de duração e também de F0. A síntese é feita por meio da aplicação do modelo de síntese senoidal [45][44], que providencia uma evolução gradual da amplitude, freqüência e fase entre duas marcas adjacentes. No caso de variações de F0, deve-se fazer uma reamostragem da envoltória espectral, a fim de obter as novas harmônicas, múltiplas da freqüência fundamental $b[i]w_0[i]$. A evolução gradual entre dois conjuntos de parâmetros é conseguida por meio de uma interpolação cúbica para a fase e uma interpolação linear para a amplitude.

Ao se fazer uma alteração de duração as novas marcas *m[i]* não correspondem mais, necessariamente, às marcas de *pitch*. Um número não inteiro de marcas de *pitch* pode ocorrer entre duas marcas de síntese; por isso, ao iniciar-se a síntese para cada nova marca deve-se efetuar uma correção de fase (atraso $t_0[i]$) a fim de evitar descontinuidades. O cálculo da correção associado à k-ésima harmônica é dado pelo conjunto de equações a seguir:

$$
\begin{aligned}
&q'_k(i) = q_k(i) - kb(i)w_0(i)t_0(i) \\
&t_0(i) = t_0(i-1) + L(i)pitch'(i) - Int\{a(i)pitch(i)\} \\
&pitch(i) = pm(i) - pm(i-1) \\
&pitch'(i) = pitch(i)/b(i)
\end{aligned}
$$

onde *L(i)* é um inteiro escolhido de forma a minimizar o atraso $t_0[i]$.

A Figura 7-8 ilustra as alterações de duração e de F0 feitas na síntese híbrida.

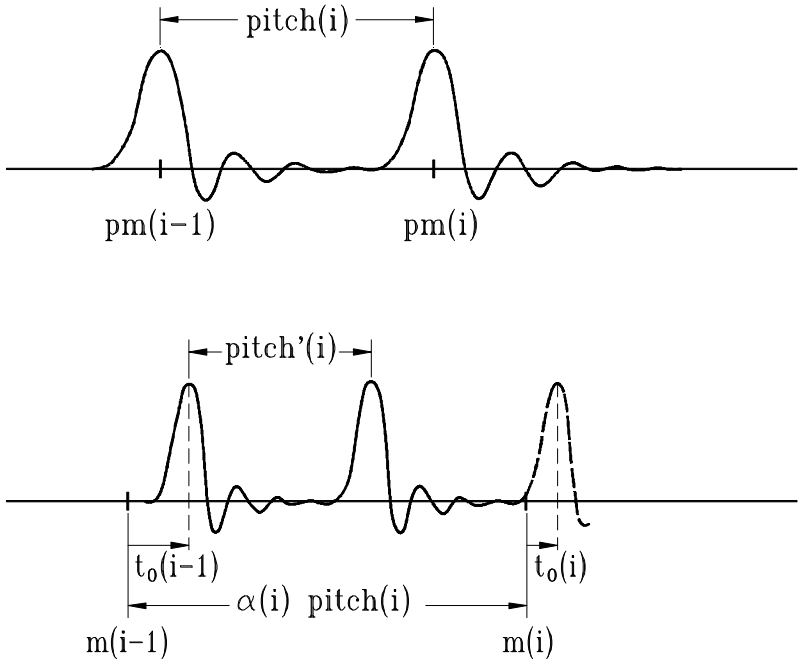



**Figura 7-8 Síntese híbrida com alteração de duração e F0.**

A síntese híbrida apresenta algumas vantagens em relação à síntese TD-PSOLA. Em primeiro lugar, as alterações de duração e de F0 são feitas de forma independente, diferentemente do que acontece com técnica TD-PSOLA, na qual as alterações em F0 introduzem automaticamente uma alteração de duração que deve ser compensada. Além disso, a técnica híbrida permite efetuar alterações de duração mais precisas, com erro de apenas uma amostra. No caso da técnica TD-PSOLA, vale lembrar que as alterações na duração do sinal devem ser sempre múltiplas do período de *pitch.*

Um dos efeitos colaterais da técnica TD-PSOLA, que é a introdução de ruído metálico quando se faz o aumento de duração em segmentos não sonoros, também não ocorre na síntese híbrida. Por fim, a técnica híbrida mostra-se mais robusta quando há a necessidade de efetuar alterações prosódicas largas, pois nesse caso as distorções introduzidas são menores em relação àquelas introduzidas pela síntese TD-PSOLA.

A técnica híbrida apresenta, porém, algumas desvantagens. Em primeiro lugar, ela é mais sensível a erros de marcação de *pitch*; além disso, a determinação incorreta das componentes harmônica e de ruído do sinal original podem introduzir alterações graves na qualidade do sinal sintetizado, pois elas são submetidas a processamentos bastante diferentes durante a etapa de variação dos parâmetros prosódicos. Por fim, ela é computacionalmente mais custosa que a técnica TD-PSOLA.

## 7.4 Síntese articulatória

Os métodos de síntese até aqui discutidos trabalham com o objetivo principal de gerar um sinal de fala sintético com características acústicas as mais próximas possíveis das de um sinal de fala natural. Não existe, no entanto, a preocupação de que os mecanismos através dos quais esse sinal sintético é construído se aproximem do processo natural de produção do sinal

acústico de fala pelo aparelho fonador humano. Eles se preocupam apenas em modelar a saída do mecanismo de produção da fala.

Os modelos baseados na *síntese articulatória* diferenciam-se dos anteriores por procurarem simular de maneira mais realista o processo de geração do sinal acústico de fala. A intenção de tal abordagem é a de que um modelo construído dessa maneira alcance também um maior realismo quanto à qualidade do sinal sintetizado.

O funcionamento de um sistema de síntese articulatória baseia-se na construção de um modelo físico o mais realista possível do aparelho fonador humano, capaz de mimetizar a dinâmica dos diversos articuladores no processo de produção da fala. As posições desses articuladores (língua, mandíbula, lábios, osso hióide, véu palatino, etc.) correspondem às variáveis do modelo. Tal modelo incorpora um controle individual sobre os articuladores, que se movimentam de maneira semi-independente uns dos outros. Essa relativa independência permite o modelamento da superposição de gestos articulatórios que normalmente ocorre no processo natural de produção da fala. Por outro lado, deve-se levar em conta também os limites fisiológicos de movimentação dos articuladores, bem como as interações na movimentação dos articuladores entre si. Por isso, a movimentação de certos articuladores é expressa tendo em vista o seu posicionamento relativo a um outro articulador (o movimento da ponta da língua, por exemplo, é dependente do movimento do corpo da língua, cujo posicionamento, por sua vez, depende do grau de abertura da mandíbula). Tais interações, inerentes ao modelo, limitam o número de configurações possíveis do espaço articulatório: essa é uma característica típica desse mecanismo de síntese, pois a geração do sinal de saída está limitada apenas aos padrões fisicamente realizáveis. As técnicas de síntese vistas anteriormente não forneciam nenhum tipo de mecanismo de controle equivalente.

O modelo acima exposto permite especificar os diferentes sons da língua, cada um deles caracterizado por uma dada configuração do espaço articulatório. A observação da movimentação dos articuladores durante os eventos de produção da fala natural permitem a derivação de regras de controle para o modelo do sintetizador, as quais são utilizadas na

geração do sinal de saída. Ou seja, uma vez definidos alvos correspondentes à seqüência fonética a ser sintetizada, procura-se, a partir do fornecimento de um conjunto de regras de controle, determinar uma trajetória no espaço articulatório que passe por esses valores alvo, obedecendo as especificações espaciais e temporais apropriadas.

Os estudos a respeito da síntese articulatória são relativamente recentes [15] [46]. Os sistemas até então desenvolvidos confirmam o potencial de produção de um sinal de fala de alta qualidade por meio de tal abordagem, mas limitam-se, por enquanto, à geração de segmentos curtos de fala [57]. Ainda existe um longo caminho a ser percorrido até que seja possível construir um sistema de conversão texto-fala eficiente baseado em síntese articulatória.

Alguns dos fatores que dificultam o desenvolvimento de tais modelos são a extrema dificuldade de obtenção de dados sobre a movimentação dos articuladores durante o processo de produção da fala natural, bem como a altíssima complexidade computacional associada à simulação dos modelos articulatórios. Além disso, no caso dos sistemas de conversão texto-fala, podemos ressaltar a dificuldade de se obter os comandos de controle do modelo articulatório a partir da transcrição fonética e da análise prosódica do texto de entrada. No caso desta última, dever-se-ia representar a prosódia em termos articulatórios, sobretudo duração e amplitude.

# 8 Implementação de um sistema de conversão texto-fala para o português falado no Brasil

## 8.1 Introdução

Ao longo dos capítulos anteriores procurou-se apresentar os sistemas de conversão texto-fala de maneira genérica. Dentro dessa abordagem, foi introduzido, inicialmente, o problema que um sistema de conversão texto-fala se propõe a resolver (a transformação de um texto genérico num sinal de fala inteligível e natural). A partir da exposição do problema, foram discutidas as principais dificuldades que tornam a implementação de um sistema desse tipo uma tarefa extremamente complexa. Vimos que muitas dessas dificuldades provêm da dificuldade de se realizar uma análise lingüística adequada a partir do texto de entrada, observação esta que nos fez constatar a importância da inclusão de módulos de processamento lingüístico operando dentro do sistema de conversão.

Apresentado o problema e expostas as suas dificuldades principais procurou-se, em seguida, entrar em detalhes a respeito das etapas que um sistema de conversão texto-fala deve percorrer para transformar o texto de entrada no sinal de fala correspondente. Para cada etapa foram apresentadas, detalhadamente, as dificuldades inerentes, bem como diferentes estratégias para a resolução do problema em questão. Não foi dado, até agora, nenhum tipo de enfoque diferenciado a qualquer tipo de abordagem de resolução, pois o objetivo até este ponto da dissertação foi o de apresentar o estado da arte em sistemas de síntese de fala a partir de texto.

A partir deste capítulo da tese será apresentado o trabalho que vem sendo desenvolvido no Laboratório de Processamento Digital de Fala do Departamento de Comunicações da Faculdade de Engenharia Elétrica e de Computação da UNICAMP, em conjunto com o

Laboratório de Fonética Acústica e Psicolingüística Experimental (LAFAPE) do Instituto de Estudos da Linguagem (IEL), também da UNICAMP, no sentido de implementar um sistema de conversão texto-fala para o português do Brasil.

Poucas são as atividades dentro da comunidade científica na direção da construção de sistemas de síntese para a língua portuguesa. No Brasil, a maior concentração de esforços acontece justamente no Laboratório de Processamento Digital de Fala da UNICAMP. Dentre os trabalhos desenvolvidos pela equipe da UNICAMP, podemos destacar os de Egashira [22], na construção de um protótipo de um sistema de conversão texto-fala, e o de Silva [62], na elaboração de um modelo prosódico por regras para utilização dentro do sistema de conversão desenvolvido por Egashira. Fora da UNICAMP, podemos destacar o trabalho de Solewicz [63], realizado no Centro de Estudos em Telecomunicações da PUC/RJ (CETUC), bem como alguns trabalhos iniciais desenvolvidos na Universidade Federal da Paraíba [26] [27] e na Escola Politécnica da Universidade de São Paulo. Fora do Brasil vale citar o trabalho de Oliveira [50], desenvolvido para o português europeu.

O presente trabalho procura dar continuidade às atividades anteriormente desenvolvidas na UNICAMP (em especial ao protótipo de conversor texto fala desenvolvido por Egashira [22]), tendo por principal objetivo a implementação de um sistema de síntese de fala a partir de texto genérico, operando para o português do Brasil. Ao longo do restante desta dissertação, estaremos descrevendo com detalhes o trabalho desenvolvido nesse sentido. A discussão doravante apresentada concentrar-se-á nos aspectos práticos, uma vez que as considerações de natureza teórica envolvidas na implementação do sistema já foram apresentadas nos capítulos anteriores.

## 8.2 Especificação geral do sistema

O sistema implementado utiliza o método de síntese concatenativa. Para tanto, ele faz uso de um dicionário de polifones composto por 2450 unidades obtidas a partir da gravação da voz de um locutor do sexo masculino. O sistema pode utilizar duas técnicas de síntese distintas para a geração do sinal de fala: TD-PSOLA e síntese híbrida. O sistema utiliza como entrada um texto genérico em formato ASCII escrito em português e gera como saída um sinal de fala no formato WAV, amostrado a 16kHz e com precisão de 16 bits.

Diferentes versões do sistema de síntese foram construídas durante a elaboração deste trabalho: inicialmente ele foi projetado para trabalhar em ambiente MS-DOS; mais tarde o sistema foi transformado em um aplicativo do tipo EasyWin (emulação de uma janela do MS-DOS dentro do ambiente Windows). A versão atual do sistema de conversão texto-fala é um aplicativo 32 bits para Windows 95. Ele é capaz de trabalhar com múltiplas janelas de texto, sendo que a síntese será sempre efetuada sobre a janela corrente. O aplicativo é dotado de menus para a manipulação dos arquivos-texto de entrada, execução da síntese e da reprodução dos arquivos de saída, bem como opções para a manipulação dos parâmetros de configuração do sistema (essas opções permitem ao usuário escolher a técnica de síntese e o dicionário de polifones utilizados, os arquivos de regras de pré-processamento e de conversão ortográfico-fonética, o nome do arquivo .WAV de saída, dentre outras alternativas). O aplicativo possui ainda uma barra do tipo Media Player, com botões para reprodução, pausa, interrupção, avanço e recuo do arquivo de áudio de saída.

A figura a seguir ilustra a aparência do aplicativo criado:

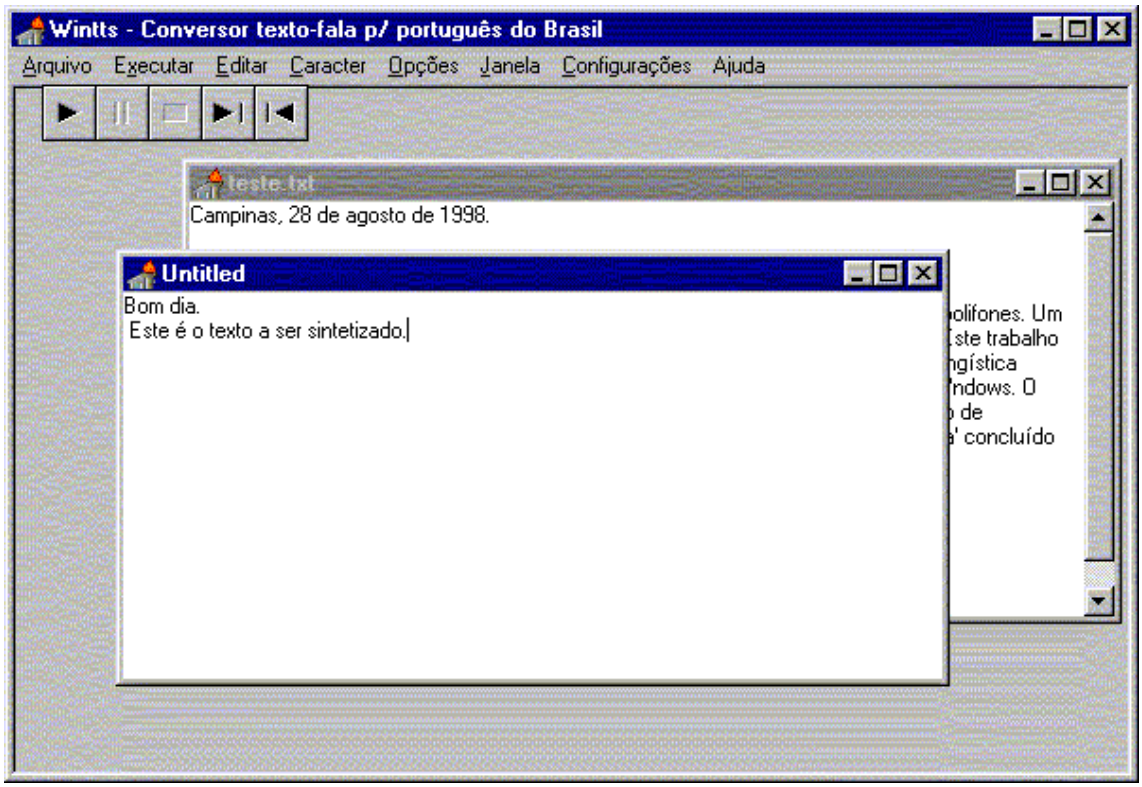



**Figura 8-1 Instantâneo do aplicativo de conversão texto-fala**

O aplicativo desenvolvido funciona adequadamente num Pentium 100 com pelo menos 16 Mb de RAM, equipado com uma placa de som do tipo SoundBlaster e com um disco rígido com aproximadamente 55 Mb livres (essa quantidade de espaço em disco é necessária basicamente para abrigar os dicionários de unidades: 9Mb para o dicionário PSOLA e 45Mb para o dicionário híbrido).

