FERNANDO SANTANA PACHECO

TÉCNICAS DE PROCESSAMENTO DE SINAIS PARA ALTERAÇÃO DE PARÂMETROS PROSÓDICOS APLICADAS A UM SISTEMA DE CONVERSÃO TEXTO-FALA PARA A LÍNGUA PORTUGUESA FALADA NO BRASIL

FLORIANÓPOLIS
2001

UNIVERSIDADE FEDERAL DE SANTA CATARINA

PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA

TÉCNICAS DE PROCESSAMENTO DE SINAIS PARA ALTERAÇÃO DE PARÂMETROS PROSÓDICOS APLICADAS A UM SISTEMA DE CONVERSÃO TEXTO-FALA PARA A LÍNGUA PORTUGUESA FALADA NO BRASIL

Dissertação submetida à
Universidade Federal de Santa Catarina
como parte dos requisitos para a
obtenção do grau de Mestre em Engenharia Elétrica.

FERNANDO SANTANA PACHECO

Florianópolis, Abril de 2001.

TÉCNICAS DE PROCESSAMENTO DE SINAIS PARA ALTERAÇÃO DE PARÂMETROS PROSÓDICOS APLICADAS A UM SISTEMA DE CONVERSÃO TEXTO-FALA PARA A LÍNGUA PORTUGUESA FALADA NO BRASIL

Fernando Santana Pacheco

'Esta Dissertação foi julgada adequada para obtenção do Título de Mestre em Engenharia Elétrica, Área de Concentração em *Circuitos e Instrumentação Eletrônica*, e aprovada em sua forma final pelo Programa de Pós-Graduação em Engenharia Elétrica da Universidade Federal de Santa Catarina.'

Prof. Rui Seara, Dr.
Orientador

Prof. Aguinaldo Silveira e Silva, Ph.D.
Coordenador do Programa de Pós-Graduação em Engenharia Elétrica

Banca Examinadora:

Prof. Rui Seara, Dr.
Presidente

Prof. Hans Helmut Zürn, Ph.D.

Eng. Orlando José Tobias, Dr.

Prof. Sidnei Noceti Filho, Dr.

À minha família

AGRADECIMENTOS

A Deus, por sempre me guiar pelos melhores caminhos.

A toda minha família, pelo incentivo e fé no sucesso.

À Gisele, pela alegria.

Ao Prof. Rui, pela orientação primorosa.

Aos membros da banca, pelas valiosas correções e sugestões.

À equipe de desenvolvimento do TTS: Neco, Sandra, Simone e Izabel, pelo excepcional convívio.

Ao Eng. Walter.

Ao Elton e a todos os amigos do LINSE.

Ao CNPq, pelo apoio financeiro.

A todos que, de alguma forma, contribuíram para a realização deste trabalho.

Resumo da Dissertação apresentada à UFSC como parte dos requisitos necessários para a obtenção do grau de Mestre em Engenharia Elétrica.

TÉCNICAS DE PROCESSAMENTO DE SINAIS PARA ALTERAÇÃO DE PARÂMETROS PROSÓDICOS APLICADAS A UM SISTEMA DE CONVERSÃO TEXTO-FALA PARA A LÍNGUA PORTUGUESA FALADA NO BRASIL

Fernando Santana Pacheco

Abril/2001

Orientador: Prof. Rui Seara, Dr.
Área de Concentração: Circuitos e Instrumentação Eletrônica.
Palavras-chave: Processamento de Sinais; Conversão Texto-Fala; Parâmetros Prosódicos; Mudança de *Pitch*; Síntese de Fala.
Número de Páginas: 100.

RESUMO: Sistemas de conversão texto-fala têm como objetivo a transformação de um texto com vocabulário irrestrito em uma mensagem falada. Esse processo consiste de duas etapas básicas. Na primeira, técnicas de processamento lingüístico realizam a extração de uma representação simbólica dos parâmetros acústicos a partir do texto de entrada. A representação simbólica é transformada em sinal de fala através de técnicas de processamento de sinais. Um dos métodos de síntese de fala é o de concatenação de segmentos de fala previamente gravados. No entanto, para conferir maior naturalidade à fala sintetizada, faz-se necessário alterar de forma dinâmica os parâmetros prosódicos (*pitch*, duração e energia) dos segmentos durante a operação de síntese. O presente trabalho apresenta o desenvolvimento de uma técnica baseada em análise/ressíntese LPC com excitação residual para alteração de parâmetros prosódicos. O objetivo é aplicá-la a um sistema de conversão texto-fala baseado em síntese concatenativa para a língua portuguesa falada no Brasil. Nesta técnica, simples operações de cópia e corte são realizadas no sinal de resíduo, permitindo a alteração do *pitch*. A alteração da duração é efetuada eliminando ou copiando quadros inteiros de análise. Essa técnica apresenta uma carga computacional reduzida, possibilitando a implementação em tempo real. Análises objetivas e testes perceptuais preliminares mostraram um bom desempenho da técnica.

Abstract of Dissertation presented to UFSC as a partial fulfillment of the requirements for the degree of Master in Electrical Engineering.

SIGNAL PROCESSING TECHNIQUES FOR PROSODIC MODIFICATION OF SPEECH APPLIED TO A TEXT-TO-SPEECH SYSTEM FOR PORTUGUESE LANGUAGE SPOKEN IN BRAZIL

Fernando Santana Pacheco

April/2001

Advisor: Prof. Rui Seara, Dr.
Area of Concentration: Circuits and Electronic Instrumentation.
Keywords: Speech Processing; Speech Synthesis; Text-to-Speech Systems; Prosodic Modification; Voice Response Systems.
Number of Pages: 100.

ABSTRACT: The aim of text-to-speech synthesizers is to convert unrestricted text input to speech. The procedure consists of two main stages. In the first one, the input text is converted to a symbolic representation, by means of natural language processing. In the second stage, that representation is transformed into speech by use of signal processing techniques. One common speech synthesis method is based on concatenation of pre-recorded speech segments. To allow more naturalness and intelligibility in the synthetic speech, dynamic prosodic modification is required during the synthesis operation. This work presents the development of a prosodic modification technique based on residual-excited LPC processing, applied to a text-to-speech system for the Portuguese language spoken in Brazil. In that technique, simple copy and cut operations are made on the residual signal, allowing pitch modification. Changing the duration of segments is performed by copy or exclusion of frames. This technique offers a low computational load, and real-time implementation is achieved. Objective analysis and perceptual tests have shown a good performance for the proposed approach.

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

Capítulo 1

Introdução

1.1 Considerações Iniciais

—Parla!

Assim tentou Michelangelo dar voz ao mármore inanimado, ao concluir a escultura de Moisés. O mármore não respondeu... E o desejo de dar fala a um objeto, a uma máquina, só seria realizado alguns séculos depois. Os primeiros sistemas de produção de fala artificial surgiram no século XVIII. Eram mecânicos, difíceis de operar e não geravam mais do que alguns poucos sons da fala. No entanto, serviram como ferramentas de experimentação para o estudo do mecanismo de produção da fala. Com o avanço tecnológico, sistemas eletro-eletrônicos e *softwares* de síntese de fala foram sendo desenvolvidos. Mas, somente na década de 60, foi possível gerar fala a partir de um texto. A idéia, que no início parecia uma brincadeira, foi tomando corpo e encontra um extenso campo de aplicações no mundo atual.

A escrita tem um papel fundamental como forma de comunicação humana. Entretanto, isso não significa que a mensagem escrita seja sempre a forma mais conveniente de se obter acesso a informações [1]. Em diversas circunstâncias, não se pode parar a atividade que se está realizando para ler um texto. Mas pode-se ouvi-lo, se for falado de forma correta e agradável. Por exemplo, ao dirigir não se pode desviar a atenção dos olhos e mãos para ler o jornal, mas pode-se ouvir as notícias no rádio. Na interação homem-máquina, mensagens de alerta faladas possivelmente são mais eficientes do que respostas visuais. Na cabine de comando de um avião, com centenas de instrumentos, a

forma mais conveniente de emitir avisos críticos é através da resposta sonora. Além disso, a resposta falada pode dar um aspecto mais humano à interação com a máquina [1].

Usando um telefone comum, a única forma de acesso a informações é através da interação vocal. Sistemas que realizem a passagem do domínio fala para texto e vice-versa permitem o desenvolvimento de diversas aplicações em que o único meio de entrada e saída é a fala. O acesso a informações como saldo bancário, previsão de tempo e acompanhamento de processos torna-se viável usando apenas o telefone.

Com estes exemplos, fica clara a necessidade de um processo automático de transformação de informações escritas em mensagens faladas. Esse mapeamento do texto para a fala é o objetivo dos sistemas de síntese de fala.