## 8.3 Ferramentas utilizadas na implementação do sistema

Para criar a interface gráfica do aplicativo de conversão texto-fala foi utilizado o Borland Delphi 3.0. A razão da escolha dessa ferramenta foi a sua facilidade de utilização e a

possibilidade que esta oferecia para a criação dos mecanismos de interface apropriados, uma vez que o próprio Delphi possui um ambiente de trabalho que permite a criação de interfaces dentro do padrão do Windows, com janelas, menus, barras de tarefa, objetos de mídia, etc.

O Delphi foi utilizado apenas para implementar a interface do aplicativo com o usuário. Para a elaboração do software propriamente dito foi utilizada a linguagem C++, através da ferramenta Borland C++ 5.01. O C++ foi escolhido por apresentar os recursos de programação apropriados e por ser uma linguagem extremamente flexível (é uma das linguagens de programação mais utilizadas atualmente). Além disso, trata-se de uma linguagem orientada a objeto, o que se encaixava perfeitamente à característica altamente modular do sistema de conversão texto-fala a ser implementado.

O software foi compilado dentro do Borland C++ como uma DLL com funções exportáveis que pudessem ser ativadas dentro do aplicativo desenvolvido a partir do Delphi. Esse tipo de estratégia precisou ser adotado pois o Delphi trabalha com Object Pascal, e a comunicação com módulos desenvolvidos em C++ não é direta.

Para a gravação e segmentação das unidades constituintes do inventário para a síntese concatenativa foi utilizada a ferramenta CSL (Computerized Speech Laboratory), da Kay Elemetrics, modelo 4300B, equipamento este pertencente ao Laboratório de Fonética Acústica e Psicolingüística Experimental (LAFAPE) do Instituto de Estudos da Linguagem (IEL) da UNICAMP. A escolha dessa ferramenta deveu-se basicamente aos recursos por ela disponibilizados (observação da forma de onda e do espectrograma dos sinais de fala, análise LPC, FFT, dentre outras), os quais sem dúvida foram essenciais para a elaboração correta do inventário de unidades.

## 8.4 Estrutura do software de conversão texto-fala

O diagrama de blocos a seguir ilustra a estrutura geral do sistema de conversão texto-fala construído:

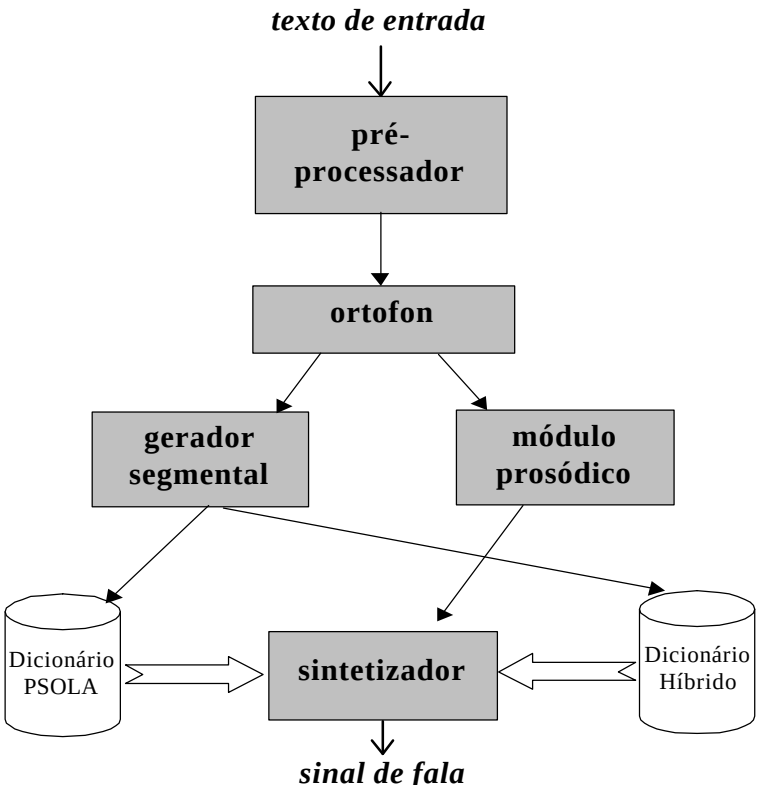



**Figura 8-2 Estrutura geral do software de conversão texto-fala**

Obviamente, essa figura é bastante semelhante àquela apresentada no Capítulo 4 (Figura 4.1). Como o sistema desenvolvido é um sistema de síntese concatenativa, é necessária a existência de pelo menos um dicionário de unidades, a partir do qual serão selecionados os segmentos de fala pré-gravados utilizados no processo de síntese. Essa tarefa de seleção é executada pelo módulo gerador segmental.

No sistema aqui apresentado, dois dicionários estão disponíveis: um para a síntese PSOLA e outro para a síntese híbrida. O dicionário a ser utilizado depende da técnica de síntese selecionada pelo usuário anteriormente à execução do processo de conversão.

Ao longo do restante deste capítulo, estaremos descrevendo com detalhes o funcionamento de cada um dos módulos que compõem a Figura 8-2.

## 8.5 O pré-processador

O pré-processador é o primeiro módulo do sistema que entra em funcionamento ao iniciar-se o processo de conversão texto-fala. Ele utiliza como entrada o texto contido na janela ativa e gera uma saída, também em formato textual, que será utilizada pelo módulo de transcrição ortográfico-fonética.

O funcionamento do módulo pré-processador é baseado na existência de um *compilador de regras*, já presente no protótipo do sistema de conversão texto-fala implementado por Egashira. O compilador de regras permite a especificação das regras de pré-processamento por meio de um arquivo-texto simples, denominado *arquivo de regras.* A vantagem da utilização de um compilador de regras é que tal abordagem permite a inclusão, exclusão e modificação de regras de pré-processamento sem que haja a necessidade de "reprogramar" o sistema. As regras são inteiramente independentes do sistema, por isso pode-se usar arquivos de regras distintos dependendo do tipo de texto com o qual se pretende trabalhar (pois como vimos, o resultado do pré-processamento pode variar de acordo com o contexto).

O compilador de regras é um objeto dentro do programa, cuja função é a de aplicar as transformações contidas no arquivo de regras, de forma a transformar o texto de entrada num texto de saída normalizado. O resultado dessa normalização depende do conteúdo do arquivo de regras. Para que possa ser corretamente interpretado pelo compilador, o arquivo de regras

deve ser escrito segundo uma sintaxe pré-definida. Basicamente, ele deve seguir a estrutura a seguir:

**_GRUPOS_**
*Definição de classes*

**_REGRAS_**
*Definição de regras*

A primeira parte do arquivo corresponde à definição de classes. As classes constituem os elementos básicos do texto que serão manipulados pelas regras. Um elemento pertencente a uma classe corresponde a uma seqüência de um ou mais caracteres que se encaixam na especificação da definição da classe.

Há várias maneiras de definir uma classe. Considere os exemplos:

**Digito[0-9]**
**Digito_não_nulo [1-9]**

Tais definições indicam que qualquer caracter contido entre 0 e 9 pode pertencer à classe *dígito*. Já um *dígito_não_nulo* corresponde a um caracter contido entre 1 e 9. Note que uma determinada seqüência de caracteres pode pertencer, simultaneamente, a mais de uma classe.

Uma classe já definida pode ser utilizada no corpo de definição de uma outra classe. Por exemplo:

**Numero[{Digito}]+**

indica que um número é uma seqüência de um ou mais dígitos.

A segunda parte do arquivo de regras corresponde à definição das regras propriamente ditas. A estrutura geral de uma regra é dada a seguir:

**<*contexto_esquerdo*> objeto_de_análise <*contexto_direito*> := <objeto resultante>**

Os elementos das regras são constituídos por uma seqüência de caracteres, sendo que as classes definidas anteriormente também podem ser utilizadas nas definições das regras. Os contextos esquerdo e direito são opcionais na definição de uma regra e podem ser omitidos.

O lado esquerdo da regra (localizado antes do operador ":=") constitui o padrão a ser encontrado no texto de entrada. Sempre que ocorre um casamento entre uma seqüência de caracteres da entrada e o lado esquerdo de uma regra, esta regra entrará em ação. A atuação da regra é bastante simples: ela apenas substitui a seqüência de caracteres correspondente ao objeto de análise por uma outra seqüência (que pode inclusive ser nula) correspondente ao objeto resultante. Considere os exemplos a seguir:

**(S|s)r. := senhor;**
**2 < {digito_não_nulo} > := ("vinte e");**

A primeira regra indica que as seqüências de caracteres "*Sr.*" e "*sr.*" devem ser transformadas na seqüência "*senhor*". A segunda regra, por sua vez, indica que o número *2*, quando sucedido por um e apenas um dígito não nulo, deve ser substituído por "*vinte e*".

A ordem de execução das regras afeta o resultado final do pré-processamento, e depende da ordem em que estas estão especificadas dentro do arquivo de regras. À medida em que elas são executadas, os elementos do texto de entrada vão sendo consumidos progressivamente, de forma que, ao final da execução de todas as regras, tem-se o texto normalizado.

O arquivo de regras de pré-processamento utilizado pelo sistema de conversão texto-fala apresentado neste trabalho encontra-se no Apêndice I. As regras nele definidas são capazes de processar os seguintes elementos do texto:

**números:** os números cardinais menores que 1 quintilhão são tratados adequadamente pelo módulo de pré-processamento (esse teto de 1 quintilhão é mais do que suficiente, pois raramente números tão grandes aparecem num texto sob a forma cardinal: normalmente eles já aparecem por extenso). O módulo de pré-processamento assume que, na representação cardinal, os milhares aparecem separados por um ponto (Ex.: 1.999). Números decimais também são tratados adequadamente (Ex.: 14,25 é lido como *quatorze vírgula vinte e cinco*). Outras formas numéricas, como datas, horas, números de telefone, números ordinais, etc., não são tratadas atualmente pelo pré-processador. Vale aqui citar a colaboração do trabalho de iniciação científica de Anthony Gerarde Alves Stanton, aluno de graduação em Engenharia de Computação, na definição das regras de processamento de números [66].

**abreviaturas:** o módulo de pré-processamento é capaz de identificar quando o caracter *ponto* (.) pertence a uma abreviatura, ao invés de constituir um indicador de final de sentença. Portanto, é possível fazer o tratamento de abreviaturas terminadas ou não em ponto. Por hora, não é possível resolver o caso de abreviaturas que possuam mais de uma forma de expansão possível, pois nesse caso é necessário efetuar uma análise semântica mais complexa. Durante a confecção do arquivo de regras, deve-se escolher a forma de expansão mais comum, que será executada toda vez que a referida abreviatura for encontrada.

**siglas:** palavras escritas em letras maiúsculas são interpretadas como siglas e, como tal, são soletradas. Caso se queira fazer o tratamento de siglas não soletráveis, deve-se escrever uma regra específica para ela.

**Ex.: PUCCAMP := (pucâmpi)**

**símbolos especiais:** semelhantemente ao caso das abreviaturas, só é possível efetuar o tratamento dos símbolos especiais que possuem uma única forma de expansão possível. Nos outros casos, deve-se selecionar uma regra que realize a expansão mais usual.

**maiúsculas:** as letras maiúsculas que não correspondem a siglas (como as que ocorrem em início de sentença ou em nomes próprios) são convertidas em letras minúsculas, por uma exigência do módulo de transcrição ortográfico-fonética.

## 8.6 O *Ortofon*

Após passar pelo pré-processador, o texto de entrada encontra-se em formato normalizado, ou seja, é composto somente por letras minúsculas e por sinais de pontuação. Esse texto normalizado dá entrada ao módulo seguinte do sistema de conversão, denominado *ortofon.*

A função do *ortofon* é reescrever a seqüência ortográfica de entrada utilizando uma notação que represente a seqüência fonética correspondente à sentença a ser sintetizada (o nome *ortofon* indica justamente essa mudança de notação, de uma representação ortográfica para uma representação fonética).

Uma das diretrizes que orientou a construção do sistema de conversão texto-fala aqui apresentado foi a incorporação de modelos lingüísticos realistas, tendo por objetivo garantir a alta qualidade do sinal sintetizado. O *ortofon* constitui um dos pontos básicos do processamento lingüístico efetuado pelo sistema. Ele foi construído pela equipe do LAFAPE e, muito embora já esteja incorporado ao sistema de conversão, possui alguns detalhes a respeito de seu funcionamento que ainda estão em fase de elaboração.

O ponto de partida para a construção de um módulo de transcrição fonética consiste em determinar a notação a ser empregada, ou seja, o conjunto dos símbolos utilizados para

representar os fonemas que podem ocorrer em uma sentença. Obviamente, o conjunto de símbolos deve ser suficiente para representar os diversos sons produzidos na língua. Mais do que isso, no entanto, é preciso decidir até que ponto os fenômenos de alofonia e coarticulação estarão representados dentro da própria notação.

Numa notação fônica *estreita* os detalhes fonéticos de cada segmento são totalmente explicitados, o que significa dizer, por exemplo, que é necessário utilizar vários símbolos para representar o mesmo fonema em contextos fonéticos distintos. Ao se utilizar uma notação fônica mais *abstrata,* no entanto, muitos desses detalhes estarão subespecificados.

A estratégia utilizada na notação empregada pelo *ortofon* não segue à risca nenhuma das duas abordagens. Ela supõe a existência de duas classes de fonemas distintas: *plenos* e *reduzidos.*

Os segmentos plenos são aqueles que ocorrem em ambientes prosódicos fortes, e são menos sujeitos à coarticulação. Dentre eles, podemos destacar as vogais tônicas e pré-tônicas, bem como as consoantes de início de sílaba.

Os segmentos fracos ou reduzidos, por sua vez, ocorrem em ambientes prosódicos mais fracos. Tais segmentos possuem, em geral, duração menor que a dos segmentos fortes e são mais suscetíveis à coarticulação com segmentos subjacentes, portanto apresentam maior variabilidade de acordo com o contexto em que estão inseridos. Como exemplos de segmentos reduzidos podemos citar as vogais pós-tônicas e semivogais, as consoantes de final de sílaba e as líquidas de encontros consonantais, como em *“prato”* e *“placa”* [6].

A notação empregada pelo *ortofon* utiliza letras minúsculas para representar os segmentos plenos e letras maiúsculas para representar os reduzidos. Considere o exemplo abaixo:

*casa -> kazA*

A primeira vogal (“*a*”) é uma vogal tônica, e portanto constitui um segmento pleno. Já a segunda (“*A*”) é uma vogal pós-tônica, e portanto corresponde a um segmento reduzido. No exemplo, as duas consoantes são plenas, por isso são representadas por letras minúsculas.

O conjunto dos símbolos utilizados encontra-se no Apêndice II. Uma discussão mais detalhada a respeito dos critérios adotados na elaboração dessa notação pode ser encontrada em [1].

Uma vantagem adicional dessa notação é que ela já traz implícita a posição da vogal tônica: ela sempre corresponde à última vogal minúscula da palavra transcrita. Os exemplos a seguir ilustram essa característica:

*prático -> pRatIkO*

*pratico -> pRatikO*

O módulo de transcrição ortográfico-fonética (*ortofon*) pode ser dividido em duas partes distintas: um aplicador de regras de transcrição e um dicionário de exceções.

As regras de transcrição tratam das correspondências regulares entre letras e sons. Como o português é uma língua razoavelmente fonêmica, as regras são capazes de dar conta da transcrição da maioria das palavras.

A maneira ideal de implementar o aplicador de regras de transcrição seria por meio de um compilador de regras, de maneira análoga à que foi utilizada na implementação do módulo pré-processador. Essa estratégia permitiria modificar as regras de transcrição sem a necessidade de reprogramar o sistema, pois estas estariam especificadas dentro do arquivo de regras, que é um arquivo-texto simples.

Como o *ortofon* ainda está em construção, algumas das regras que o compõem ainda estão em fase de preparação pela equipe do LAFAPE. Por serem exatamente as regras mais trabalhosas, surgiu a dúvida sobre se o compilador de regras atualmente utilizado pelo sistema

seria suficiente para dar conta de tais regras, visto que o compilador emprega uma gramática relativamente simples. Dependendo da complexidade das regras faltantes, poderemos chegar à conclusão de que há necessidade de reformular também o funcionamento do compilador de regras atual. Por esse motivo preferiu-se, por ora, implementar o *ortofon* como um objeto integrante do sistema, ou seja, as regras de transcrição foram implementadas como código de programa. Muito embora perca-se em flexibilidade, essa estratégia nos pareceu mais adequada, pois permitiu testar de forma rápida os resultados gerados pelo *ortofon*. Como o módulo foi escrito em linguagem C++, este pôde ser facilmente incorporado ao restante do sistema.

Obviamente, trata-se apenas de uma estratégia provisória pois, assim que as regras de transcrição forem totalmente especificadas, poder-se-á tomar uma decisão final a respeito da necessidade ou não de reformular o funcionamento do compilador de regras. Uma vez tomada essa decisão, o compilador poderá ser também utilizado na aplicação das regras de transcrição.

O segundo componente do módulo *ortofon* é o dicionário de exceções. Atualmente, ele é composto pelos 1383 verbetes do minidicionário Aurélio [25] para os quais as regras de transcrição ainda continuam falhando, mesmo após sucessivos refinamentos. O dicionário de exceções entra em funcionamento *antes* do aplicador de regras de transcrição, ou seja, as regras são aplicadas apenas naquelas palavras que não constam do dicionário.

Após passar pelo dicionário de exceções e pelo aplicador de regras de transcrição, a seqüência ortográfica de entrada é transformada na seqüência fonética equivalente. Considere, por exemplo, a sentença a seguir, já devidamente normalizada pelo módulo de pré-processamento:

*"joão romão foi, dos treze aos vinte e cinco anos, empregado de um vendeiro que enriqueceu entre as quatro paredes de uma suja e obscura taverna nos refolhos do bairro do botafogo; e tanto economizou do pouco que ganhara nessa dúzia de anos, que, ao retirar-se o patrão para a terra, lhe deixou, em pagamento de ordenados vencidos, nem só a venda com o*

*que estava dentro, como ainda um conto e quinhentos em dinheiro"*. (Aluízio Azevedo, *O Cortiço*).

Após passar pelo *ortofon*, a seqüência ortográfica acima seria transcrita da seguinte forma:

*zhoaNU romaNU foI, doS tRezE aOS viNtE e siNkO anOS, eNpRegadO de uN veNdeIRO ke eNrikeseU eNtRE aS kUatRO paRedES de umA suzhA e obSkuRA tavehRnA noS refolhOS do baIrO do bohtafogO, e taNtO ekonomizoU do poUkO ke gaNnhaRA nehsA duzIA de anOS, ke, aU retiRaR sE o patRaNU paRA a tehrA, lhe deIshoU, eN pagameNtO de oRdenadOS veNsidOS, neN soh a veNdA koN o ke eStavA deNtRO, komO aiNdA uN koNtO e kinheNtOS eN dinheIRO.*

Obviamente, as regras de transcrição e o dicionário de exceções não são suficientes para dar conta da transcrição correta de todas as palavras do texto. A existência de um *parser* é essencial para resolver alguns problemas de transcrição, como o já citado caso dos vocábulos homógrafos heterofônics (ex.: verbo *piloto* X substantivo *piloto*).

O sistema atual ainda não contém um *parser*; no entanto algum trabalho nesse sentido já vem sendo realizado pela equipe do LAFAPE. Um grupo de pesquisadores está trabalhando na construção de um módulo de análise morfossintática: a construção de tal módulo baseia-se na existência de um léxico digitado manualmente, contendo os 27.078 verbetes do minidicionário Aurélio. As flexões dos verbos regulares foram calculadas de forma automática e também fazem parte do léxico; já as flexões dos verbos irregulares foram inseridas manualmente.

O *ortofon* trata apenas dos aspectos relacionados à fonologia *lexical*, ou seja, ele faz a transcrição de cada palavra isoladamente. Muitas vezes, no entanto, as palavras podem interagir entre si dentro do texto, e ter algumas de suas características fonéticas alteradas em função das palavras que lhe são adjacentes.

O fone [**s**] em final de palavra, por exemplo, é normalmente uma fricativa não-sonora; no entanto ele pode se sonorizar quando o primeiro fone da palavra seguinte for sonoro, como em *"velhos marinheiros"*. O fenômeno de *sândhi externo*, no qual a última vogal de uma palavra se funde com a primeira vogal da palavra seguinte (ex.: *"casa azul"*), também constitui um caso de ajuste pós-lexical. Há ainda casos em que o fone final de uma palavra pode se modificar ou inclusive desaparecer totalmente, devido à coarticulação com o fone inicial da próxima palavra, como ocorre em *"se o"*, que se pronuncia *"si o"*, e em *"cara idoso"*, que se pronuncia *"caridoso"*.

A fonologia *pós-lexical* trata desses ajustes que ocorrem nas palavras quando elas estão inseridas em um contexto mais amplo. É preciso, portanto, que o sistema de conversão contenha um módulo, posterior ao *ortofon*, capaz de lidar com esses ajustes.

A versão atual do *ortofon* não faz nenhum tipo de tratamento pós-lexical. No entanto, já existe também um trabalho sendo realizado junto à equipe do LAFAPE no sentido de implementar um novo módulo, que aqui chamaremos de *processador métrico*, e que será brevemente incorporado ao sistema de conversão texto-fala. O objetivo desse novo módulo é exatamente o de prover um tratamento pós-lexical à saída do *ortofon*.

No caso dos exemplos de ajuste pós-lexical anteriormente citados, a atuação do processador métrico manifestar-se-ia através da alteração da saída gerada pelo *ortofon* da seguinte forma:

*velhos marinheiros -> vehlhOS maRinheIROS -> vehlhOz maRinheIROS*

*casa azul -> kazA azuL -> kazazuL*

*se o -> sE o -> sIO*

*cara idoso -> caRA idozO -> kaRidozO*

## 8.7 O módulo de processamento prosódico

Terminada a fase de processamento lingüístico, a etapa seguinte a ser executada pelo sistema de conversão texto-fala diz respeito ao processamento prosódico. Durante essa etapa são determinados valores de duração e curvas de freqüência fundamental para cada um dos segmentos fonéticos determinados durante a etapa de transcrição fonética.

O modelo aqui empregado adota uma reta simples para representar a curva de F0 de um segmento fonético, ou seja, cada segmento possui três parâmetros prosódicos a ele associados: duração (expressa em milissegundos), *F0* inicial e *F0* final (expressos em Hz). O valor de F0 final de um segmento fonético deve ser sempre igual ao valor de F0 inicial do segmento fonético seguinte, de forma que a curva de F0 associada à sentença como um todo, constituída pela junção dos segmentos de reta, seja contínua. A amplitude do sinal não foi considerada, por hora, como parâmetro prosódico relevante, por isso ela não é tratada no seu aspecto frasal dentro do módulo prosódico.

Vale aqui ressaltar as diferenças entre o que, neste trabalho, chamamos respectivamente de *fone* e *segmento fonético*. Fones dizem respeito a unidades básicas de fala: no caso da notação empregada pelo *ortofon*, os fones correspondem aos 34 símbolos que se encontram no Apêndice II. Segmentos fonéticos, por sua vez, dizem respeito às unidades às quais são associados os parâmetros prosódicos (duração, F0 inicial e F0 final), e que não correspondem necessariamente aos fones.

Um segmento fonético pode se composto por um *ou mais* fones. Como exemplos de segmentos fonéticos compostos podemos citar os ditongos (ex.: *aI, aU, eI, eU,* etc.) e os tritongos (ex.: *UaI, UaU,* etc.), que são encontros vocálicos formados, respectivamente, por dois e três fones distintos. Cada ditongo e cada tritongo é tratado como um segmento fonético único. Isso porque, durante a realização de um encontro vocálico desse tipo, é difícil determinar o ponto exato onde ocorre a transição entre uma vogal e a seguinte. Faz muito

mais sentido encarar o encontro vocálico como um segmento único com padrões formânticos que variam ao longo do tempo e atribuir parâmetros prosódicos ao segmento como um todo.

As vogais nasalizadas também constituem segmentos fonéticos individuais, formados pela junção de uma vogal oral com o arquifonema **"N"** (ex.: *aN, eN, iN, oN, uN, AN, EN, IN, ON, UN*). Na verdade o **"N"** funciona, em nossa notação, muito mais como um traço nasal do que como um fonema propriamente dito, por isso ele jamais ocorre como um segmento fonético isolado.

Alguns segmentos fonéticos nasais são constituídos por três fones distintos: Veja os exemplos abaixo:

*pão -> p**aNU***

*põe -> p**oNI***

*mãe -> m**aNI***

*sótão -> soht**ANU***

Nesses casos o **"N"** aparece entre a vogal e a semivogal, e não após a semivogal. Isso porque ele funciona como um traço nasal cujo raio de atuação atinge sobretudo a vogal precedente, fenômeno este que pode ser verificado a partir da observação de espectrogramas correspondentes a tais segmentos fonéticos.

Para efetuar a criação do dicionário de unidades para concatenação foram gravadas diversas frases-veículo contendo todas as unidades a serem incorporadas ao dicionário (esse trabalho será melhor descrito na seção seguinte). O início e o fim de cada um dos segmentos fonéticos constituintes das unidades contidas nas frases-veículo foram marcados, de forma que assim pudemos contar com uma amostra grande de instâncias de segmentos fonéticos extraídos de fala natural (vale aqui ressaltar que estamos nos referindo a segmentos fonéticos inteiros contidos nas frases-veículo, e não aos pedaços de segmentos contidos nos polifones).

Esse conjunto de segmentos fonéticos foi utilizado como espaço amostral para calcular a duração média e o desvio-padrão de cada um dos segmentos fonéticos da nossa língua (obviamente os valores de duração média e os desvios são uma característica própria do locutor que emprestou a voz para a gravação das unidades).

O conjunto completo dos segmento fonéticos, com as respectivas durações médias e os desvios-padrão, encontra-se no Apêndice III.

O papel do módulo de processamento prosódico dentro do sistema de conversão texto-fala consiste em determinar uma seqüência de segmentos fonéticos a partir da seqüência de fones obtida pelo módulo de transcrição e determinar os três parâmetros prosódicos (duração, F0 inicial e F0 final) de cada um desses segmentos fonéticos.

A versão atual de nosso sistema de conversão texto-fala ainda não possui um algoritmo capaz de calcular tais parâmetros prosódicos de maneira automática. Um modelo de duração desenvolvido por Barbosa [11] já apresenta resultados satisfatórios e deverá ser, muito em breve, incorporado ao sistema. Tal modelo de duração deverá se unir a um modelo entoacional (ainda a ser construído) para compor o módulo de processamento prosódico de nosso sistema de conversão texto-fala.

A fim de suprir a ausência momentânea do módulo de processamento prosódico, o sistema foi construído para aceitar não apenas entrada no formato textual, mas também um outro tipo de entrada, que aqui chamaremos de *modo de entrada prosódica*, na qual os parâmetros prosódicos associados a cada segmento fonético já aparecem especificados. Essa entrada, gerada manualmente, é equivalente à saída que seria produzida pelo módulo de processamento prosódico, e dá entrada diretamente ao módulo de síntese.

O exemplo a seguir ilustra uma entrada prosódica correspondente à seqüência ortográfica *"Bom dia"*.

0: / [ 120 ] 333 [ 120 ]

1: b [ 95 ] 97 [ 89]

2: oN [ 89 ] 163 [ 79 ]

3: d [ 79 ] 90 [ 77 ]

4: i [ 77 ] 149 [ 74 ]

5: A [ 74 ] 107 [ 52 ]

6: / [ 120 ] 330 [ 120 ]

Os valores entre colchetes correspondem aos valores de F0 inicial e F0 final do segmento fonético, em Hz, nessa ordem; o outro valor corresponde à duração, em milissegundos, do segmento fonético. O valor de 120 Hz p/ o segmento de silêncio (“/”) no início e no final da seqüência fonética é arbitrário, pois na verdade não faz sentido falar em valor de F0 no caso de um segmento não-sonoro.

## 8.8 Criação do inventário de unidades

Antes de entrarmos na descrição do módulo seguinte do sistema de conversão texto-fala, responsável pela síntese do sinal propriamente dita, estaremos discutindo em detalhes, nesta seção, o trabalho de criação do inventário de unidades para concatenação. Trata-se de um dos pontos fundamentais deste trabalho, pois os critérios lingüísticos adotados na determinação da estrutura das unidades componentes do inventário, bem como as estratégias seguidas durante as etapas de gravação e de segmentação das unidades, constituem fatores de influência preponderante na qualidade da fala sintetizada.

### 8.8.1 Critérios utilizados na elaboração do conjunto de unidades

O sistema de conversão texto-fala aqui apresentado baseia-se no método de síntese concatenativa, ou seja, o sinal de fala sintetizada é construído a partir da junção e da modificação dos parâmetros prosódicos de segmentos de fala previamente gravados. Tais segmentos de fala encontram-se em um banco de dados, que aqui denominaremos de *inventário de unidades* ou *dicionário de unidades.* Para que seja possível realizar a conversão texto-fala de um texto genérico, é preciso que o inventário contenha unidades suficientes que permitam gerar qualquer combinação de sons existente na língua.