Os sistemas de síntese de fala podem ser divididos em duas classes, definidas pelo tamanho do vocabulário e pelo campo de aplicação. Na primeira classe estão os sistemas utilizados em aplicações que requerem pouca interação com o usuário, representados pelos sistemas de resposta vocal. Na segunda, a necessidade de interação com o usuário é alta, exigindo a utilização de sistemas de conversão texto-fala (*text-to-speech systems*). Os sistemas de resposta vocal operam com um vocabulário limitado [1] em aplicações, por exemplo, de serviços telefônicos como hora certa e despertador automático. Em uma primeira etapa, as mensagens requeridas para o serviço são definidas, gravadas e armazenadas. A operação de síntese de fala é realizada pela simples combinação e reprodução do que foi gravado. Em um sistema de saldo bancário, por exemplo, frases introdutórias como “bom dia”, “boa tarde”, “digite sua senha”, “seu saldo é” e palavras básicas para a formação dos valores monetários como “um”, “cem”, “mil” são combinadas de forma adequada para a geração da resposta falada. Como vantagens dessa técnica, pode-se citar a alta qualidade que pode ser atingida e a pequena carga de processamento. Entretanto, o domínio é restrito e bem definido e a capacidade de armazenamento das mensagens é limitada pela memória disponível do sistema.

Os sistemas de conversão texto-fala produzem fala sintetizada a partir de um texto de entrada com vocabulário irrestrito. Como o vocabulário é ilimitado, não é possível armazenar todas as combinações possíveis de palavras para posterior reprodução. A solução é realizar, inicialmente, uma análise de texto que identifique os sons correspondentes à representação escrita e associe parâmetros de entonação e ritmo. Em um segundo passo, a transformação dessa representação simbólica intermediária em sinal de fala é efetuada através de técnicas de processamento de sinais. Problemas ocorrem nas duas etapas: a análise de texto é uma tarefa difícil, pois nem sempre a mensagem escrita permite a especificação de todas as informações importantes para a fala, e a síntese do sinal, limitada por aspectos como a complexidade computacional, usualmente não permite a produção de fala com a mesma qualidade da natural.

A avaliação dos métodos de síntese de fala em diferentes aplicações é realizada através de três parâmetros básicos [2]:

a) a qualidade, medida subjetivamente em termos de inteligibilidade e naturalidade;
b) a flexibilidade, relacionada à capacidade de síntese de mensagens com diferentes palavras e diferentes entonações, velocidades e ênfases;
c) a complexidade, medida em relação à carga de processamento computacional e capacidade de armazenamento requerida.

A interação destes três fatores em um espaço tridimensional e a localização das duas abordagens apresentadas são ilustradas na Fig. 1.1. O sistema ideal proveria uma saída de alta qualidade, praticamente indistinguível da fala natural; produziria mensagens com qualquer padrão de entonação e ritmo de forma adequada; teria baixa complexidade para permitir a integração a um pequeno custo em qualquer ambiente de aplicação. Infelizmente, não há nenhum sistema nos dias atuais que atenda completamente esses três requisitos. Os sistemas de resposta vocal têm baixa complexidade e alta qualidade, mas não são capazes de lidar com texto irrestrito. Os sistemas *text-to-speech* (TTS), por sua

vez, têm um custo computacional mais elevado e uma qualidade mais baixa. Mas são a única alternativa para a transformação de qualquer texto em uma representação falada.

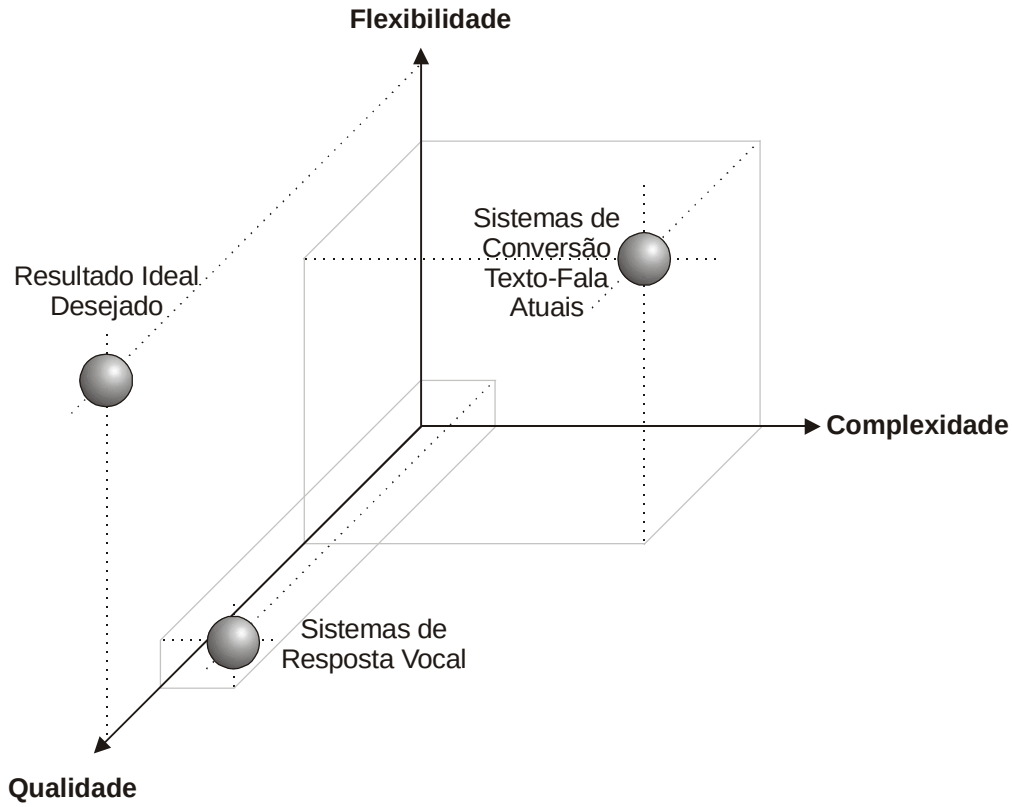

Fig. 1.1: Técnicas de síntese de fala em termos de flexibilidade, qualidade e complexidade.

1.2 Justificativa e Objetivos do Trabalho

Para a língua portuguesa falada no Brasil, já existem sistemas de conversão texto-fala, como os desenvolvidos em [1,3]. A avaliação desses sistemas, feita na época em que foram desenvolvidos, indicava uma qualidade satisfatória. Entretanto, com o vislumbre de aplicações cada vez mais sofisticadas, o grau de exigência flexibilidade-qualidade-complexidade tem aumentado sobremaneira. É importante, então, estudar, desenvolver e

aperfeiçoar métodos e técnicas que, com relação aos parâmetros mostrados na Fig. 1.1, estejam posicionados mais próximos ao sistema ideal.

A alteração dos parâmetros prosódicos[1] dos segmentos durante a operação de síntese é necessária para conferir, principalmente, maior naturalidade à fala sintetizada. Diversas técnicas, utilizando diferentes modelagens, se propõem a efetuar essas alterações [4-10]. De modo similar ao exposto para as técnicas de síntese, três requisitos básicos estão associados a uma técnica de modificação prosódica:

a) a flexibilidade das alterações, permitindo o atendimento das exigências de variação dos parâmetros prosódicos imposta pelo módulo de processamento lingüístico;
b) a manutenção da qualidade dos segmentos, ou seja, a não inserção de degradações perceptíveis (como ruídos) pelo processo de alteração;
c) a exigência de uma carga computacional reduzida, para possibilitar a implementação em tempo real.

O presente trabalho tem os seguintes objetivos:

a) a apresentação e discussão de técnicas de processamento de sinais que permitam a alteração dos parâmetros prosódicos *pitch* e duração;
b) o desenvolvimento de uma nova técnica de alteração prosódica a ser aplicada em um ambiente de síntese de fala por concatenação de segmentos para a língua portuguesa falada no Brasil.

1.3 Estrutura

No Capítulo 1, será apresentada uma visão geral dos sistemas de conversão texto-fala, incluindo possíveis aplicações. Ainda nesse capítulo, uma revisão histórica do desenvolvimento dos sistemas de síntese de fala mostrará a evolução desses sistemas.

[1] Agrupados no estudo de prosódia estão os processos lingüísticos relacionados à entonação, ao acento e ao ritmo da fala, acusticamente percebidos pela variação dos parâmetros prosódicos *pitch*, quantidade e volume. O Capítulo 2 tratará do aspecto prosódico da fala.

No Capítulo 2, será apresentado o processo de produção da fala pelo aparelho fonador humano, considerando também os aspectos relacionados à prosódia e à modelagem do sistema.

O Capítulo 3 apresenta a primeira etapa de processamento dos sistemas de conversão texto-fala, o estágio de análise de texto.

As estratégias de síntese do sinal de fala são apresentadas no Capítulo 4. Modelos baseados nas abordagens de sinal e sistema são discutidos.

Técnicas de alteração de parâmetros prosódicos baseadas em diferentes modelos são apresentadas no Capítulo 5.

No Capítulo 6 é proposta uma técnica de alteração de parâmetros prosódicos baseada na modificação do sinal residual da análise LPC, que estabelece um satisfatório compromisso entre custo computacional e qualidade.

No Capítulo 7 são apresentadas conclusões e propostas para trabalhos futuros.