Conforme vimos no Capítulo 7, a utilização de *fones* como unidades básicas para concatenação, muito embora seja a alternativa mais intuitiva, não é capaz de gerar resultados satisfatórios. Vimos que a alternativa mais razoável consiste em utilizar *polifones* como unidades básicas para concatenação. Os polifones são segmentos de fala constituídos por dois ou mais fones em seqüência. O início e o fim de um polifone ocorrem sempre na porção espectralmente estável dos fones inicial e final, de forma que as transições entre fones estão sempre inteiramente contidas no interior da unidade. Por isso, grande parte dos fenômenos de coarticulação entre fones estão contidos no interior da unidade.

O sistema de conversão texto-fala aqui apresentado utiliza-se de polifones como unidades básicas para concatenação. São utilizadas tanto unidades demissilábicas como também intersilábicas. Os critérios utilizados na criação dessas unidades foram elaborados no LAFAPE e encontram-se descritos com mais detalhes em [2].

Um desses critérios foi o de não efetuar cortes no interior de encontros vocálicos, ou seja, mantê-los intactos dentro das unidades. Isso porque as vogais dentro do encontro vocálico são extremamente coarticuladas entre si. Além disso, normalmente não existe uma região espectralmente estável no interior de um encontro vocálico que permita efetuar o corte de forma satisfatória.

Outro critério adotado foi o de diferenciar os segmentos plenos dos reduzidos. Isso está em acordo com a notação fônica empregada na etapa de transcrição ortográfico-fonética. Conforme já vimos anteriormente, os segmentos plenos correspondem às vogais tônicas e pré-tônicas, bem como às consoantes de *onset* silábico (início de sílaba). Os segmentos reduzidos, por sua vez, correspondem às vogais pós-tônicas, às semivogais, às consoantes de coda silábica (final de sílaba) e as líquidas de *onset* silábico (como em *p**R**a, p**R**e, p**R**i, p**L**a, p**L**e, p**L**i,* etc.).

Os segmentos reduzidos são aqueles que ocorrem em ambientes prosódicos mais fracos. Por esse motivo eles têm, em média, duração menor que os segmentos plenos, raramente apresentando uma região espectralmente estável. Além disso, eles são extremamente suscetíveis à coarticulação com os segmentos subjacentes.

Muito embora a natureza das vogais tônicas e pré-tônicas não seja inteiramente similar [6], optou-se por não diferenciá-las dentro das unidades. Os segmentos reduzidos, por sua vez, possuem características acústicas bastante distintas das dos seu equivalentes plenos, o que levou à necessidade de utilizar unidades específicas que contivessem esse tipo de segmento.

Um critério básico na elaboração das unidades constituintes do inventário seria, portanto, não efetuar jamais o corte no interior dos segmentos reduzidos, ou seja, tais segmentos devem estar, em princípio, inteiramente contidos no interior das unidades. Com isso, minimiza-se a ocorrência de descontinuidades espectrais nas junções das unidades durante o processo de concatenação; além disso, boa parte dos fenômenos coarticulatórios que se manifestam predominantemente nos segmentos reduzidos estariam contidos nas próprias unidades.

A adoção à risca de tal critério, bem como a inclusão de vogais pré-tônicas distintas das tônicas, nos leva a um problema de natureza prática: o número de unidades necessário para compor o inventário seria da ordem de 20.000. Comparando-se este número com o número de

unidades de um dicionário composto exclusivamente por difones, que é da ordem de 1.000, percebe-se uma diferença significativa.

Tendo isso em vista, tornou-se necessária a adoção de alguns critérios que levassem à diminuição do número total de unidades constituintes do inventário. Procurou-se encontrar algumas condições dentro das quais fosse possível efetuar cortes no interior dos segmentos reduzidos, sem no entanto degradar de forma significativa a qualidade final da fala sintetizada. A seguir, estaremos descrevendo algumas das estratégias adotadas:

Em primeiro lugar, as vogais pós-tônicas em posição de núcleo silábico são segmentadas. Nesse caso são utilizadas unidades demissilábicas. O corte é efetuado no final da transição com a consoante precedente, pois dessa forma as descontinuidades espectrais oriundas do processo de concatenação se tornam menos perceptíveis. Considere o exemplo abaixo:

*ótimo -> /ohtimo/ -> /oh + oht + **tI** + **Im** + mO + O/*

Como o corte é efetuado logo após a transição da vogal pós-tônica com a consoante precedente, ocorre que a vogal estará contida quase que inteiramente na unidade seguinte. No exemplo acima, apenas a transição entre o ***"I"*** e o ***"t"*** está contida no polifone ***"tI".*** Toda a porção estável da vogal se encontra no interior da unidade ***"Im"***.

Outra estratégia adotada foi a de permitir que vogais nasais e ditongos nasais fossem concatenados com *onsets* (inícios) orais. Com isso, evitou-se a necessidade de criar unidades específicas contendo os *onsets* nasais. Muito embora tenha-se contrariado o princípio de que a junção das unidades só pode ocorrer entre segmentos de natureza equivalente, não houve degradações significativas no resultado da concatenação. Isso porque as vogais nasalisadas, em português brasileiro, iniciam-se com uma fase oral [64]. A minimização de distorção foi conseguida por meio da utilização de unidades demissilábicas, semelhantes àquelas já descritas para o caso das vogais pós-tônicas em posição de núcleo silábico. O corte na vogal de *onset* oral foi efetuado logo após a transição desta com a consoante precedente. Portanto

tais unidades podem ser concatenadas tanto com unidades contendo vogais e ditongos orais como com aquelas contendo vogais e ditongos nasais, pois a característica oral ou nasal da rima silábica está quase que inteiramente contida na unidade seguinte. Os exemplos abaixo ilustram o tipo de concatenação utilizado:

*bomba -> /boNbA/ -> /b + **bo** + **oNb** + bA + A/*

*põe -> /poNI/ -> /p + **po** + **oNI/***

Nos exemplos acima, a junção ocorre entre o segmento oral ***"o"*** e os segmentos nasais ***"oN"*** e ***"oNI"***, respectivamente.

A concatenação de *onsets* orais com rimas nasais funciona bem para quase todos os casos. A aplicação desse procedimento nos segmentos ***aN***, ***aNU*** e ***aNI***, no entanto, produz resultados pobres. Isso porque o segmento oral ***"a"*** apresenta características espectrais muito diferentes das do início oral de ***"aN"***, ***"aNU "***e ***"aNI".*** A solução adotada nesse caso foi a de efetuar a concatenação das rimas nasais não com a vogal tônica ***"a"***, mas sim com a pós-tônica ***"A"***, que apresenta um padrão formântico mais próximo ao do início oral desses segmentos nasais. O exemplo abaixo ilustra a estratégia adotada:

*mão -> /maNU/ -> /m + **mA** + **aNU/***

Ainda na tentativa de procurar minimizar o tamanho do inventário de unidades, optou-se por segmentar ***"S"***, ***"R"*** e ***"L"*** quando estes ocorrem no início de encontros consonantais. Alguns testes auditivos foram feitos, os quais demonstraram que o corte no interior desses segmentos não ocasionava uma degradação significativa na qualidade do sinal sintetizado. Veja nos exemplos a seguir os tipos de unidades que são utilizadas na formação de tais encontros consonantais:

*pasta -> /paStA/ -> /p + pa + **aS** + **St** + tA/*

*porta -> /pohRtA/ -> /p + poh + **ohR** + **Rt** + tA/*

Uma outra característica específica do português do Brasil foi explorada a fim de economizar mais algumas unidades do inventário. No caso, o segmento ***"L"*** em posição de coda silábica normalmente é pronunciado de forma idêntica à semivogal ***"U"***, como por exemplo em palavras como *"mal"* e *"possível"*. Nesse caso, ao invés de utilizar unidades específicas contendo o ***"L"*** em posição de coda silábica, são utilizadas as unidades contendo a semivogal. O exemplo a seguir mostra essa adaptação:

*calma -> /kaLmA/ -> /kaUmA/ -> /k + ka + aUm + mA + A/*

### 8.8.2 Gravação das unidades

Uma vez definidos os critérios a serem seguidos na determinação do conjunto de unidades constituintes do inventário, passou-se para a parte prática do trabalho, ou seja, a gravação das unidades propriamente dita.

O primeiro passo nesse processo consistiu em selecionar um locutor responsável por "emprestar" a sua voz durante o processo de gravação. Optou-se pela utilização de um locutor masculino, principalmente devido à maior facilidade em efetuar análises de natureza espectral numa voz com padrão de F0 mais grave.

Procurou-se seguir alguns critérios na seleção do locutor mais apropriado. Em primeiro lugar, tal pessoa deveria possuir uma dicção bastante clara, a fim de garantir a articulação correta de cada um dos fones constituintes das unidades durante o processo de leitura. As características da voz também foram importantes no processo de escolha, pois havia preferência por uma voz limpa, sem a presença de fenômenos como rouquidão, aspiração ou diplofonia, os quais dificultariam sensivelmente a análise espectral. Além disso o locutor deveria possuir um bom controle vocal, que lhe permitisse manter o mesmo padrão de leitura (entonação, taxa de elocução) ao longo de todo o extenuante processo de gravação.

O então aluno de doutorado Zaldo Rocha Filho, do Instituto de Estudos da Linguagem, apresentava os requisitos acima descritos e por isso foi escolhido como locutor.

Uma vez selecionada a voz a ser utilizada partiu-se para o passo seguinte, que foi a escolha de um *contexto fonético-prosódico* adequado para a gravação das unidades. Gravá-las simplesmente efetuando a leitura isolada das unidades, uma a uma, não constitui uma boa estratégia. O ideal, nesse caso, é que as unidades ocorram dentro de um ambiente o mais neutro possível, tanto do ponto de vista fonético como do ponto de vista prosódico.

Para garantir um ambiente prosódico neutro é importante que, durante a gravação, as unidades ocorram no interior de um enunciado maior. A neutralidade fonética, por sua vez, é garantida por meio da escolha apropriada do contexto fonético adjacente à unidade em questão dentro do enunciado. Deve-se escolher um ambiente fonético que se coarticule de maneira mínima com os fones da unidade, pois esta, durante o processo de concatenação, deverá ser usada em um contexto fonético distinto daquele do qual foi extraída.

Para a gravação da maioria das unidades foram utilizadas palavras sem sentido (*logatomas).* Na formação desses logatomas optou-se pela utilização de consoantes bilabiais (***p*** e ***b***) e da vogal central (***a***). A opção por esses segmentos foi com o intuito de tentar minimizar a coarticulação com os segmentos constituintes da unidade. Para gerar a unidade ***"teh"***, por exemplo, utilizou-se o logatoma *patéba;* para gerar ***"it"*** usou-se *pita,* e assim por diante.

Em alguns casos, no entanto, optou-se pela utilização de palavras verdadeiras ao invés de logatomas. Um exemplo de situação em que isso se mostra necessário é o caso de certas unidades que nunca ocorrem no interior de uma palavra, mas sim em fronteiras de palavras. Encontros vocálicos do tipo *vogal pós-tônica + vogal tônica*, por exemplo, não podem nunca ocorrer dentro de uma palavra, por isso a criação de um logatoma com essa estrutura seria anti-natural, e faria com que a unidade em questão fosse pronunciada de forma hiperarticulada. Em casos como esse, torna-se mais interessante a utilização de duas palavras que contenham, entre elas, a unidade em questão. Para gerar a unidade ***"A#o***", por exemplo, pode-se usar a

combinação de palavras *"bola oca"*, de transcrição fonética *"bol**A#o**kA"* (nesse caso, o símbolo **#** indica uma fronteira de palavra).

Os logatomas não foram gravados isoladamente. Ao invés disso, foram inseridos no interior de *frases-veículo*. O objetivo da utilização das frases-veículo foi o de evitar a leitura dos logatomas com *entonação de lista*, e também evitar o fenômeno de alongamento final, que é mais forte na palavra final do enunciado. A ocorrência desses fenômenos comprometeria a naturalidade da fala. A escolha das frases-veículo também procurou seguir o critério de propiciar um ambiente fonético-prosódico tão neutro quanto possível. As seguintes frases-veículo foram utilizadas:

**Digo <*logatoma*> baixinho.**

**Baixinho digo <*.logatoma*>.**

**<*Logatoma*> digo baixinho.**

A segunda e a terceira estruturas de frases-veículo foram utilizadas para gerar unidades de final e início de enunciado, respectivamente. Maiores detalhes a respeito dos critérios lingüísticos utilizados na seleção dos logatomas e das frases-veículo podem ser encontrados em [6].

Cada uma das frases-veículo foi impressa num cartão e apresentada ao locutor para leitura. Todo o processo de gravação foi executado no Laboratório de Fonética Acústica e Psicolingüística Experimental (LAFAPE). Para garantir a qualidade do sinal gravado foi utilizada uma cabine com isolamento acústico, e o resultado final da gravação foi armazenado em fitas DAT.

Após a gravação, cada uma das frases obtida foi cuidadosamente conferida a fim de assegurar a sua correção. Em seguida, elas foram digitalizadas e armazenadas em disco rígido. Como durante o processo de leitura houve grande variação de intensidade entre as frases e, por

conseguinte, entre as unidades, utilizou-se um processo de *escalonamento* nas formas de onda digitalizadas, a fim de prover a normalização das amplitudes.

O Apêndice IV apresenta a descrição das unidades que constituem o inventário aqui descrito.

### 8.8.3 Geração dos dicionários híbrido e PSOLA

A gravação das frases-veículo constituiu apenas o primeiro passo na elaboração do inventário de unidades. Fazia-se ainda necessário isolar as unidades contidas nas frases-veículo e fazer a montagem do dicionário propriamente dita.

As duas técnicas de síntese utilizadas (PSOLA e híbrida) possuem a característica de serem síncronas com o período de *pitch*. Portanto, era necessário que as frases gravadas fossem submetidas a um processo de *marcação de pitch* [59]. Esse processo foi feito de forma automática por meio de um programa de computador desenvolvido por Violaro em nosso laboratório. O algoritmo de marcação de *pitch* calcula uma seqüência de marcas espaçadas pelo período de *pitch* e posicionadas nos picos do sinal nas porções sonoras, e espaçadas de 10 ms entre si nas porções não sonoras.

O passo seguinte à marcação de *pitch* é o da *segmentação* das unidades. O processo de segmentação consiste das seguintes tarefas:

- localização das marcas de transição, ou seja, das fronteiras entre os segmentos internos às unidades. Dessa forma, um difone deve conter uma marca de transição, um trifone deve conter duas, e assim por diante.

- localização dos pontos de corte das unidades. Os pontos de corte correspondem às posições de início e de final do polifone, ou seja, os pontos onde acontece a junção com outra unidade no processo de concatenação.

Normalmente o corte é efetuado na porção central (mais estável) do segmento. No caso das unidades demissilábicas, no entanto, já vimos que o corte é efetuado de maneira diferente: logo após a transição da vogal com a consoante precedente. As vogais plenas são segmentadas no ponto correspondente ao seu quinto período, e as vogais reduzidas no ponto correspondente ao terceiro período (para este locutor, na taxa de elocução que foi usada).

O processo de segmentação das unidades constituintes do inventário foi feito de forma inteiramente manual pela equipe do LAFAPE (Patrícia Aquino e Plínio Barbosa). Muito embora existam alguns algoritmos visando a efetuar a segmentação de forma automática [61], o resultado nunca é exato e precisa sempre passar por um processo de ajuste manual posterior. Normalmente não existe um critério para determinar o ponto exato onde ocorre a transição entre dois segmentos.

A segmentação manual das unidades é uma tarefa extremamente trabalhosa, porém o resultado produzido é mais confiável. Tal tarefa foi executada a partir da análise tanto dos espectrogramas como das formas de onda dos sinais gravados, com o auxílio da ferramenta CSL (Computerized Speech Laboratory), do LAFAPE. Posteriormente à segmentação manual, as unidades passaram por um processo de verificação e correção também manuais. Este trabalho, bem como todos as etapas posteriores da criação do inventário, foram realizadas no LPDF (Laboratório de Processamento Digital de Fala).

Além de localizar as marcas de transição e os pontos de corte, foi necessário fazer a *sincronização* desses elementos com as marcas de *pitch* anteriormente calculadas. Isso porque tanto os pontos de corte como as marcas de transição foram calculados manualmente, ao passo que as marcas de *pitch* foram determinadas de forma automática, ou seja, não havia, necessariamente, coincidência entre esses elementos. A sincronização visou justamente a garantir o alinhamento com as marcas de *pitch* do sinal.

Uma vez calculadas as marcas de *pitch* e marcas de transição, e extraídas as unidades do interior das frases-veículo, partiu-se para a última etapa, que foi a montagem do dicionário

propriamente dito. Dois dicionários foram montados: um para a síntese PSOLA e outro para a síntese híbrida.

O dicionário PSOLA contém simplesmente a forma de onda de cada uma das unidades, além das marcas de *pitch* e marcas de transição a ela associadas. Cada marca de transição é um número inteiro simples, representando a distância relativa (em amostras) entre a posição da marca e o início da unidade. As marcas de *pitch* contêm, por sua vez, além do valor de distância relativa, um valor indicando se a marca corresponde a uma porção sonora ou não-sonora do sinal. O dicionário híbrido também contém as marcas de transição e de *pitch* das unidades; no entanto, a unidade em si não é representada pela sua forma de onda, mas sim por meio dos *parâmetros híbridos* descritos na seção 7.3.2.

A montagem dos dicionários foi feita de forma automática com o auxílio de dois programas desenvolvidos em nosso laboratório especificamente para esse fim. O primeiro, exclusivo para a montagem do dicionário híbrido, foi responsável pelo cálculo dos parâmetros híbridos (harmônicos e de ruído) de cada uma das unidades do dicionário. O segundo, utilizado na montagem de ambos os dicionários, tinha por função efetuar a geração de dois arquivos distintos: um *arquivo de sinal*, correspondente ao dicionário propriamente dito, contendo as marcas de transição, as marcas de *pitch* e as unidades (forma de onda no caso do dicionário PSOLA e parâmetros híbridos no caso do dicionário híbrido); e um arquivo de índices, contendo o nome de cada uma das unidades contidas no dicionário e a posição relativa de cada uma dessas unidades dentro do arquivo de sinal.

Os dicionários criados contam, atualmente, com 2041 unidades. Não se trata ainda de uma versão definitiva: algumas unidades sofreram problemas de natureza diversa durante a etapa de gravação e segmentação (*clipping*, presença de ruído, leitura incorreta ou mal articulada, erros durante a segmentação, etc.). Tais unidades ainda precisam ser regravadas ou ressegmentadas. Houve ainda algumas redefinições a respeito de certas unidades que devem constar do inventário. Essas redefinições ocorreram posteriormente ao processo de gravação, e decorreram de um conjunto de testes por nós efetuados sobre textos diversos. Por isso algumas

outras unidades precisarão ainda ser geradas. A versão final do dicionário deverá contar com aproximadamente 2450 unidades.

## 8.9 O módulo de síntese

O último módulo a entrar em funcionamento durante o processo de conversão texto-fala é o módulo responsável pela síntese propriamente dita. Ele tem por objetivo determinar, a partir do conteúdo do inventário de unidades do sistema, qual é a seqüência de polifones correspondente à seqüência fonética a ser sintetizada, seqüência essa previamente calculada pelo módulo de transcrição ortográfico-fonética. Uma vez determinada a seqüência de polifones, estes devem ser concatenados, e os parâmetros prosódicos dos segmentos fonéticos constituintes da sentença obtida por meio da concatenação devem ser alterados, de forma a se adequarem aos parâmetros calculados durante a etapa de processamento prosódico.

Num sistema de síntese por concatenação de sub-unidades o processo de segmentação da seqüência fonética em uma seqüência de polifones é de responsabilidade do *gerador segmental*. Obviamente, o resultado dessa segmentação depende do conjunto de polifones contido no interior do dicionário de unidades. O gerador segmental deve sempre selecionar, dentre as unidades existentes no inventário, uma seqüência que corresponda exatamente à seqüência fonética desejada. O número de seqüências possíveis nem sempre é único; no entanto, a segmentação deve ser feita de forma que as unidades selecionadas sejam sempre as maiores possíveis. Isso porque quanto maior for a unidade, mais preservada estará a coarticulação entre os elementos que lhe são internos.

O gerador segmental para o novo dicionário de polifones (2450 unidades) foi construído inicialmente como parte de um trabalho de curso, sendo mais tarde incorporado ao sistema de conversão. Ele calcula a seqüência de polifones ótima a partir da aplicação de um conjunto de *regras de formação*. Para isso ele faz a interpretação de um *arquivo de regras*,

onde cada regra é responsável pela descrição de um conjunto de unidades. Há uma regra específica para a formação dos polifones do tipo *"cv"* (consoante + vogal tônica), outra regra para polifones do tipo *"cV"* (consoante + vogal pós-tônica), e assim por diante. O arquivo de regras completo, descrevendo a formação de todos os grupos de polifones, é mostrado no Apêndice V.

O objetivo da utilização das regras de formação foi o de tornar o funcionamento do gerador segmental o mais independente possível do inventário de unidades do sistema. De fato, ele pode ser utilizado com um inventário qualquer, desde que haja um conjunto de regras apropriado descrevendo esse inventário. A inserção de um novo grupo de unidades a um inventário já existente acarreta a inserção de uma única linha ao arquivo de regras, contendo a regra de formação desse grupo específico de unidades. A tarefa de segmentação em si também se torna mais rápida, pois ao invés de efetuar buscas pelos polifones válidos, o algoritmo de segmentação realiza buscas ao longo das regras de formação. Obviamente, o número de regras de formação é extremamente menor que o número de polifones contidos no inventário.

Um exemplo do resultado produzido pelo gerador segmental se encontra a seguir. Ele corresponde à seqüência fonética apresentada na seção 8.6 deste capítulo.

*/zh zho oA aNUr ro om mA aNUf fo oI/ /d do oS St tRe ez zE Ea aO OS Sv vi iNt tE Ee es si iNk kO Oa an nO OS/ /e eNp pRe eg ga ad dO Od de eu uNv ve eNd de eIRO Ok ke ee eNr ri ik ke es se eUe eNt tRE Ea aS Sk kUa at tRO Op pa aRe ed dE ES Sd de eu um mA As su uzh zhA Ae eo ob bI IS Sk ku uRA At ta av veh ehR Rn nA An no oS Sr re ef fo olh lhO OS Sd do ob ba aIr rO Od do ob boh oht ta af fo og gO O/ /e et tA aNt tU Ue ek ko on no om mi iz zo oUd do op po oUk kO Ok ke eg gA aNnh nha aRA An neh ehs sA Ad du uz zI IAd de ea an nO OS/ /k ke e/ /a aUr re et ti iRa aR Rs sE Eo op pa at tRA aNUp pa aRA Aa at teh ehr rA A/ /lh lhe ed de eIsh sho oU/ /e eNp pa ag ga am me eNt tO Od de eo oR Rd de en na ad dO OS Sv ve eNs si id dO OS/ /n ne eNs soh oha av ve eNd dA Ak ko oNo ok ke ee eS St ta av vA Ad de eNt tRO O/ /k ko om mO Oa ai iNd dA Au uNk ko oNt tO Oe ek ki inh nhe eNt tO OSe eNd di inh nhe eIRO O/*

Uma vez determinado o conjunto de unidades a serem concatenadas, resta apenas ao sistema efetuar a síntese propriamente dita. O aplicativo de conversão texto-fala implementado permite ao usuário selecionar duas técnicas de síntese distintas: TD-PSOLA e síntese híbrida. De acordo com a opção escolhida o sistema seleciona as unidades do dicionário apropriado e realiza a concatenação, bem como a alteração dos parâmetros prosódicos (F0 e duração), seguindo os procedimentos descritos nas seções 7.3.1 e 7.3.2.

Ao final da etapa de síntese, tem-se um arquivo de fala sintetizada na forma digital, com precisão de 16 bits e amostrado a 16 kHz, no formato WAV.

# 9 Conclusões

## 9.1 Considerações sobre o trabalho desenvolvido

O presente trabalho procurou, em primeiro lugar, apresentar o problema da síntese de fala a partir de texto de maneira bastante genérica. Desse modo, mostrou-se que um sistema de conversão texto-fala apresenta uma estrutura bastante modular, e que os seus módulos principais (módulos de pré-processamento, de transcrição ortográfico-fonética, de processamento prosódico e de síntese do sinal) têm cada um uma importância específica no funcionamento do sistema. A função de cada um dos módulos foi descrita com detalhes, como também foram apresentadas algumas abordagens para cada uma das etapas do processo de conversão texto-fala.

Em seguida foi apresentada uma estratégia de implementação de um sistema de conversão texto-fala para o português falado no Brasil. O sistema implementado é fruto não apenas do trabalho de mestrado relatado nesta tese, mas também de um esforço conjunto do Laboratório de Processamento Digital de Fala, pertencente ao Departamento de Comunicações da Faculdade de Engenharia Elétrica e de Computação da UNICAMP, e do Laboratório de Fonética Acústica e Psicolingüística Experimental, do Instituto de Estudos da Linguagem da UNICAMP.

O sistema de conversão aqui apresentado baseia-se no método de síntese concatenativa. Um inventário de 2041 unidades de fala para concatenação, a ser ampliado para 2450 unidades, foi construído para esse fim. Para a criação das unidades foram gravadas frases-veículo cuidadosamente selecionadas, de forma que houvesse um ambiente fonético-prosódico neutro em torno das unidades a serem segmentadas. A segmentação dessas unidades foi feita de forma manual, a fim de garantir maior precisão na realização dos cortes.

A etapa de processamento lingüístico realizada pelo sistema consiste primeiramente de uma normalização do texto de entrada (tratamento de números, siglas, abreviaturas e símbolos

especiais), efetuada por um módulo de pré-processamento, cujo funcionamento está baseado na existência de um compilador de regras. A etapa de transcrição ortográfico-fonética, por sua vez, é realizada por dois módulos distintos: um aplicador de regras de transcrição, responsável pelo tratamento das correspondências regulares entre letras e sons; e um dicionário de exceções, responsável pela transcrição das palavras para as quais as regras falham. O ortofon tem uma taxa de acerto de cerca de 96%, que é um valor considerado muito bom.

Essa taxa de erro diz respeito à conversão das palavras em isolado. No entanto, muitos dos erros cometidos durante a etapa de transcrição ocorrem devido à ausência do módulo de análise pós-lexical (processador métrico), ainda não implementado, responsável por determinar as alterações na transcrição convencional das palavras que ocorrem devido à interação com as palavras vizinhas.

A notação fonética utilizada pelo sistema tem como característica principal a distinção entre segmentos fonéticos *plenos* e *reduzidos*, os quais se distinguem pelo nível pelo qual estão sujeitos aos fenômenos de coarticulação. Quanto ao inventário de unidades utilizado pelo sistema, utilizou-se o critério básico de preservar intactos os segmentos reduzidos e os encontros vocálicos. Como a adoção rígida desse critério levaria à criação de um inventário excessivamente grande, foram utilizadas algumas condições de contorno que permitiram efetuar a segmentação no interior de segmentos reduzidos:

- vogais pós-tônicas foram segmentadas no final da transição com a consoante precedente.
- vogais e ditongos nasais são concatenados com onsets orais, evitando-se a necessidade de unidades específicas para os onsets nasais.
- os segmentos fracos S, R e L, quando no início de encontros consonantais, são segmentados.

- segmento reduzido L, quando em final de sílaba, é transformado na semivogal U, o que evita a necessidade de unidades específicas contendo L.

O sistema foi construído de forma a comportar a existência de um módulo de processamento prosódico, responsável pela determinação automática dos parâmetros prosódicos (F0 e duração) da sentença a ser sintetizada. Por ora esse módulo ainda não foi incorporado ao sistema (o de duração já foi implementado no LAFAPE), mas o sistema permite fornecer os parâmetros prosódicos de forma manual, através de um arquivo contendo valores de duração e de F0 inicial e F0 final para cada um dos segmentos da sentença.

O sistema permite a utilização de duas técnicas de síntese: TD-PSOLA e síntese híbrida, sendo que a técnica selecionada é responsável pelo processo de concatenação e de modificação dos parâmetros prosódicos da sentença.

Pelo fato de o sistema ainda não estar totalmente implementado, nenhuma análise a respeito dos resultados por ele produzidos pode ser considerada de caráter definitivo. Além de o sistema não contar com os módulos de análise pós-lexical (processador métrico) e de processamento prosódico, o inventário de unidades ressente-se da falta de cerca de 400 unidades, que ainda precisam ser incorporadas ao conjunto.

Mesmo assim, procurou-se efetuar alguma forma de avaliação simples, que pudesse ao menos mostrar o potencial do sistema até aqui desenvolvido e indicar se as estratégias adotadas até o presente momento foram adequadas.

A avaliação consistiu basicamente em fornecer um conjunto de 10 frases ao sistema e fazer com que ele as sintetizasse. As frases escolhidas foram baseadas em uma lista de frases foneticamente balanceadas que consta do trabalho de Alcaim *et al.* [3]. Além disso, as frases foram escolhidas de forma que pudessem ser completamente geradas a partir da versão atual (ainda incompleta) do inventário de unidades do sistema. As dez frases são listadas a seguir:

*Os maiores picos da terra ficam em baixo d'água.*

*A inauguração ali na vila é Quinta-feira.*

*Só vota quem tiver o título de eleitor.*

*É fundamental buscar a razão e o sentido da existência.*

*A temperatura só é boa mais cedo.*

*Em Cuba e muitas outras regiões a população está quase diminuindo.*

*Nunca se pode esquecer de ficar em cima do muro.*

*Prá quem vê de fora o panorama é desolador.*

*É bom te ver colhendo plantas.*

*Eu me banho no lago ao amanhecer.*

As etapas de pré-processamento e de transcrição fonética foram feitas de forma automática pelo sistema, mas os parâmetros prosódicos (F0 inicial, F0 final e duração dos segmentos) foram inseridos manualmente.

A primeira intenção do teste foi fazer uma avaliação preliminar das duas técnicas de síntese utilizadas pelo sistema (síntese híbrida e TD-PSOLA). Para tanto, as dez frases foram sintetizadas utilizando-se ambas as técnicas, gerando-se um total de vinte frases. Essas frases foram submetidas a uma avaliação acústica informal e subjetiva, por parte de alguns ouvintes.