1.4 Visão Geral dos Sistemas de Conversão Texto-Fala

Os sistemas de conversão texto-fala têm por objetivo a transformação de *qualquer* texto em fala. Realizam a produção de fala por meio da "fonetização" automática do texto [11]. Esse processo é bastante complexo e só pode ser resolvido de forma multidisciplinar. Os conhecimentos envolvidos na resolução do problema levam a uma divisão praticamente natural do processo em duas etapas:

a) passagem do domínio texto para um domínio de representação intermediário, baseada em técnicas de processamento de linguagem natural;
b) passagem do domínio intermediário para o domínio acústico do sinal de fala, baseada em técnicas de processamento de sinais.

Na Fig. 1.2, é mostrado um diagrama dos blocos fundamentais de processamento envolvidos na tarefa de conversão texto-fala. Na primeira etapa, em que estão relacionados fundamentalmente aspectos lingüísticos, o texto é analisado, sendo gerada uma

representação fonética associada a informações prosódicas da fala que será sintetizada. Esse estágio de processamento é fortemente dependente do idioma a que se propõe o sistema de conversão e envolve, dentre outros, módulos de:

a) pré-processamento do texto de entrada, com a separação de blocos de análise, identificação e expansão de abreviaturas, siglas, algarismos;
b) transcrição ortográfico-fonética;
c) separação silábica e determinação da tonicidade;
d) análise sintática, com a classificação gramatical das palavras;
e) modelagem prosódica, que determina padrões de entonação e ritmo, acusticamente relacionados à freqüência fundamental, duração e intensidade do sinal.

Ao final do processamento lingüístico, os sons que devem ser sintetizados estão definidos. Por exemplo, os sons da palavra “casa” são, numa transcrição fonética, [k], [a], [z] e [e]. Além desse nível de descrição fonética, parâmetros prosódicos estarão associados aos sons.

A síntese propriamente dita do sinal de fala é realizada na etapa de processamento de sinais. Um modelo de síntese deve permitir a geração dos sons e a alteração dos parâmetros prosódicos de acordo com o que foi prescrito na etapa de análise lingüística. Os modelos que realizam a síntese podem ser classificados em dois paradigmas, de acordo com o domínio em que atuam [11]:

a) abordagem de sistema. Também chamada de síntese articulatória, nessa abordagem o próprio mecanismo de produção da fala é modelado, com maior ou menor detalhamento fisiológico;
b) abordagem de sinal. Também conhecida como *terminal-analogue synthesis*, modela o próprio sinal de fala, utilizando quaisquer meios convenientes. Oposta à abordagem de sistema, não implica na modelagem dos gestos

articulatórios e sim na representação do sinal acústico gerado pelo processo de produção da fala.

As duas abordagens evoluíram de forma independente, com resultados mais rápidos tendo sido obtidos com a modelagem do sinal, devido à relativa simplicidade [11]. Enquanto a modelagem do complexo mecanismo de produção da fala é ainda um problema a ser resolvido, técnicas no domínio do sinal, como a de síntese por formantes e por predição linear, são empregadas em sistemas comerciais desde os anos 70.

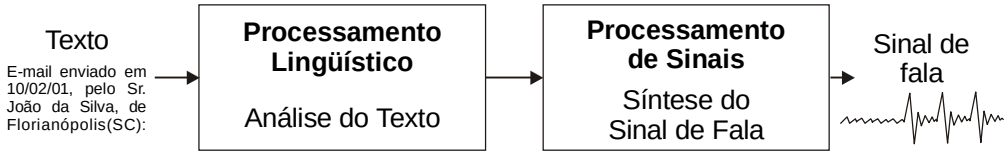

Fig. 1.2: Diagrama básico do processo de conversão de texto em fala.

Uma das técnicas da abordagem de sinal que apresenta melhores resultados é a de síntese por concatenação de segmentos de fala. Nessa técnica, segmentos do sinal de fala de tamanhos diversos são previamente gravados por um locutor e posteriormente concatenados para a geração de fala sintética. A idéia lembra um pouco a dos sistemas de resposta vocal, mas aqui os segmentos formam um conjunto que permite a síntese de qualquer texto. No processo de gravação desses segmentos, naturalmente estão associadas uma entonação e ritmo relacionados ao contexto no qual o segmento está inserido. Para conferir maior inteligibilidade e, principalmente, naturalidade à fala sintetizada, uma simples operação de concatenação dos segmentos não é suficiente. Torna-se, então, necessário modificar os parâmetros da fala associados à entonação e ao ritmo através de técnicas de processamento de sinais.

1.5 Aplicações dos Sistemas de Conversão Texto-Fala

As aplicações que requerem a utilização de sistemas de conversão texto-fala são aquelas que exigem o tratamento de texto irrestrito em ambientes de interação homem-máquina. Os sistemas de conversão texto-fala são uma alternativa interessante em situações em que [12]:

a) o texto é imprevisível e dinâmico. Existem situações em que as mensagens que se deseja sintetizar são curtas, mas o conteúdo varia significativamente e não pode ser enquadrado em um formato padrão que permita a utilização de um sistema de resposta vocal. Nesses casos, o único método viável de síntese é o de conversão texto-fala;

b) é necessário acesso a um grande banco de dados. Não é viável realizar a gravação de todo o conteúdo de grandes bancos de dados, devido aos custos de gravação e armazenagem. Além disso, as informações estão sujeitas a alterações constantes;

c) a saída é relativamente estável, mas o custo de provimento e o tempo de resposta são críticos. Em sistemas telefônicos de atendimento, algumas mensagens permanecem constantes por longos períodos, mas, em certas situações, pode ser necessário modificá-las. A manutenção da mesma voz e o curto tempo disponível para a mudança favorecem o uso de sistemas TTS, quando comparados a novas gravações;

d) a consistência da voz é requerida. Muitos sistemas requerem a manutenção da voz para todas as mensagens. Não há problemas se, uma vez operando, não houver modificações. Entretanto, se a possibilidade de melhoramentos futuros for planejada, deve-se prever a disponibilidade do mesmo locutor. Nessas situações, a utilização de sistemas de conversão texto-fala deve ser considerada;

e) pequena ocupação de banda de transmissão é necessária. A transmissão de informação através de texto e posterior conversão para fala emprega uma banda de comunicação extremamente pequena.

Apresentadas as características das aplicações alvo, pode-se citar algumas delas [1,2,11,13,14]:

a) auxílio a portadores de deficiências. Incapacidades no processo de fala têm causas mentais ou motoras. Para o caso de problemas motores, os sistemas de conversão texto-fala podem atuar como um importante suporte. Com o auxílio de um teclado especial e um programa de montagem de sentenças, a geração de fala sintetizada pode permitir a comunicação. As aulas do astrofísico Stephen Hawking são proferidas dessa forma. Pessoas com deficiência visual podem ter acesso a informações escritas em formato eletrônico através de sistemas TTS. Aqueles com incapacidades auditivas e/ou de fala podem fazer ligações telefônicas e "conversar" normalmente se em cada extremo for utilizado um sistema de conversão de texto em fala e de fala em texto (reconhecimento de fala);

b) pesquisa básica e aplicada. Sintetizadores de fala são uma ferramenta muito interessante para lingüistas, por uma característica peculiar: provêem um ambiente de total controle, permitindo que experimentos repetidos produzam resultados idênticos, o que é praticamente impossível com seres humanos. Assim, investigações relacionadas a modelos prosódicos, por exemplo, podem ser realizadas. Os sistemas TTS que são baseados nos parâmetros do trato vocal têm sido extensivamente utilizados por foneticistas para o estudo do processo de fala;

c) monitoramento com resposta vocal. Em certas situações, uma resposta vocal é mais eficiente do que uma mensagem escrita. Avisos de atenção ou perigo dados na forma falada têm um apelo mais forte. Poderiam ser utilizados, por

exemplo, quando alguém se aproximasse de equipamentos ou áreas que oferecessem risco. A sobrecarga de informações visuais nas cabines de comando de aviões poderia ser aliviada com algumas mensagens faladas;

d) ensino de idiomas. Sistemas de alta qualidade podem ser utilizados para o aprendizado de idiomas, constituindo uma ferramenta muito valiosa;

e) livros e brinquedos falantes;

f) serviços em telecomunicações. Geralmente os serviços telefônicos usam bases de dados com informações que variam constantemente, tornando adequado o emprego de sistemas de conversão texto-fala. O número de aplicações é muito grande e, dentre outras, pode-se citar:

- acesso às mensagens de correio eletrônico;
- auxílio à lista telefônica;
- informações sobre cursos, classificação em provas;
- resultados de exames médicos;
- acesso a informações como previsão meteorológica, eventos esportivos e culturais, feiras, exposições, programação de teatro e cinema;
- agenda e despertador automático;
- acesso a dicionários, enciclopédias e manuais de equipamentos;
- acompanhamento de processos ou pedidos de compras;
- informações bancárias.