De maneira geral, os melhores resultados foram obtidos por meio da aplicação do TD-PSOLA. O sinal produzido pela técnica híbrida foi considerado inferior por apresentar a presença constante de um ruído de fundo, inexistente no sinal produzido pelo TD-PSOLA, ruído este que confere à fala sintetizada um aspecto rouco e abafado. Quanto ao sinal gerado pelo TD-PSOLA, mostrou-se de muito boa qualidade, apresentando, contudo, o problema típico dessa técnica de síntese, que é o de conferir à fala sintetizada uma qualidade levemente

metálica, principalmente em condições de variação prosódica mais larga. Nesse aspecto, a técnica híbrida se mostrou superior ao TD-PSOLA, por não introduzir esse tipo de efeito; no entanto, a avaliação deixa evidente que a versão atual do algoritmo implementado para a técnica híbrida ainda precisa ser aperfeiçoado, a fim de amenizar o ruído de fundo que foi observado. Vale aqui ressaltar que numa avaliação anterior, com uma versão diferente do dicionário de polifones (1200 unidades), construído com a voz de outro locutor, a síntese híbrida chegou a apresentar resultados melhores [70].

Como a preferência de qualidade recaiu sobre a síntese TD-PSOLA, esta foi a técnica utilizada para gerar as frases utilizadas na avaliação geral do desempenho do sistema de conversão texto-fala. O mesmo conjunto de dez frases foi analisado.

No tocante à inteligibilidade (qualidade que diz respeito ao grau de compreensão do texto relativo à sentença sintetizada), pode-se dizer que esta foi praticamente total, pois todas as sentenças foram bem compreendidas pelas pessoas que as ouviram. Trata-se de um indicador do acerto na escolha dos critérios de elaboração do inventário de unidades, pois a utilização de unidades mal-elaboradas, que não contemplassem os fenômenos de coarticulação entre segmentos de maneira satisfatória, certamente levaria a dificuldades de compreensão. A inteligibilidade se mostrou também bastante alta para outras frases que foram geradas em testes no nosso laboratório.

Quanto à naturalidade, mostrou-se razoável (dentro do esperado). Obviamente a avaliação de naturalidade é bastante prejudicada pelo fato de a prosódia das frases sintetizadas ter sido determinada de forma arbitrária e inserida manualmente. A presença de um módulo prosódico atuante no sistema sem dúvida implicaria em ganhos consideráveis na qualidade da fala sintetizada. De qualquer forma, a naturalidade se mostrou satisfatória, tendo-se em vista o fato de que o inventário foi construído de forma a preservar uma boa parte dos fenômenos prosódicos no interior das próprias unidades [6], principalmente devido à preservação dos segmentos reduzidos, que são exatamente os mais sensíveis ao ambiente prosódico no qual estão inseridos.

Quanto à qualidade da concatenação, mostrou-se extremamente satisfatória. Não foram detectadas, praticamente, descontinuidades espectrais perceptíveis à audição, o que avaliza os critérios utilizados na seleção dos pontos de corte e na concatenação propriamente dita. Os poucos casos em que foram observadas descontinuidades resultaram em reformulações do inventário de unidades.

## 9.2 Propostas para trabalhos futuros

O sucesso na implementação do sistema de conversão texto-fala apresentado nesta dissertação, muito mais do que representar o fechamento de um trabalho, abre caminhos para a realização de novas tarefas, as quais deverão dar continuidade ao trabalho até aqui realizado. Apontaremos aqui algumas sugestões de atividades visando a alcançar uma maior integração do sistema como um todo, bem como a melhoria da qualidade final da saída gerada.

Uma primeira tarefa importante, relacionada à etapa de processamento lingüístico, seria a implementação de alguns aperfeiçoamentos no módulo de pré-processamento. A versão atual é capaz de lidar apenas com os casos de conversão mais simples. Conforme discussão apresentada no capítulo 5, no entanto, pudemos constatar que o processamento de certos elementos do texto é inerentemente ambíguo, e requer mecanismos de análise do conteúdo da mensagem contida no texto. A inclusão de tais mecanismos não é tarefa simples, mas certamente trará maior eficiência ao módulo de pré-processamento.

Já o módulo de transcrição ortográfico-fonética mostrou ser capaz de produzir resultados bastante coerentes, com uma taxa de erro muito baixa. Essa taxa de erro pode ser diminuída ainda mais com a inclusão de algumas regras de transcrição que ainda não foram implementadas, por ainda estarem em fase de estudos pela equipe de trabalho do LAFAPE. Além disso, seria conveniente implementar o módulo de transcrição de maneira semelhante àquela que foi utilizada na implementação do módulo de pré-processamento, ou seja, por meio da utilização de um compilador de regras. Tal abordagem permitiria a atualização das regras de transcrição de maneira trivial, sem a necessidade de alterar o código do programa, o que

seria útil tanto para a realização de testes de avaliação como também para a inclusão de modificações efetivas no processo de transcrição. A definição, por parte do LAFAPE, das regras ainda não implementadas, permitirá avaliar a necessidade ou não de reformulação da versão atual do compilador de regras, a fim de que este possa ser efetivamente utilizado na implementação do módulo de transcrição. Isto está sendo providenciado através do trabalho de iniciação científica de Marcelo Rebelo, continuado por Daniel Araújo, ambos do Instituto de Computação, orientado pelo Prof. Plínio Barbosa.

Ainda com relação à etapa de processamento lingüístico, falta ao sistema um módulo de análise pós-lexical, ao qual denominamos processador métrico, responsável por determinar as modificações na transcrição das palavras devido à interação destas com as palavras vizinhas. A presença de tal módulo representaria um refinamento importante à saída produzida pelo ortofon, e não se trata de um quesito opcional, pois a sua ausência implica em alguns erros na etapa de determinação das unidades a serem concatenadas. Isso porque a não aplicação dessa análise pós-lexical pode levar à geração de seqüências fonéticas que não existem na língua e que, portanto, não podem ser geradas a partir do inventário de unidades do sistema.

Outra tarefa essencial para o funcionamento correto do sistema é a conclusão da confecção do inventário de unidades para concatenação. O inventário atual conta com 2041 unidades, mas a versão final deverá conter cerca de 2450 unidades. Somente com a presença do inventário completo será possível realizar a síntese de qualquer texto apresentado como entrada do sistema.

Dentre todas as tarefas aqui apontadas, a mais urgente e sem dúvida a mais essencial é a inclusão no sistema de um módulo de processamento prosódico. O sistema atual provê apenas um mecanismo de inserção de prosódia manual. A existência de um módulo responsável pela determinação dos parâmetros prosódicos da sentença a ser sintetizada tornará o sistema verdadeiramente automático. Além disso, o cálculo correto desses parâmetros prosódicos é essencial para garantir a inteligibilidade e principalmente a naturalidade da fala

sintetizada. Um modelo de duração já foi construído pela equipe de trabalho do LAFAPE, bastando apenas adaptá-lo às mudanças de notação e de locutor introduzidas no sistema de conversão texto-fala de 1995 para cá e inseri-lo no aplicativo. Já o modelo entoacional ainda precisa ser elaborado.

No que diz respeito às técnicas de síntese, seria interessante desenvolver um trabalho de aperfeiçoamento do algoritmo de síntese híbrida atualmente implementado, visto que este ainda não apresentou resultados com a qualidade que dele se esperava. Uma outra tarefa a ser realizada nesse sentido seria a redução da complexidade do algoritmo de síntese híbrida, uma preocupação que não foi levada em conta na implementação atual, mas que é importante para a redução do tempo total de síntese. Atualmente, a técnica híbrida está longe de ser aplicável em sistemas que trabalhem em tempo real.

Por fim, a modularidade do sistema implementado permitirá que outras técnicas de síntese que venham a ser implementadas possam também ser testadas. Como sugestão de atividade nesse sentido, poder-se-ia trabalhar em cima da implementação da síntese MBROLA que, segundo a literatura, apresenta resultados superiores àqueles obtidos por meio do TD-PSOLA tradicional.

# Apêndice I - Arquivo de regras de pré-processamento

__GROUPS__

nada [""]

digito [0-9]

nondigit ![0-9]

naonulo [1-9]

Pontuacao [",",";",".", "!", "?"]

Separador [" ", \n, \t, "-", \", "(", ")"]

FimDePalavra [{Pontuacao}, {Separador}]

__RULES__

" "[" "]+ := " ";

//<[{digito}]+> " " :=;

<[a-z]|[A-Z]> "-" <[a-z]|[A-Z]> := " "; // hifen

"-" := ","; // travessao

"(" := ",";

")" := ",";

";" := ",";

"\t" := " ";

"\n" := " ";

"©" := *;

A := a;

B := b;

C := c;

D := d;

E := e;

F := f;

G := g;

H := h;

I := i;

J := j;

K := k;

L := l;

M := m;

N := n;

O := o;

P := p;

Q := q;

R := r;

S := s;

T := t;

U := u;

V := v;

W := w;

X := x;

Y := y;

```
Z := z;


// Algarismos


// zeros e seus efeitos


<.> 000 := ;
<.> 001 <. {digito} {digito} {digito} {FimDePalavra} {nondigit}> :=;
<.> 000. <[{digito}]+> := ;
<.> 0 <{naonulo} {digito} {FimDePalavra} {nondigit}> := (e);
<.> 00 <{naonulo} {FimDePalavra} {nondigit}> := (e);


// milhares


.0 <{naonulo} {digito}> := ("mil e");
.00 <{naonulo}> := ("mil e");
. <{digito} {digito} {digito}> := (mil);
<{nondigit}> 1.00 <{naonulo} {FimDePalavra} {nondigit}> := ("mil e") ;
<{nondigit}> 1.0 <{naonulo} {digito} {FimDePalavra} {nondigit}> := ("mil e") ;
<{nondigit}> 1. <{digito} {digito} {digito} {FimDePalavra} {nondigit}> := (mil) ;
```

// milhao

<{nondigit} 1> . <{digito} {digito} {digito} . {digito} {digito} {digito}> := (milhão);

<001> . <{digito} {digito} {digito} . {digito} {digito} {digito}> := (milhão);

. <{digito} {digito} {digito} . {digito} {digito} {digito}> := (milhões);

// bilhao

<{nondigit} 1> . <{digito} {digito} {digito} . {digito} {digito} {digito} . {digito} {digito} {digito}> := (bilhão);

<001> . <{digito} {digito} {digito} . {digito} {digito} {digito} . {digito} {digito} {digito}> := (bilhão);

. <{digito} {digito} {digito} . {digito} {digito} {digito} . {digito} {digito} {digito}> := (bilhões);

// trilhao

<{nondigit} 1> . <{digito} {digito} {digito} . {digito} {digito} {digito} . {digito} {digito} {digito} . {digito} {digito} {digito}> := (trilhão);

<001> . <{digito} {digito} {digito} . {digito} {digito} {digito} . {digito} {digito} {digito} . {digito} {digito} {digito}> := (trilhão);

. <{digito} {digito} {digito} . {digito} {digito} {digito} . {digito} {digito} {digito} . {digito} {digito} {digito}> := (trilhões);

// quadrilhao

<{nondigit} 1> . <{digito} {digito} {digito} . {digito} {digito} {digito} . {digito} {digito} {digito} . {digito} {digito} {digito} . {digito} {digito} {digito}> := (quatrilhão);

<001> . <{digito} {digito} {digito} . {digito} {digito} {digito} . {digito} {digito} {digito} . {digito} {digito} {digito} . {digito} {digito} {digito}> := (quatrilhão);

. <{digito} {digito} {digito} . {digito} {digito} {digito} . {digito} {digito} {digito} . {digito} {digito} {digito} . {digito} {digito} {digito}> := (quatrilhões);

0 := (zero);

1 := (um);

2 := (dois);

3 := (três);

4 := (quatro);

5 := (cinco);

6 := (seis);

7 := (sete);

8 := (oito);

9 := (nove);

10 := (dez);

11 := (onze);

12 := (doze);

13 := (treze);

14 := (catorze);

15 := (quinze);

16 := (dezesseis);

17 := (dezessete);

18 := (dezoito);

19 := (dezenove);

20 := (vinte);

30 := (trinta);

40 := (quarenta);

50 := (cinqüenta);

60 := (sessenta);

70 := (setenta);

80 := (oitenta);

90 := (noventa);

0 < {digito} > := ;

2 < {digito} > := ("vinte e");

3 < {digito} > := ("trinta e");

4 < {digito} > := ("quarenta e");

5 < {digito} > := ("cinqüenta e");

6 < {digito} > := ("sessenta e");

7 < {digito} > := ("setenta e");

```
8 < {digito} > := ("oitenta e");

9 < {digito} > := ("noventa e");


100 := (cem);

200 := (duzentos);

300 := (trezentos);

400 := (quatrocentos);

500 := (quinhentos);

600 := (seiscentos);

700 := (setecentos);

800 := (oitocentos);

900 := (novecentos);


<.> 100 <{FimDePalavra} {nondigit}> := ("e cem");

<.> 200 <{FimDePalavra} {nondigit}> := ("e duzentos");

<.> 300 <{FimDePalavra} {nondigit}> := ("e trezentos");

<.> 400 <{FimDePalavra} {nondigit}> := ("e quatrocentos");

<.> 500 <{FimDePalavra} {nondigit}> := ("e quinhentos");

<.> 600 <{FimDePalavra} {nondigit}> := ("e seiscentos");

<.> 700 <{FimDePalavra} {nondigit}> := ("e setecentos");

<.> 800 <{FimDePalavra} {nondigit}> := ("e oitocentos");

<.> 900 <{FimDePalavra} {nondigit}> := ("e novecentos");
```

1 < {digito} {digito} > := ("cento e");

2 < {digito} {digito} > := ("duzentos e");

3 < {digito} {digito} > := ("trezentos e");

4 < {digito} {digito} > := ("quatrocentos e");

5 < {digito} {digito} > := ("quinhentos e");

6 < {digito} {digito} > := ("seiscentos e");

7 < {digito} {digito} > := ("setecentos e");

8 < {digito} {digito} > := ("oitocentos e");

9 < {digito} {digito} > := ("novecentos e");

// Abreviaturas

(a|A)v. := (avenida);

<{Separador}> cm <{FimDePalavra}> := (centímetros);

(d|D)r. := (doutor);

(d|D)ra. := (doutora);

<{Separador}> etc. := (etcetera);

<{Separador}> (j|J)r. := (júnior);

"%" := (porcento);

(p|P)rof. := (professor);

(s|S)r. := (senhor);

(s|S)ra. := (senhora);

// Siglas

<[A-Z]+> A <[A-Z]|{FimDePalavra}> := (á);
<[A-Z]|{FimDePalavra}> A <[A-Z]+> := (á);

<[A-Z]+> B <[A-Z]|{FimDePalavra}> := (bê);
<[A-Z]|{FimDePalavra}> B <[A-Z]+> := (bê);

<[A-Z]+> C <[A-Z]|{FimDePalavra}> := (cê);
<[A-Z]|{FimDePalavra}> C <[A-Z]+> := (cê);

<[A-Z]+> D <[A-Z]|{FimDePalavra}> := (dê);

<[A-Z]|{FimDePalavra}> C <[A-Z]+> := (dê);

<[A-Z]+> E <[A-Z]|{FimDePalavra}> := (ê);

<[A-Z]|{FimDePalavra}> E <[A-Z]+> := (ê);

<[A-Z]+> F <[A-Z]|{FimDePalavra}> := (éfe);

<[A-Z]|{FimDePalavra}> F <[A-Z]+> := (éfe);

<[A-Z]+> G <[A-Z]|{FimDePalavra}> := (gê);

<[A-Z]|{FimDePalavra}> G <[A-Z]+> := (gê);

<[A-Z]+> H <[A-Z]|{FimDePalavra}> := (agá);

<[A-Z]|{FimDePalavra}> H <[A-Z]+> := (agá);

<[A-Z]+> I <[A-Z]|{FimDePalavra}> := (í);

<[A-Z]|{FimDePalavra}> I <[A-Z]+> := (í);

<[A-Z]+> J <[A-Z]|{FimDePalavra}> := (jóta);

<[A-Z]|{FimDePalavra}> J <[A-Z]+> := (jóta);

<[A-Z]+> K <[A-Z]|{FimDePalavra}> := (cá);

<[A-Z]|{FimDePalavra}> K <[A-Z]+> := (cá);

<[A-Z]+> L <[A-Z]|{FimDePalavra}> := (éle);

<[A-Z]|{FimDePalavra}> L <[A-Z]+> := (éle);

<[A-Z]+> M <[A-Z]|{FimDePalavra}> := (eme);

<[A-Z]|{FimDePalavra}> M <[A-Z]+> := (eme);

<[A-Z]+> N <[A-Z]|{FimDePalavra}> := (ene);

<[A-Z]|{FimDePalavra}> N <[A-Z]+> := (ene);

<[A-Z]+> O <[A-Z]|{FimDePalavra}> := (ó);

<[A-Z]|{FimDePalavra}> O <[A-Z]+> := (ó);

<[A-Z]+> P <[A-Z]|{FimDePalavra}> := (pê);

<[A-Z]|{FimDePalavra}> P <[A-Z]+> := (pê);

<[A-Z]+> Q <[A-Z]|{FimDePalavra}> := (quê);

<[A-Z]|{FimDePalavra}> Q <[A-Z]+> := (quê);

<[A-Z]+> R <[A-Z]|{FimDePalavra}> := (érre);

<[A-Z]|{FimDePalavra}> R <[A-Z]+> := (érre);

<[A-Z]+> S <[A-Z]|{FimDePalavra}> := (ésse);

<[A-Z]|{FimDePalavra}> S <[A-Z]+> := (ésse);

<[A-Z]+> T <[A-Z]|{FimDePalavra}> := (tê);

<[A-Z]|{FimDePalavra}> T <[A-Z]+> := (tê);

<[A-Z]+> U <[A-Z]|{FimDePalavra}> := (ú);

<[A-Z]|{FimDePalavra}> U <[A-Z]+> := (ú);

<[A-Z]+> V <[A-Z]|{FimDePalavra}> := (vê);

<[A-Z]|{FimDePalavra}> V <[A-Z]+> := (vê);

<[A-Z]+> W <[A-Z]|{FimDePalavra}> := (dábliu);

<[A-Z]|{FimDePalavra}> W <[A-Z]+> := (dábliu);

<[A-Z]+> X <[A-Z]|{FimDePalavra}> := (xis);

<[A-Z]|{FimDePalavra}> X <[A-Z]+> := (xis);

<[A-Z]+> Y <[A-Z]|{FimDePalavra}> := (ípsilon);

<[A-Z]|{FimDePalavra}> Y <[A-Z]+> := (ípsilon);

<[A-Z]+> Z <[A-Z]|{FimDePalavra}> := (zê);

<[A-Z]|{FimDePalavra}> Z <[A-Z]+> := (zê);

// Simbolos Especiais

"+" := (mais);

"=" := (igual);

< [{digito}]+ > "," < [{digito}]+ > := ( vírgula );

__END__

# Apêndice II - Notação fônica utilizada pelo Ortofon

| Consoantes plenas | |
|---|---|
| **p** | pata (**p**atA) |
| **b** | bala (**b**alA) |
| **f** | faca (**f**akA) |
| **v** | vela (**v**ehlA) |
| **m** | mola (**m**ohlA) |
| **t** | tapa (**t**apA) |
| **d** | data (**d**atA) |
| **s** | sela (**s**ehlA) |
| **z** | zona (**z**onA) |
| **n** | nada (**n**adA) |
| **r** | rato (**r**atO) |
| **l** | lata (**l**atA) |
| **k** | casa (**k**azA) |
| **g** | gota (**g**otA) |
| **sh** | cheque (**sh**ehkE) |
| **zh** | jogo (**zh**ogO) |
| **nh** | vinho(vi**nh**O) |
| **lh** | molho (mo**lh**O) |

| Consoantes reduzidas | |
|---|---|
| **S** | mesmo (me**S**mO) |
| **N** | ponte (po**N**tE) |
| **L** | placa (p**L**akA) |
| **R** | prato (p**R**atO) |

| Vogais plenas | |
|---|---|
| **i** | pipa (p**i**pA) |
| **e** | medo (m**e**dO) |
| **eh** | sela (s**eh**lA) |
| **a** | lata (l**a**tA) |
| **oh** | bola (b**oh**lA) |
| **o** | bolo (b**o**lO) |
| **u** | mula (m**u**lA) |

| Vogais reduzidas | |
|---|---|
| **I** | rápido (rap**I**dO) |
| **E** | trôpego (tRop**E**gO) |
| **A** | casa (kaz**A**) |
| **O** | pérola (pehR**O**lA) |
| **U** | glóbulo (gLohb**U**lO) |

# Apêndice III - Segmentos Fonéticos

| FONE | DURAÇÃO (ms) | DESVIO PADRÃO (ms) |
|---|---|---|
| i (p**i**pa) | 87 | 19 |
| e (m**e**do) | 110 | 19 |
| eh (s**e**la) | 113 | 17 |
| a (s**a**la) | 132 | 45 |
| oh (b**o**la) | 119 | 18 |
| o (b**o**bo) | 111 | 20 |
| u (b**u**la) | 103 | 14 |
| I (pân**i**co) | 55 | - |
| A (bal**a**) | 59 | 15 |
| U (óv**u**lo) | 47 | 15 |
| E (trôp**e**go) | 51 | 13 |
| O (pér**o**la) | 51 | 14 |
| aN (b**an**co) | 134 | 15 |
| eN (p**en**te) | 128 | 15 |
| iN (v**in**te) | 128 | 16 |
| oN (p**on**te) | 137 | 18 |
| uN (b**um**bo) | 128 | 10 |
| AN (ím**ã**) | 93 | 34 |
| EN (híf**en**) | 82 | 26 |
| IN (ínter**im**) | 89 | 25 |
| ON (mórm**on**) | 88 | 19 |
| UN (fór**um**) | 98 | 16 |
| p (**p**ato) | 80 | 16 |
| t (**t**atu) | 81 | 20 |
| k (**c**asa) | 80 | 22 |
| b (**b**ola) | 59 | 18 |
| d (**d**ado) | 58 | 22 |
| g (**g**ato) | 52 | 20 |
| f (**f**oca) | 89 | 28 |
| s (**s**apo) | 96 | 24 |
| sh (bi**ch**o) | 104 | 19 |
| v (**v**ela) | 57 | 19 |
| z (ro**s**a) | 64 | 25 |
| zh (**g**elo) | 65 | 17 |

| | | |
|---|---|---|
| m (**m**ala) | 61 | 13 |
| n (**n**eto) | 51 | 17 |
| nh (vi**nh**o) | 84 | 27 |
| r (**r**ato) | 78 | 24 |
| l (**l**ata) | 47 | 13 |
| lh (i**lh**a) | 72 | 24 |
| S (a**s**ma) | 74 | 31 |
| R (p**r**ato) | 30 | 10 |
| L (p**l**aca) | 41 | 11 |
| oU (l**ou**co) | 133 | 14 |
| ehI (id**éi**a) | 134 | 13 |
| ohI (r**ói**) | 133 | 12 |
| ohU (s**ol**) | 138 | 19 |
| aI (b**ai**le) | 135 | 12 |
| aU (**au**la) | 145 | 15 |
| oNI (p**õe**) | 120 | 10 |
| aNI (m**ãe**) | 143 | 18 |
| aNU (m**ão**) | 140 | 13 |
| ehU (r**éu**) | 143 | 15 |
| uI (f**ui**) | 122 | 13 |
| eI (p**ei**to) | 126 | 13 |
| oI (l**oi**ra) | 131 | 14 |
| eU (m**eu**) | 134 | 17 |
| iU (ca**iu**) | 145 | 23 |
| uU (az**ul**) | - | - |
| Ui (sag**üi**) | - | - |
| Ue (seq**üe**strei) | - | - |
| Ueh ((seq**üe**stro) | - | - |
| Ua (q**ua**se) | - | - |
| Uoh (q**uó**rum) | - | - |
| Uo (q**uo**ciente) | - | - |
| IO (pát**io**) | 61 | 16 |
| UO (árd**uo**) | 63 | 18 |
| IE (sér**ie**) | 49 | 14 |
| UE (tên**ue**) | 106 | 14 |
| IA (pátr**ia**) | 80 | 17 |
| UA (ág**ua**) | 70 | 16 |
| AU (asdrúb**al**) | - | - |

| | | |
|---|---|---|
| EU (amáv**el**) | - | - |
| IU (út**il**) | - | - |
| OU (álc**ool**) | - | - |
| ANU (sót**ão**) | 96 | 26 |
| UaI (parag**uai**) | - | - |
| UaU (ig**ual**) | - | - |
| Uei (enxag**üei**) | - | - |
| Ui U (arg**üiu**) | - | - |
| UoU (enxag**uou**) | - | - |
| UaN (q**uan**do) | - | - |
| UeN (ag**üen**ta) | - | - |
| UAN(enxag**uam**) | - | - |
| UEN(enxag**üem**) | - | - |
| UaNU (sag**uão**) | - | - |
| UoNI (sag**uõe**s) | - | - |

As durações e desvios-padrão em branco ainda não foram calculados.

# Apêndice IV – Estrutura do inventário de unidades

**Grupo A – Tônicas e pré-tônicas**

**( V = vogais e grupos vocálicos, C = consoantes e grupos consonantais)**

| Código do Grupo | Estrutura do Grupo | Observações | Exemplos |
|---|---|---|---|
| A1 | CV | | pa, peh, ba, beh...<br>pRa, pReh, pLa, pLeh... |
| A2 | VC | | ap, ehp, ab, ehb...<br>aNp, eNp, aIp, aUp, aNUp...<br>iS, eS, iNS,eNS, iUS, eUS, aNUS...<br>iR, eR... |
| A3 | VC(#)V | (#) = fronteira de palavra(opcional)<br>C = R | iR(#)i, iR(#)u...<br>aIRi, aIRu...<br>iRE, iRU...<br>aIRI, aIRU... |
| A4 | $C_1C$ | $C_1$ = S,R | Sp, Sb<br>Rp, Rb... |
| A5 | VC#V | #=fronteira de palavra<br>C = S | iS#i, iS#u...<br>aIS#i, aIS#u, aNS#i, aNUS#i... |
| A6 | V/ | / = final de enunciado | i/, e/ a/...<br>aI/, aU/, iN/, aN/, aNU/... |
| A7 | VC/ | / = final de enunciado<br>C = S,R | iS/, uS/,iNS/, uNS/, aNUS/<br>iR/, eR/... |
| A8 | /V | / = início de enunciado | /i, /e, /a... |

**Grupo B – Pós-tônicas**

**( V = vogais e grupos vocálicos, C = consoantes e grupos consonantais)**

| Código do grupo | Estrutura do grupo | Observações | Exemplos |
|---|---|---|---|
| B1 | CV | | pI, pA, bI, bA...<br>pRI, pLA, pLI, pLA... |
| B2 | V(#)C | (#) = fronteira de palavra(opcional) | I(#)p, U(#)p...<br>IO(#)p, UO(#)p, IN(#)p, ANU(#)p...<br>IS, US, INS, UNS, ANUS...<br>IR#, ER#, AR# |
| B3 | VC(#)V | (#) = fronteira de palavra(opcional)<br>C = R | IRE#, URE#..<br>IR#i, IR#u... |
| B4 | VN#V | #=fronteira de palavra | IN#i, EN#i... |
| B5 | $C_1$#C | #=fronteira de palavra<br>$C_1$ = R, S, L | S#p. S#b..<br>R#p, R#b...<br>L#P, L#b... |
| B6 | VC(#)V | (#) = fronteira de palavra(opcional)<br>C = S, L | IS(#)i, AS(#)i, IS(#)u, AS(#)u...<br>INS(#)i, ANS(#)i, ANUS(#)i...<br>IL#i, AL#i... |
| B7 | V/ | / = final de enunciado | I/, A/, U/..<br>IO/, UO/, IN/, AN/, ANU/ |
| B8 | VC/ | / = final de enunciado<br>C = R, S, L | IS/, AS/, IOS/, UOS/, ANS/, ANUS/...<br>IR/, ER/...<br>IL/, EL/... |

**Grupo C – Hiatos e tritongos**

**( V = vogais e grupos vocálicos)**

| Código do Grupo | Estrutura do Grupo | Observações | Exemplos |
|---|---|---|---|
| C1 | VV internas | | iE, iO, oE.. |
| C2 | V(N)#V | #=fronteira de palavra | a#i, a#u, A#i, A#u...<br>aN#i, aN#u, AN#i, NA#u.. |
| C3 | VVV internas | | eIA, ehIA, aIA... |
| C4 | V(N)#V | #=fronteira de palavra | uI#i, iU#i, aNU#i...<br>IO#i, EU#i, ANU#i |

**Grupo D – Outros**

**( V = vogais e grupos vocálicos, C = consoantes e grupos consonantais)**

| Código do Grupo | Estrutura do Grupo | Observações | Exemplos |
|---|---|---|---|
| D1 | /C | | /p, /b, /f... |
| D2 | CV | Encontros consonantais raros | vLa, dLa, tLe, tLeh, tLaN, tLA |
| D3 | “kU”V | | kUi, kUe, kUeh... |
| D4 | “gU”V | | gUi, gUe, gUeh.. |

# Apêndice V - Regras de formação dos polifones

// Simbolos:

// c: consoantes (p,b,f,v,m,t,d,s,z,n,r,l,k,g,sh,zh,nh,lh)

// v: vogais tonicas (i,e,eh,a,oh,o,u)

// V: vogais pos-tonicas (I,E,A,O,U)

// W: vogais assilabicas (I,U)

// U: L c/ som de U (U) ou U de gua, gue, gui, kua, kue kui

// N: traco nasal (N)

// S: "s" reduzido (S)

// R: "r" reduzido (R)

// L: "l"reduzido (L)

// G: "k" e "c" em unidades do tiop gUa, gUe, gUi, kUa, kUe, kUi...

// /: pausa (//)

// Os fonemas que podem ter mais de 1 símbolo a eles

// associados sao I e U.