1.6 Histórico

A potencialidade de aplicações de sistemas de síntese de fala despertou, há longo tempo, um interesse nessa área. Mas, só é possível uma aplicação prática se um padrão mínimo de qualidade for atingido. O acréscimo de qualidade só pode ser obtido com o desenvolvimento gradual de diversas áreas do conhecimento. O histórico apresentado a

seguir, baseado nas literaturas [1,11,15-18], mostra a evolução dos sistemas de síntese de fala.[2]

1.6.1 Sistemas Mecânicos

Uma das primeiras tentativas de geração de fala sintética ocorreu em 1779, na Academia Imperial de São Petersburgo, na Rússia. O professor Christian Kratzenstein recebeu o prêmio anual ao explicar as diferenças fisiológicas entre cinco vogais longas ([a], [e], [i], [o] e [u]) e construir uma série de ressoadores acústicos[3]. A estrutura básica desses ressoadores é mostrada na Fig. 1.3. Esses dispositivos eram similares à configuração do trato vocal humano e emitiam sons pelo uso de palhetas, como em instrumentos musicais. Na Fig. 1.4, é mostrada a configuração do trato vocal e o respectivo ressoador para a vogal [e].

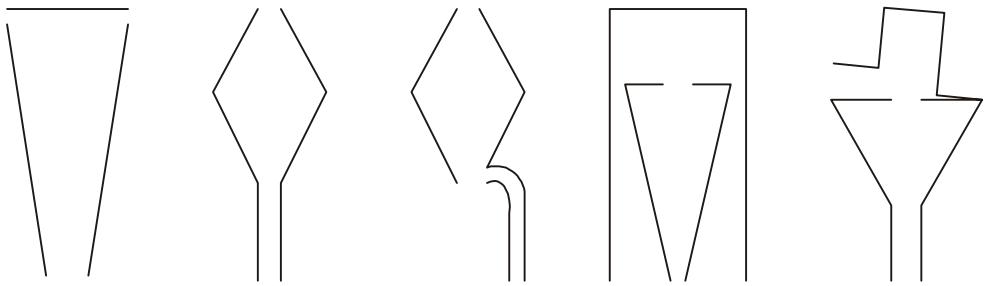

Fig. 1.3: Ressoadores de Kratzenstein [16].

[2] Gravações de áudio de alguns dos sistemas descritos nesta seção podem ser ouvidas no CD anexo ou obtidas no endereço eletrônico do LINSE (www.linse.ufsc.br).

[3] Uma versão moderna dos ressoadores de Kratzenstein pode obtida no endereço eletrônico *Vocal Vowels* [19].

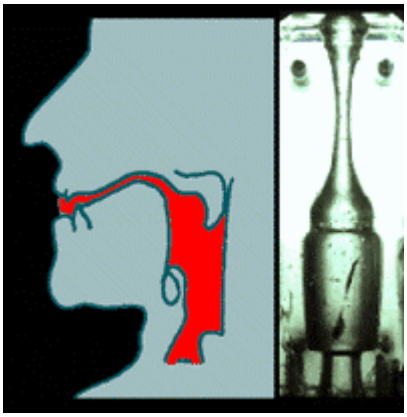

Fig. 1.4: Configuração do trato vocal e respectivo ressoador para a vogal [e] [19].

Em Viena, em 1791, Wolfgang von Kempelen apresentou o resultado de mais de 20 anos de pesquisa ao publicar o livro “O Mecanismo da Fala Humana e a Construção de uma Máquina Falante”. A máquina, um equivalente mecânico do sistema articulatório, era capaz de produzir não só vogais, como palavras e até frases completas. As partes essenciais do dispositivo eram um fole, equivalente aos pulmões, uma palheta vibratória, atuando como as cordas vocais, e um tubo de couro, simulando o trato vocal. Alterando o formato do tubo, era possível produzir diferentes vogais. Obstruções feitas com os dedos em quatro pequenas passagens de ar permitiam a geração de sons consonantais. Na Fig. 1.5, é apresentado um esboço da máquina de von Kempelen.

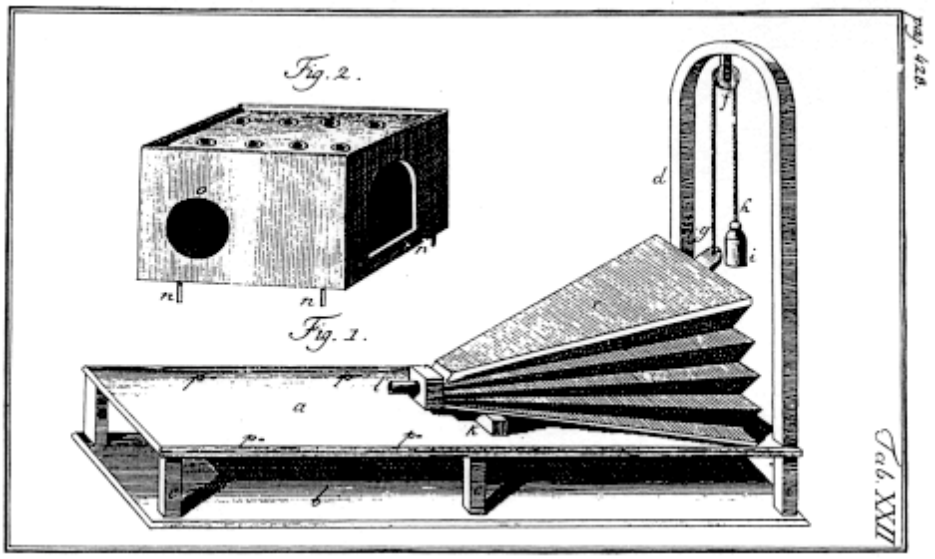

Fig. 1.5: Dispositivo mecânico de síntese de fala de von Kempelen [16].

A máquina falante não foi levada tão a sério na época devido a um acontecimento que marcou negativamente seu criador. Enquanto trabalhava na construção da máquina falante, von Kempelen prometeu para a imperatriz Maria Theresa a criação de uma máquina automática para jogar xadrez. Em seis meses, ela estava pronta e operando [20]. Infelizmente, o mecanismo principal da máquina era um hábil jogador de xadrez colocado no interior.

Na metade do século XIX, Charles Wheatstone construiu uma versão da máquina falante de von Kempelen. Esta, um pouco mais complexa, era capaz de produzir mais sons e combinações de sons. A conexão entre os sons vocálicos e a geometria do trato vocal foi estudada por Willis em 1838. Com ressoadores semelhantes aos de instrumentos musicais chamados órgãos de tubos, ele sintetizou diferentes vogais. Joseph Faber, em 1846, desenvolveu um sintetizador que, com maior controle de *pitch*[4], permitiu cantar *God Save the Queen*, em uma apresentação em Londres. No final do século XIX, Alexander Graham Bell e seu pai construíram também uma máquina de fala. Controversos foram os

[4] *Pitch* é um aspecto subjetivo de um som, relacionado à percepção da freqüência.

experimentos que Bell realizou com seu cão quando fazia estudos para a construção da máquina. Colocava-o entre as pernas, fazia-o rosnar e alterava a conformação do trato vocal com as mãos.

Outros experimentos baseados em sistemas mecânicos e semi-elétricos foram realizados até os anos 60, mas sem muito sucesso.

1.6.2 Sistemas Eletro-Eletrônicos

O primeiro sintetizador eletro-eletrônico de fala foi desenvolvido por Stewart, em 1922. Dois circuitos ressonantes, excitados por uma cigarra elétrica, modelavam as duas freqüências de ressonância mais baixas do trato vocal, gerando sons vocálicos. Não eram, no entanto, sintetizadas consoantes nem transições entre as vogais, impossibilitando a geração de palavras ou sentenças.

O primeiro dispositivo eletro-eletrônico de síntese de fala capaz de gerar sons conectados foi desenvolvido nos Laboratórios Bell e apresentado por Homer Dudley e Richard Riesz na Feira Mundial de 1939, em Nova York. Chamado de VODER (*Voice Operating Demonstrator*), era também conhecido pelos cientistas como Pedro, em alusão ao imperador Dom Pedro II, que em 1876, ao usar um telefone em uma demonstração, exclamou: “Meu Deus! Ele fala!” [21]. O VODER consistia de chaves para seleção de uma fonte sonora ou de ruído, com controle da freqüência fundamental através de um pedal. O sinal da fonte era transmitido através de dez filtros passa-banda, com amplitudes controladas manualmente. Três chaves adicionais introduziam transientes, reproduzindo as consoantes plosivas. Um operador experiente e bem treinado era capaz de produzir frases. A inteligibilidade estava longe de ser considerada boa, mas o potencial de geração de fala sintética estava demonstrado. Um esquema do VODER é ilustrado na Fig. 1.6.