SIMBOLOS:

/ = $

i = v

e = v

eh = v

a = v

oh = v

o = v

u = v

I = V,W

E = V

A = V

O = V

U = V,W,U

N = N

S = S

R = R

L = L

p = c

b = c

f = c

v = c

m = c

t = c

d = c

s = c

z = c

n = c

l = c

r = c

k = c,G

g = c,G

sh = c

zh = c

nh = c

lh = c

REGRAS:

cv //A1

cRv

cLv

vc //A2.1

vWc

vNc

vNWc

vS //A2.2

vWS

vNS

vNWS

vR //A2.3

vRv //A3.1

vWRv //A3.2

vRV //A3.3

vWRV //A3.4

Sc //A4.1.1

Rc //A4.1.2

vSv //A5

vWSv

vNSv

vNWSv

v$ //A6

vW$

vN$

vNW$

vS$  //A7.1

vWS$

vNS$

vNWS$

vR$  //A7.2

$v  //A8

cV  //B1

cLV

cRV

Vc  //B2.1

WVc  //B2.2

VNc

VNWc

VS //B2.3

WVS

VNS

VNWS

VR //B2.4

VU

VRV //B3.1

VRv //B3.2

//B4 está contido em C2

//Sc //B5.1

//Rc

Uc

VSv //B6.1

WVSv

VNSv

VNWSv

VUv

V$  //B7.1

WV$  //B8.2

VN$

VNW$

VS$  //B8.1

WVS$  //B8.2

VNS$

VNWS$

VR$  //B8.3

VU$

vV  //C1

vv  //C2.1

vNv

Vv  //C2.2

VNv

vWV  //C3

vWv  //C4.1

vNWv

WVv  //C4.2

VNWv

$c  //D1

GUv  //D2

$$  //duplo silencio p/ substituir polifones invalidos

END.

# Bibliografia


[1] Albano, E., Moreira, A.A., Silva, A.H.P., Aquino, P.A., Kakinohana, R.K.; "Um conversor ortográfico-fônico e uma notação prosódica mínima para síntese de fala em língua portuguesa"; a sair em E.Scarpa, Estudos de prosódia do Brasil, Editora da UNICAMP.

[2] Albano, E., Aquino, P.A.; "Linguistic criteria for building and recording units for concatenative speech synthesis in Brazilian Portuguese"; 5th European Conference on Speech Communication and Technology (Eurospeech 97), vol.2, pp.729-732, 1997.

[3] Alcaim, A., Solewicz, J.A., Moraes, J.A.; “Freqüência de ocorrência dos fones e listas de frases foneticamente balanceadas no português falado no Rio de Janeiro”; Revista da Sociedade Brasileira de Telecomunicações, vol. 7, n°1, pp. 23-41, Dezembro de 1992.

[4] Allen, J., Hunnicut, S., Klatt, D.H.; *From text to speech: the MITalk system*; Cambridge, UK, 1987.

[5] Allen, J.; "Synthesis from unrestricted text"; em *Speech Synthesis*, Flanagan, J.L. & Rabiner, L.R., Eds. Dowden, Hutchinson & Ross, Inc. Strousbourg, Pennsilvania, 1975.

[6] Aquino, P.; "O papel das vogais reduzidas pós-tônicas na construção de um sistema de síntese concatenativa para o português do Brasil"; Tese de mestrado, Instituto de Estudos da Linguagem da UNICAMP, Dezembro de 1997.

[7] Atal, B.S., "Linear predictive coding of speech" em *Computer Speech Processing*, Fallside,F., Woods, W.A., Prentice-Hall International, University of Cambridge, 1983.

[8] Atal, B.S.; "Automatic recognition of speaker from their voices"; Proceedings of the IEEE 64(4), pp. 460-475, Abril de 1976.

[9] Barbosa, P. A.; "A model of segment (and pause) duration generation for Brazilian Portuguese text-to-speech synthesis"; 5th European Conference on Speech Communication and Technology (Eurospeech 97), vol.5, pp.2655-2658, 1997.

[10] Barbosa, P.A.; "At least two macrorhythmic units are necessary for modeling Brazilian Portuguese duration: emphasis on segmental duration generation"; *Caderno de Estudos Lingüísticos*, 31, pp.33-53, UNICAMP, 1996.

[11] Barbosa, P.A.; "Revelar a estrutura rítmica de uma língua construindo máquinas falantes: pela integração de ciência e tecnologia de fala"; a sair em E.Scarpa, *Estudos de prosódia do Brasil*, Editora da UNICAMP, abril de 1999.

[12] Campbell, W.N., Isard, D.; "Segmental duration in syllable frames"; Journal of Phonetics 19, pp. 37-47, 1991.

[13] Capovilla, F.C.; "Sistemas especialistas de multimídia em educação especial" em *Prevenção e intervenção em educação especial*, Nunes, L.R. (org), vol 1(4), pp.124-150, Rio de Janeiro, 1996.

[14] Charpentier, F. ;"Traitement de la parole par analyse-synthèse de Fourier – Aplication à la synthèse par diphones"; Tese de doutorado, École Nationale Supérieure de Télécommunications, Julho de 1989.

[15] Cocker, C.H.; "A model of articulatory dynamics and control"; Proceedings of the

IEEE 64(4), pp. 452-460, Abril de 1976.

[16] Cooper, F.S., Liberman, A.M., Borst, J.M.; "The interconversion of audible and visible patterns as a basis for research in the perception of speech"; Proc. Natl. Acad. Sci. (US) 37, pp. 318-325, 1951.

[17] Degen, J.V.; *Mechanismus der menschlichen Sprache nebst der Beschreibung einer sprechenden Maschine - Le Mechanisme de la parole, suivi de la description d'une machine parlante*; 1791.

[18] Dudley, H., Riesz, R.P., Watkins, S.A.; "A synthetic speaker", J. Franklin Institute 227, pp. 739-764, 1939.

[19] Dudley, H.; "The vocoder"; Bell Labs Rec. 17, pp. 122-126, 1939.

[20] Dutoit, T., Leich, H.; “MBR-PSOLA: Text to Speech synthesis based on a MBE resynthesis of the segments database”; Speech Communication 13, pp. 435-440, 1993.

[21] Dutoit, T.; *An Introduction to Text-to-Speech Synthesis*; Kluwer Academic Publishers, 1997.

[22] 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.

[23] Fant, G.; "Speech communication research"; Ing. Vetenkaps Akad. Stocholm (Suécia) 24, pp. 331-337, 1953.

[24] Fant, G.; *Acoustic Theory of Speech Production*; Mouton's Gravenhague, 1960.

[25] Ferreira, A.; *Minidicionário Aurélio*; Rio de Janeiro; Nova Fronteira, 1977.

[26] Figueiredo, F.A., Naviner, L.A.B., Neto, B.G.A.; "Uma nova abordagem para o

sistema de conversão texto-fala para a língua portuguesa"; Universidade Federal da Paraíba, Anais do XV Simpósio Brasileiro de Telecomunicações, Recife, Setembro de 1997.

[27] Figueiredo, F.A., Neto, M.L.C., Naviner, L.A.B., Azevedo, J.A., Neto, B.G.A.; "Analisador de texto para um sistema de conversão texto-fala para a língua portuguesa"; Anais do III Encontro para o Processamento Computacional de Português Escrito e Falado (PROPOR'98), pp. 17-22, Novembro de 1998.

[28] Figueiredo, R. M.; "Identificação de falantes: aspectos teóricos e metodológicos"; Tese de Doutorado, LAFAPE/IEL/UNICAMP, março de 1994.

[29] Flanagan, J.L., Cocker, C.H., Rabiner, L.R., Schafer, R.W., Umeda, N.; "Synthetic voices for computers"; IEEE Spectrum 7(10), pp. 22-45, outubro de 1970.

[30] Flanagan, J.L.; "Computers that talk and listen: man-machine communication by voice"; Proceedings of the IEEE 64(40), pp.405-415, Abril de 1976.

[31] Fujimura, O., Lovins, J.; "Syllables as concatenative phonetic elements" em *Syllables and Segments*, Bell, A. & Hooper, J.B. (North Holland, N. York), pp. 107-120, 1978.

[32] Harris, C. M.; "A study of the building blocks in Speech"; Journal of the Acoustical Society of America 25, pp. 962-969, Maio de 1953.

[33] Holmes, J.N.; "The influence of the glottal waveform on the naturalness of speech from a parallel formant synthesizer"; IEEE Trans. Audio Electroacoust. AU-21, pp. 298-305, 1973.

[34] Kay, M.; "Machine Translation: The disappointing past and present"; *Survey of the State of the Art in Human Language Technology*, pp. 285-287, novembro de 1995.

[35] Kent, R.D., Read, C.; *The Acoustic Analysis of Speech*; Singular Publishing Group, Inc., San Diego, 1992.

[36] Klatt, D.H., "Review of text to speech conversion for English"; Journal of the Acoustical Society of America 82(3), pp 737-792, Setembro de 1987.

[37] Klatt, D.H., Klatt, L.C.; "Analysis, synthesis and perception of voice quality variations among female and male talkers"; Journal of the Acoustical Society of America 87(2), pp. 820-875, Fevereiro de 1990.

[38] Klatt, D.H.; "Software for a cascade/parallel formant synthesizer"; Journal of the Acoustical Society of America 67, pp. 971-995, 1980.

[39] Kurzweil, R.; "The Kurzweil reading machine: a technical overview" in *Science, Technology and the Handicapped*, editado por Redden, M.R. & Schwandt, (American Association for the Advancement of Science, Report 76-R-11, Washington D.C.), pp. 3-11, 1976.

[40] Lawrence, W.; "The synthesis of speech from signals which have a low information rate"; em *Communication Theory*, editado por Jackson, W. (Butterworths, Londres-Inglaterra), pp. 460-469, 1953.

[41] Lee, K.F., *Automatic Speech Recognition*, Kluwer Academic Publishers, 1989.

[42] Madureira, S.; “Entoação e síntese de fala: modelos e parâmetros”; a sair em E.Scarpa, *Estudos de prosódia do Brasil*, Editora da UNICAMP.

[43] Madureira, S.; “Pitch patterns in Brazilian Portuguese: na acoustic phonetics analysis”; Vth Australian International Conference of Speech Science and Technology, pp. 156-161, 5 a 9 de Dezembro, Perth, Austrália, 1994.

[44] Marques, J.S., Almeida, L.B.; "Sinusoidal modeling of voiced and unvoiced speech"; Proc. Eurospeech, Vol II, pp. 203-206, 1989.

[45] McAulay, R.J., Quartieri, T.F.; "Speech analysis/synthesis based on a sinusoidal representation"; IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-34 (4), pp.744-754, 1986.

[46] Mermelstein,P.; "Articulatory model for the study of speech production"; Journal of the Acoustical Society of America 53(4), pp 1070-1082, Julho de 1973.

[47] Moulines, E.; "Algorithmes de codage et de modification des paramètres prosodiques pour la synthèse de la parole à partir du texte"; Tese de Doutorado, École National Supérieure des Télécommunications, Fevereiro de 1990.

[48] Muthusamy, Y.K., Spitz, L.; "Automatic Language Identification"; *Survey of the State of the Art in Human Language Technology*, pp. 314-317, novembro de 1995.

[49] Olive, J.P., Nakatani, L.H.; "Rule-synthesis of speech by word concatenation: a first step"; Journal of the Acoustical Society of America 55(3), pp. 660-666, Março de 1974.

[50] Oliveira, L.M.V.V.C.; "Síntese de fala a partir de texto"; Tese de Doutorado, Universidade Técnica de Lisboa, Outubro de 1996.

[51] Pessoa, L.A.S.; "Modelos da língua para o português do Brasil aplicados ao reconhecimento de fala contínua: modelos lineares e modelos hierárquicos (*parsing*)"; Tese de Mestrado, Faculdade de Engenharia Elétrica e de Computação da UNICAMP, Fevereiro de 1999.

[52] Portele, T., Höfer, F., Hess, W.J.; “A mixed inventory structure for german concatenative synthesis” em *Progress in Speech Synthesis,* van Santen, J.P.H.,

Sproat, R.W., Olive, J.P., Hirschber, J., Springer-Verlag New York Inc., 1997.

[53] Rabiner, L., Juang B.-H.; *Fundamentals of Speech Recognition*; Englewood Cliffs, NJ, Prentice-Hall, 1993.

[54] Rabiner, L.R., Schafer, R.W.; *Digital Processing of Speech Signals*; Englewood Cliffs, NJ, Prentice-Hall, 1978.

[55] Rabiner, L.R.; "Applications of voice processing to telecommunications"; Proceedings of the IEEE 82(2), pp. 199-228, Fevereiro de 1994.

[56] Rosenberg, A.E., "Automatic speaker verification: a review"; Proceedings of the IEEE 64(4), pp. 475-487, Abril de 1976.

[57] Rubin, P., Baer, T., Mermelstein, P.; "An articulatory synthesizer for perceptual research"; Journal of the Acoustical Society of America 70(2), pp. 321-328, Agosto de 1981.

[58] Ryley, M.D.; "Free-based modeling for speech synthesis" in *Talking machines: theories, models and design*, Bailly, G. & Benoit, C., pp. 265-273, North Holland, 1992.

[59] Schäfer-Vincent, K.; "Pitch period detection and chaining: method and evaluation"; Phonetica, 40, 177-202, 1983.

[60] Sejnowski, T. J.; "Parallel networks that learn to pronounce English text"; Complex Systems 1, pp. 145-168, 1987.

[61] Silva, C.H., Nagle, E.J., Nunes, H.F.; "Automação da construção de dicionários de unidades acústicas para conversão texto-fala por concatenação", Anais do XV Simpósio Brasileiro de Telecomunicações, Recife, Setembro de 1997.

[62] Silva, C.H.; "Modelamento prosódico para conversão texto-fala do português falado no Brasil"; Tese de Mestrado, Faculdade de Engenharia Elétrica da UNICAMP, Dezembro de 1995.

[63] Solewicz, J. A.; "Síntese de voz a partir de texto para o português do Brasil"; Tese de mestrado, PUC-RJ,1993.

[64] Sousa, E. M. G.; “Towards an acoustic description of Brazilian Portuguese nasal vowels”; XIII International Congress of Phonetic Sciences, 13 a 19 de Agosto, Estocolmo, Suécia, 1995.

[65] Sproat, R.; *Multilingual text to speech synthesis – the Bell Labs approach*; Bell Laboratories/Lucent Technologies, Kuwler Academic Publishers, Dordrecht/Boston/London, 1998.

[66] Stanton, A.G.A.; “Iniciação científica em sistemas de conversão texto-fala”; relatório de bolsa pesquisa SAE, LPDF-DECOM-FEEC-UNICAMP, 2° semestre de 1997 / 1° semestre de 1998.

[67] Stella, M.; “Speech synthesis" em *Computer Speech Processing*, Fallside, F., Woods, W. A., Prentice-Hall International; University of Cambridge, 1983.

[68] Stewart, J.Q.; "An electrical analogue of the vocal organs"; Nature 110, pp.311-312, 1922.

[69] Vaissière, J.; "Speech recognition: a tutorial" em *Computer Speech Processing*, Fallside, F., Woods, W.A., Prentice-Hall International, University of Cambridge, 1983.

[70] Violaro, F., Böeffard, O.; "A hybrid model for text to speech synthesis", IEEE Transactions on Speech and Audio Processing 6(5), pp 426-434, 1998.

[71] Wang, W. S-Y., Peterson, G.E.; "Segment inventory for speech synthesis"; Journal of the Acoustical Society of America 30(8), pp. 743-746, Agosto de 1958.

[72] Witten, I. W.; *Principles of Computer Speech*; Academic Press Inc., 1982.

[73] Young, S.J., Fallside, F.; "Speech synthesis from concept: A method for speech output from information systems"; Journal of the Acoustical Society of America 66(3), pp. 685-695, Setembro de 1979.