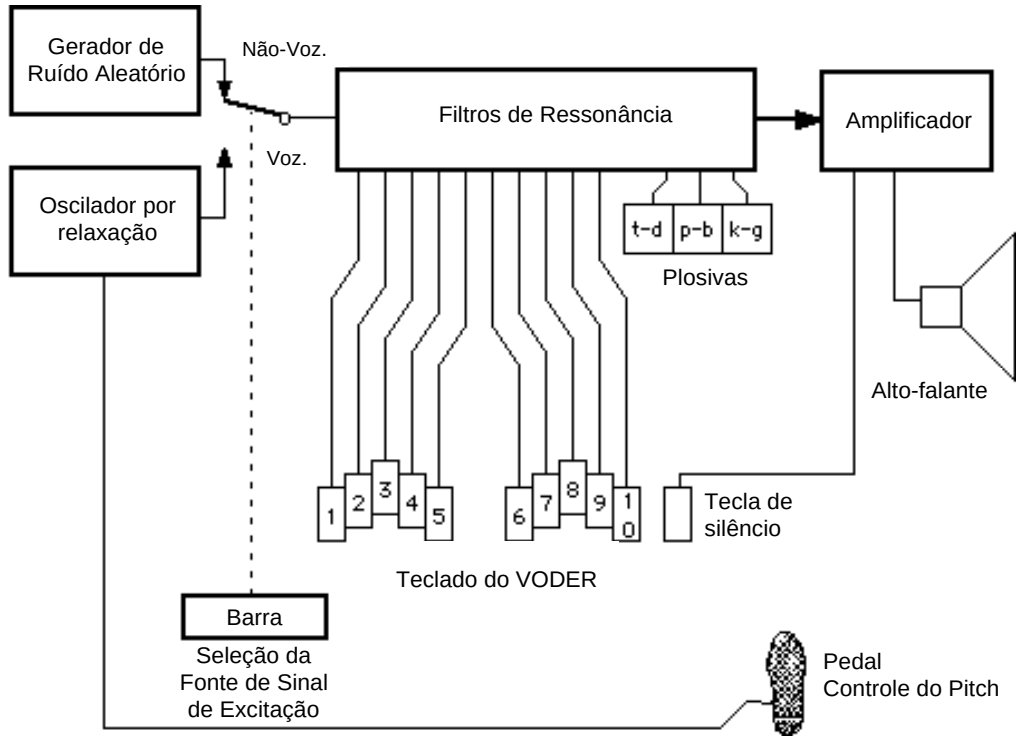

Fig. 1.6: Sintetizador VODER de 1939 [16].

Em 1951, nos Laboratórios Haskins, foi desenvolvido um sintetizador chamado *Pattern Playback*. Essa máquina realizava a função inversa de um espectrógrafo, gerando sons a partir dos padrões de um espectrograma. Na Fig. 1.7, é mostrado um diagrama esquemático do equipamento. Um espectrograma, desenhado com uma tinta especial sobre um filme transparente, era rastreado por um feixe de luz modulado por uma roda tonal. As porções de luz modulada selecionadas pelo espectrograma eram coletadas por um sistema óptico e fornecidas a um elemento fotossensível. A fotocorrente gerada era amplificada e enviada a um alto-falante. Os espectrogramas podiam ser utilizados tanto na forma original como desenhados manualmente, em formato simplificado e estilizado. Assim, era possível realizar experimentos para a determinação de evidências acústicas suficientes para a percepção de diferenças fonéticas. Uma das principais constatações foi a importância das transições entre fonemas. Apesar da naturalidade ser prejudicada pelo *pitch* constante (gerado pela roda), a inteligibilidade era bastante razoável. Palavras de um conjunto de

frases de teste alcançavam 95% de inteligibilidade se copiadas diretamente para o filme transparente e 85%, se simplificadas e estilizadas.

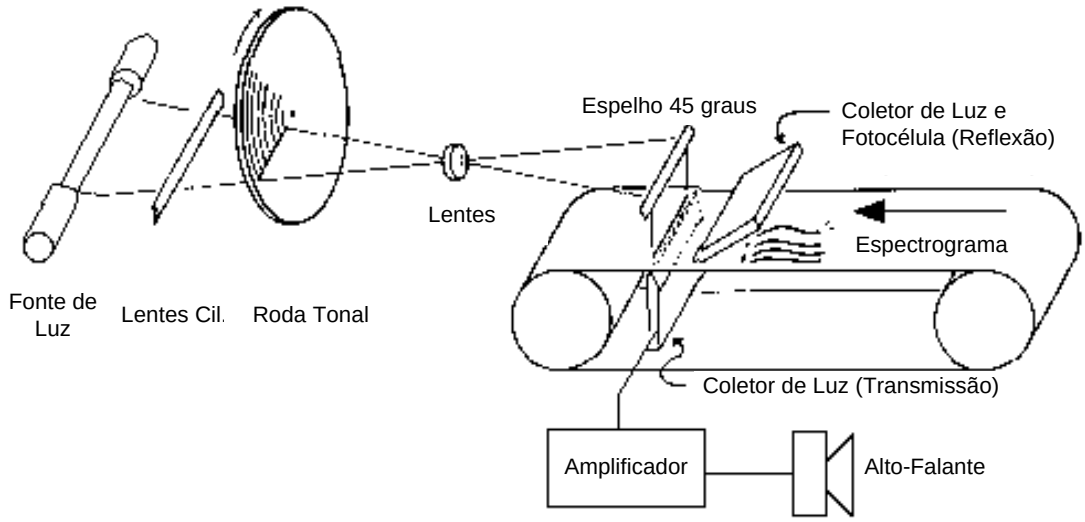

Fig. 1.7: Sistema *Pattern Playback* de 1951 [16].

O VODER e o *Pattern Playback* funcionavam através da cópia dos padrões espectrais da fala variantes no tempo. Uma melhor compreensão do processo de geração de fala, obtida com o desenvolvimento da teoria acústica da produção da fala realizado por Gunnar Fant[5], em 1960, e o conseqüente surgimento de sintetizadores articulatórios e por formantes marcaram um novo passo na história da síntese de fala.

Os primeiros sintetizadores por formantes controlados dinamicamente surgiram em 1953: o PAT (*Parametric Artificial Talker*), de Walter Lawrence e o OVE I (*Orator Verbis Electris*), de Gunnar Fant. Enquanto, no PAT, os ressonadores eram conectados em paralelo; no OVE, a operação era em série.

O primeiro sintetizador articulatório foi desenvolvido por George Rosen, em 1958, no M.I.T. O DAVO (*Dynamic Analog of the Vocal Tract*) era controlado por

[5] Considerações sobre a teoria de produção da fala serão abordadas no Capítulo 2.

gravações em fita de sinais de controle criados manualmente. Em 1968, Cecil Cooker desenvolveu regras para controle de um modelo articulatório. Paul Mermelstein e James Flanagan também trabalharam com síntese articulatória, em 1976.

Em 1968, Noriko Umeda, do Laboratório Eletrotécnico do Japão, desenvolveu o primeiro sistema completo de conversão texto-fala para a língua inglesa. Era baseado em um modelo articulatório e incluía um módulo de análise sintática. A fala era bastante inteligível, mas monótona.

Raymond Kurzweil, em 1976, criou uma máquina de leitura para cegos capaz de ler páginas de texto. Pesando 36 kg, o sistema não foi muito difundido devido ao alto custo. Em 1979, Dennis Klatt, Jonathan Allen e Sheri Hunnicut, todos do M.I.T., apresentaram o sistema *MITalk*. Dois anos depois, com uma nova e sofisticada fonte de sinal, foi lançado o *Klattalk*. Ainda em 1979, foi lançado o primeiro circuito integrado para síntese de fala: o *chip* Votrax. O circuito implementava um sintetizador de formantes em cascata.

No início dos anos 80, começaram a surgir os sistemas TTS comerciais. Baseado no *Klattalk*, foi lançado em 1982 o sistema *Prose-2000* da *Telesensory Systems*. No ano seguinte, a *Digital Equipment Corporation* lançava o *DECTalk*.

O primeiro trabalho em síntese concatenativa foi realizado em 1968 por Red Dixon e David Maxey. Difones[6] eram parametrizados por freqüências de formantes e concatenados. Em 1977, Joe Olive, nos Laboratórios Bell, concatenou difones usando predição linear. A *Texas Instruments* lançou em 1980 um sintetizador, o *Speak-n-Spell*, usando um circuito integrado que realizava síntese baseada em LPC (*Linear Predictive Coding*). Esse *chip* foi usado em um brinquedo eletrônico e recebeu bastante atenção na época.

[6] Difones são segmentos do sinal de fala obtidos da metade de um dado fonema até a metade do fonema seguinte.

Sistemas concatenativos começaram a ganhar espaço em 1985, com o desenvolvimento da técnica de modificação prosódica PSOLA (*Pitch-Synchronous Overlap-and-Add*), proposta por Moulines e Charpentier, da *France Telecom*. Nos anos 90, pesquisadores nos laboratórios do ATR (*Advanced Telecommunications Research International Institute*), no Japão, lançaram os princípios para os sistemas baseados em grandes *corpora*, abordagem utilizada nos sistemas *RealSpeak,* da *Lernout&Hauspie* [22] e *NextGen,* da *AT&T* [23].

Um resumo das etapas do desenvolvimento histórico de síntese da fala é apresentado na Fig. 1.8.

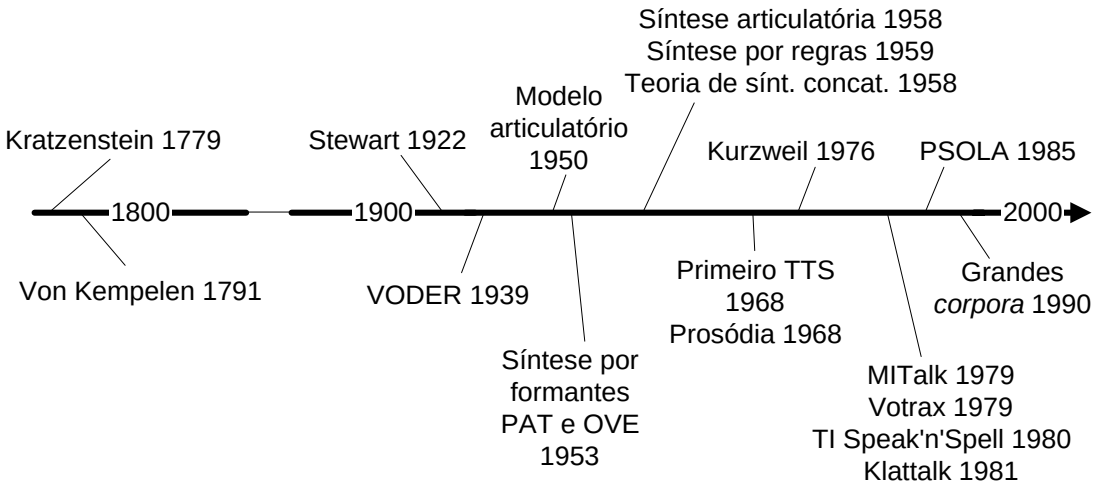

Fig. 1.8: Etapas do desenvolvimento histórico dos sistemas de síntese de fala (adaptado de [15]).

Capítulo 2

Produção da Fala

O sinal de fala produzido pelos seres humanos é formado por ondas acústicas emitidas pelo aparelho fonador. A alteração das características desse aparelho ao longo do processo de fala leva à conseqüente produção de diferentes sons. Aspectos relacionados à produção, percepção e função lingüística desses sons são discutidos neste capítulo.

2.1 Geração de Fala pelo Aparelho Fonador

O aparelho fonador humano é o responsável fisiológico pela produção dos sons da fala. Os órgãos envolvidos nesse processo podem ser divididos em três grupos: o sistema respiratório, o sistema fonatório e o sistema articulatório [24]. A Fig. 2.1 mostra os órgãos que compõem o aparelho fonador.

O sistema respiratório consiste dos pulmões, dos músculos pulmonares, dos brônquios e da traquéia. Os pulmões são a fonte do fluxo de ar que gera o sinal de fala. O fluxo de ar é levado até a laringe através da traquéia.

O sistema fonatório é constituído pela laringe. Na laringe, encontram-se músculos estriados chamados cordas vocais que podem ser posicionados em diversas configurações. A abertura entre as cordas vocais é chamada de glote. Durante a respiração, a glote está normalmente aberta. O fechamento da glote pode obstruir de forma parcial ou total o fluxo de ar vindo dos pulmões. Quando a glote está fechada, o ar que força passagem produz vibração das cordas vocais, que são ligadas e não se abrem a não ser sob

pressão periódica da massa de ar subglótico acumulado [25]. Com a glote aberta, não há vibração das cordas, pois o ar passa sem dificuldades.

O sistema articulatório é composto pela trato vocal e nasal. O trato vocal, formado pela faringe e pela cavidade oral, é o mais importante componente no processo de produção da fala [26]. O trato vocal inicia-se na glote e termina nos lábios. O trato nasal é acoplado ao trato vocal para a produção dos sons nasais. O acoplamento é realizado pelo abaixamento do véu palatino. Durante a produção de sons não-nasais, o véu palatino fecha a cavidade nasal e nenhum som é emitido pelas narinas.

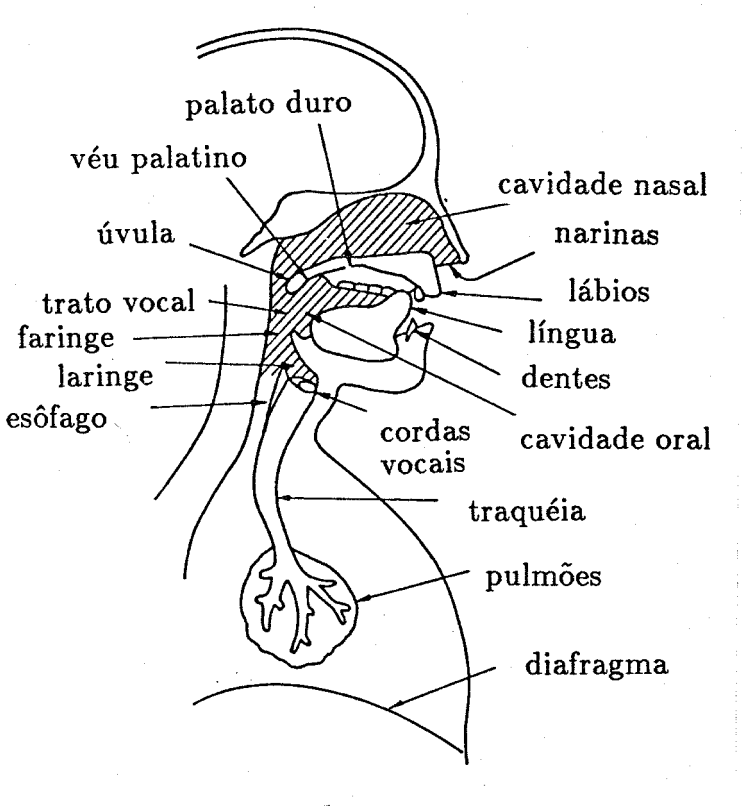

Fig. 2.1: Os órgãos do aparelho fonador humano [26].

O trato vocal é um tubo acústico com área de seção transversal não uniforme e variável com o tempo [26]. A seção transversal do trato é alterada pelo movimento dos órgãos articuladores (mandíbula, lábios, dentes, língua, cordas vocais e véu palatino). A

mudança de forma do trato modifica a resposta em freqüência. As freqüências de ressonância desse tubo acústico são chamadas formantes. Cada configuração do trato vocal é caracterizada por um conjunto de formantes. Na produção de sons nasais, além das freqüências de ressonância surgem freqüências de anti-ressonância[7].

Os sons produzidos pelo aparelho fonador podem ser classificados de forma geral em vozeados (sonoros) e não-vozeados (não-sonoros). Os sons vozeados, que nos interessam mais de perto, são produzidos com as cordas vocais em tal configuração que seja possível uma dada sonoridade. A elevação da pressão do ar nos pulmões força o fluxo de ar através da glote, fazendo com que as cordas vibrem. As cordas vocais funcionam como um oscilador aerodinâmico, controlando o fluxo de ar dos pulmões para a faringe, gerando uma seqüência quase periódica de pulsos de ar. O ciclo de abertura e fechamento da glote define a freqüência fundamental f_0. Esse termo se refere ao número de repetições dos pulsos glotais em um segundo, e é medido em Hz. A freqüência fundamental de adultos do sexo masculino tipicamente se encontra na faixa de 50-250 Hz, enquanto a de mulheres, entre 120 e 480 Hz [27]. O valor médio durante a produção de fala contínua nas línguas européias é de 120 Hz para homens, 220 Hz para mulheres e de 330 Hz para crianças [27]. Nos sons não-vozeados, em contraste aos vozeados, não há vibração das cordas vocais.

Outra forma de caracterizar os sons produzidos pelo aparelho fonador é através da presença ou não de constrições no trato vocal, permitindo a classificação em consoantes e vogais. As consoantes são produzidas com algum tipo de obstrução total ou parcial da passagem da corrente de ar em um ou vários pontos do trato vocal. As obstruções podem ser acompanhadas ou não da vibração das cordas vocais. A título de ilustração, existem aproximadamente 600 consoantes nas diferentes línguas do mundo e 98% das línguas têm as consoantes (surdas) [p], [t] e [k] [28]. As vogais, em oposição às consoantes, são

[7] A anti-ressonância é equivalente à ressonância devido aos pólos, só que agora relativa aos zeros do sistema.

produzidas sem interrupção da corrente de ar no trato vocal. As vogais são caracterizadas pelas ressonâncias do trato, os formantes.

2.2 Prosódia

"Não importa o que você fala, mas sim como você fala". Esse comentário popular mostra a importância do estudo de um conjunto de parâmetros utilizados no processo de comunicação localizados em um nível diferente do segmental. As vogais e consoantes são os segmentos distintivos da fala que combinados geram as sílabas, palavras e frases. Mas, ao mesmo tempo que são articulados esses segmentos, existem outros traços dos quais se derivam tipos de acentos, padrões de entonação, ritmos e velocidades de fala, agrupados em um nível chamado de suprassegmental. Este domínio é também conhecido como prosódia.

Os fenômenos prosódicos podem ser estudados em três níveis: o acústico, o perceptual e o lingüístico [11].

2.2.1 Nível Acústico e Perceptual

O primeiro nível de análise dos fenômenos suprassegmentais está relacionado à realização acústica dos fenômenos, observada e quantificada usando parâmetros de acústica física: freqüência fundamental, duração e intensidade. A conseqüente avaliação subjetiva desses fenômenos pode ser analisada e descrita no nível perceptual pela variação de três traços ou parâmetros prosódicos: *pitch*, *length* e *loudness*[8].

[8] Em português esses parâmetros são traduzidos como altura, quantidade e volume [29], respectivamente. No entanto, consagrou-se a manutenção do termo em inglês *pitch*, que será também utilizado aqui. Neste trabalho, os termos acústicos e perceptuais serão usados como equivalentes.

- Freqüência Fundamental e *Pitch*

Em relação à manifestação acústica, o principal componente da entonação é o *pitch*. O *pitch* é um termo perceptual, que representa a altura ou tom de um sinal sonoro, permitindo distinguir e classificar as sensações auditivas de acordo com a freqüência do sinal.

No processo de fala, a taxa de vibração das cordas vocais define acusticamente a freqüência fundamental f_0. Enquanto a freqüência fundamental é uma grandeza objetiva do sinal sonoro, medida em Hertz, o *pitch* está relacionado ao julgamento humano. Esse julgamento, entretanto, não está linearmente vinculado à freqüência fundamental. Por exemplo, uma freqüência de 1000 Hz é percebida como o dobro de 400 Hz, enquanto que 4000 Hz é percebido como o dobro de 1000 Hz. Para a fala, contudo, os valores de freqüência fundamental são relativamente baixos e para propósitos práticos o termo *pitch* pode ser utilizado como equivalente à freqüência fundamental [1].

- Duração e Quantidade

A quantidade é o parâmetro suprassegmental relacionado à duração. A produção da fala envolve uma sucessão de movimentos articulatórios que têm um intervalo de tempo associado. A percepção de um segmento como mais longo ou mais curto é chamada de quantidade. No entanto, utiliza-se duração (nível acústico) e quantidade (nível perceptual) como termos equivalentes. A medida acústica da duração, devido à natureza contínua da fala, é uma tarefa complexa. Só é possível realizá-la com a clara definição dos pontos iniciais e finais dos segmentos.

- Intensidade e Volume

Entende-se por volume a avaliação perceptual que um ouvinte faz de um som, de acordo com a intensidade do mesmo [30]. A intensidade, por sua vez, está relacionada à

energia acústica, sendo definida como o fluxo médio de energia acústica, por unidade de superfície, em direção normal à propagação. A intensidade é o parâmetro físico, medida em W/m^2, enquanto o volume representa o aspecto perceptivo. Apesar de não serem sinônimos, como não são freqüência e *pitch*, é comum empregar-se o termo intensidade com o sentido de volume.

A intensidade do sinal de fala gerado pelo aparelho fonador está relacionada à maior ou menor pressão de ar proveniente dos pulmões. Entretanto, a relevância do volume como traço prosódico é difícil de ser avaliada de forma isolada, pois o aumento da pressão de ar dos pulmões não altera somente a percepção de volume. Existe uma interrelação entre *pitch* e volume, que pode ser demonstrada com a seguinte experiência [28]: um indivíduo coloca-se diante de uma parede. Tomando ar, começa a emitir uma vogal, por exemplo [a], tentando mantê-la estável. Durante a fonação, é efetuada uma pressão logo abaixo do tórax, aumentando a quantidade de ar que é expelida pelos pulmões. O resultado é um aumento de volume, como esperado, mas com um aumento de *pitch* associado.

2.2.2 Nível Lingüístico

No nível lingüístico, os fenômenos prosódicos são classificados como pertencentes às categorias de acento, entonação, ritmo e velocidade de fala.

- Acento

Em lingüística, o acento é um processo que permite valorizar uma unidade lingüística superior ao fonema (sílaba, palavra, frase), para distingui-la das outras unidades de mesmo nível [25]. Os acentos podem ser divididos em duas categorias: acento de energia e acento de tom. O acento de energia é realizado fisicamente por meio de uma força expiratória maior e tem três funções [25]:

a) distintiva nas línguas em que ele pode ser movimentado. Em português, por exemplo, diferencia palavras como “sabiá”, “sábia” e “sabia”;

b) demarcativa nas línguas em que seu lugar não é livre. Pode indicar o início da palavra, afetando a primeira sílaba, como em tcheco, ou o final da palavra, como em francês, em que afeta a última sílaba;

c) culminativa, marcando o pico de uma unidade fonética que pode ser a palavra ou um grupo de palavras.

O acento de tom é obtido pela variação da altura melódica com um aumento ou redução de *pitch*. Tem as mesmas três funções do acento de energia. A função distintiva, por exemplo, é utilizada na língua cantonesa. Nessa língua, a sílaba [si] tem seis diferentes significados, dependendo do *pitch* empregado, como ilustrado na tabela Tabela 2.1 [28]. Nessa tabela, a primeira coluna mostra o caracter correspondente a cada palavra; a segunda, um símbolo tonal, indicando o *pitch*; a terceira, uma descrição do tom, e a quarta, o significado em português.

Tabela 2.1: Exemplo da pronúncia da sílaba [si] com diferentes tons em cantonês [28]

Sílaba [si] em Cantonês			
Caracter	Símbolo Tonal	Descrição	Significado
詩	˥˩	alta-média	poema
試	˧	nivelada média	tentar
事	˨	nivelada meio baixa	matéria
時	˩	nivelada baixa	tempo
使	˧˥	média-alta	causar
市	˨˧	baixa-média	cidade

- Entonação

A entonação é percebida como o aspecto melódico da fala, sendo relacionada basicamente às variações de *pitch*. É utilizada para veicular informações complementares à simples enunciação, desempenhando as seguintes funções [31]:

a) emocional. A função mais comum da entonação é manifestar a grande gama de sentimentos humanos: surpresa, alegria, excitação, simpatia, aborrecimento, raiva, dentre outros, através da linguagem falada. Pode-se dizer que a base da expressão emocional falada é formada pela entonação associada a outros parâmetros prosódicos;

b) gramatical. A entonação tem um importante papel na marcação dos contrastes gramaticais. A separação, por exemplo, de orações em frases geralmente é função do contorno de *pitch*. Contrastes específicos entre perguntas e respostas, afirmações e negativas são, em várias línguas, dependentes da entonação. Por exemplo, em português, a distinção entre “ele saiu ontem.” e “ele saiu ontem?” está basicamente associada a uma diferença de *pitch*;

c) semântico-pragmática. Uma das formas de atribuição de foco da mensagem é através da entonação. Foco é a parte do enunciado sobre o qual recai a informação nova [32]. Por exemplo, se alguém diz “ela ganhou um colar ontem”, com ênfase na palavra “colar”, pressupõe-se que a informação nova que se deseja transmitir é sobre o objeto (o colar) que foi ganho. Se a ênfase for dada em “ganhou”, a informação que se deseja transmitir é sobre a maneira como ela obteve o colar.

d) psicológica. A entonação pode auxiliar na organização da linguagem em unidades que são mais facilmente percebidas e memorizadas. Uma longa seqüência de números, por exemplo, é memorizada de forma mais simples

se for dividida em segmentos rítmicos. A habilidade de organizar a fala em unidades entonacionais é também um importante atributo do processo de aquisição da linguagem - ausente em casos de distúrbios da linguagem;

e) qualificatória. As características suprassegmentais têm também uma importante função de identificação pessoal. Em particular, atuam no reconhecimento de pessoas pertencentes a diferentes grupos sociais e ocupações. Pela entonação normalmente é fácil identificar um padre, um vendedor ambulante ou um soldado, por exemplo.

- Ritmo e Velocidade da Fala

O ritmo é caracterizado como a percepção de uma seqüência de eventos durante a evolução temporal dos sons. Nota-se inicialmente que o sistema auditivo humano é extremamente sensível a fenômenos que são genuinamente rítmicos [27]. Ouvintes são particularmente hábeis ao perceber a regularidade de uma batida rítmica em música, por exemplo, assim como ao notar desvios em um ritmo estabelecido. No nível lingüístico, as línguas se dividem em de ritmo silábico e de ritmo acentual [29]. As de ritmo silábico são caracterizadas por uma duração praticamente constante das sílabas. O francês é um exemplo de língua silábica. Nas línguas de ritmo acentual, a duração entre os intervalos das sílabas é isocrônica, isto é, as sílabas inacentuadas diminuem sua duração de acordo com o número delas ocorrente entre duas sílabas acentuadas [29]. A velocidade da fala, não confundida com o ritmo, se refere ao tempo total de duração de todas as elocuções em um dado turno de fala, incluídos todos os tipos de pausas.

2.3 Modelo de Produção da Fala

A relativa independência entre o trato vocal e as fontes de excitação permite a representação isolada desses dois elementos em um modelo para a produção da fala. Assim, o sistema de produção da fala pode ser representado por um sistema linear variante

no tempo e um gerador de excitação [26]. A Fig. 2.2 mostra o diagrama de blocos desse modelo.

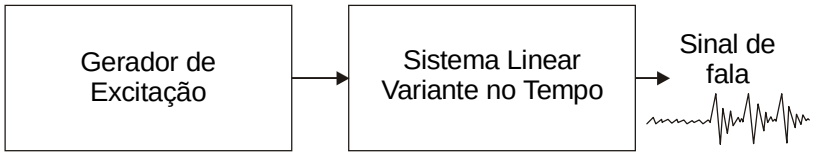

Fig. 2.2: Diagrama de blocos do modelo de produção da fala.

O gerador de excitação gera um sinal periódico para representar os sons vozeados e um sinal aleatório para modelar os sons não-vozeados. O sistema linear variante no tempo representa as ressonâncias do trato (formantes) e os efeitos de radiação nos lábios.

2.3.1 Gerador de Excitação

Os sons vozeados são produzidos pela excitação do trato por um sinal quase periódico. Esse sinal pode ser modelado por um trem de impulsos unitários passando por um filtro digital $G(z)$ que simule o efeito dos pulsos glotais. A periodicidade dos impulsos é determinada pela freqüência fundamental do sinal de fala.

A modelagem dos sons não-vozeados pode ser obtida utilizando-se uma fonte de ruído com espectro plano e um controle de ganho. Essa fonte pode ser representada por um gerador de números aleatórios com variância unitária e média zero [26].

2.3.2 Trato Vocal

O trato vocal pode ser modelado como uma associação em cascata de tubos com área de seção transversal variável. A freqüência de ressonância de cada tubo corresponderia a um formante. Desprezando-se os efeitos de radiação nos lábios, a função de transferência do trato vocal pode ser representada por [26]:

$$V(z) = \frac{G}{\prod_{i=1}^{N}(1 - p_i z^{-1})}, \tag{2.1}$$

onde o ganho G está associado à amplitude do sinal de fala e p_i são os pólos de $V(z)$, representando os formantes.

O modelo composto somente por pólos representa adequadamente os sons formados pelas ressonâncias do trato. Entretanto, na produção de sons nasais e fricativos, o trato apresenta também anti-ressonâncias. Uma representação precisa desses sons exigiria que $V(z)$ possuísse pólos e zeros. Uma aproximação pode ser realizada aumentando o número de pólos de $V(z)$, simulando a presença de zeros.

Assim, o trato vocal pode ser representado por um sistema linear, cuja função de transferência apresenta somente pólos [26].

2.3.3 Radiação

O efeito de radiação nos lábios pode ser modelado por um filtro digital passa-altas, com função de transferência:

$$R(z) = R_0(1 - z^{-1}) \tag{2.2}$$

2.3.4 Modelo Completo

A modelagem apresentada, que separa o efeito das fontes e do trato, resulta no chamado modelo fonte-filtro. O modelo completo que representa o processo de produção da fala é mostrado na Fig. 2.1.

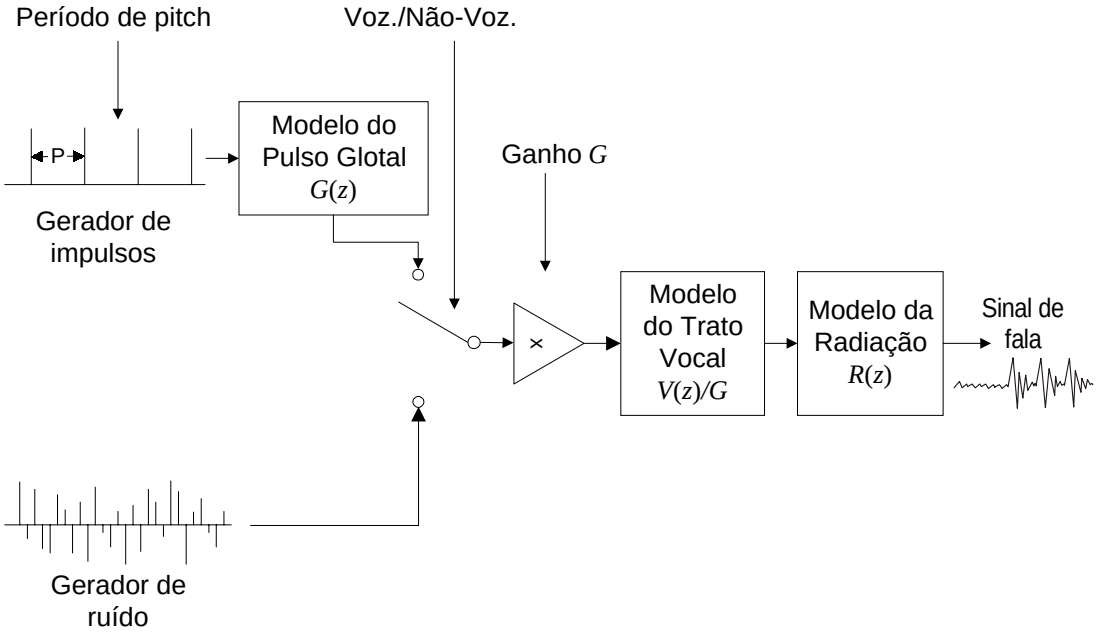

Fig. 2.3: Modelo completo para a produção da fala.

2.3.5 Modelo Simplificado

As funções de transferência do pulso glotal, do trato vocal e da radiação nos lábios podem ser combinadas em uma única função, de modo a simplificar o modelo. A função de transferência resultante é:

$$H(z) = G(z) \cdot V(z) \cdot R(z) \tag{2.3}$$

O modelo simplificado é constituído de duas fontes de excitação (um gerador de impulsos unitários periódicos e um gerador de ruído) e um sistema linear variante no tempo, um filtro digital $H(z)$. O sistema linear pode ser simplificado em uma representação apenas por pólos, sendo $H(z)$ dada por:

$$H(z) = \frac{G}{1 - \sum_{k=1}^{P} a_k z^{-k}}, \tag{2.4}$$

onde P é o número de pólos de $H(z)$ e a_k, os coeficientes do filtro digital.

Para representar a natureza de variância no tempo do sinal de fala, os coeficientes a_k e o sinal de excitação são atualizados em intervalos de tempo regulares (da ordem de 10 ms). Na Fig. 2.4, é mostrado o modelo simplificado de produção da fala.

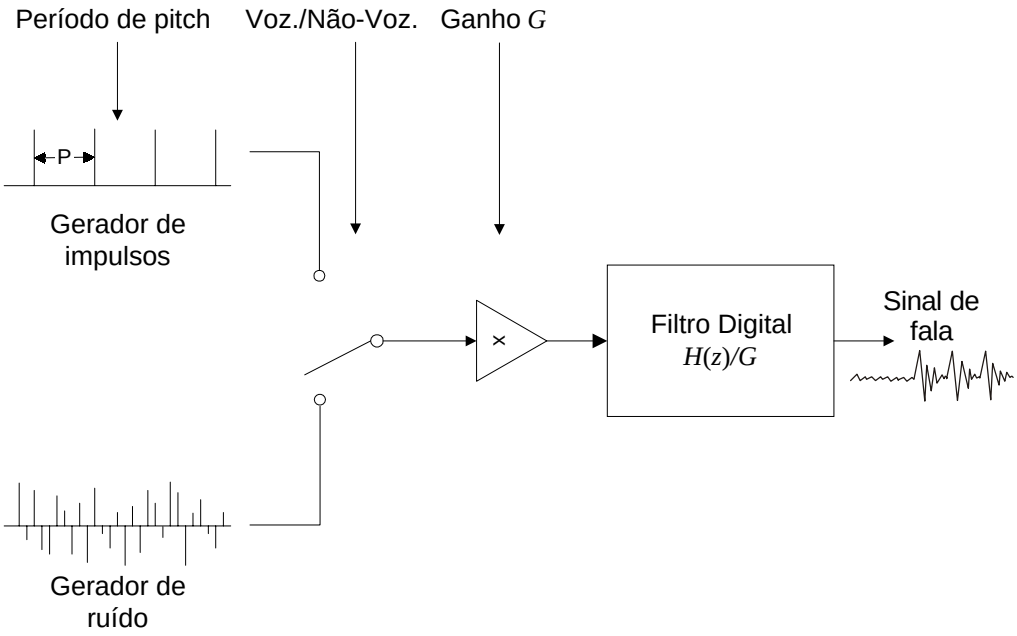

Fig. 2.4: Modelo simplificado para a produção de fala.

Capítulo 3

Processamento Lingüístico

A tarefa de transformação de texto em fala é um problema complexo e exige a divisão em etapas para sua resolução. Na primeira etapa, que envolve aspectos língüísticos, é realizada a análise do texto de entrada. O objetivo é transformar o texto em uma representação simbólica e estruturada que indique os sons que devem ser sintetizados com seus parâmetros prosódicos associados. A análise de texto é fortemente dependente do idioma a que se propõe o sistema de conversão e é subdividida em módulos. Usualmente são incluídos os seguintes estágios de processamento:

a) pré-processamento do texto de entrada;

b) transcrição ortográfico-fonética;

c) separação silábica e determinação da tonicidade;

d) análise sintática;

e) modelagem prosódica.

Na Fig. 3.1, é mostrado um diagrama de blocos das etapas envolvidas na análise de texto para conversão texto-fala.

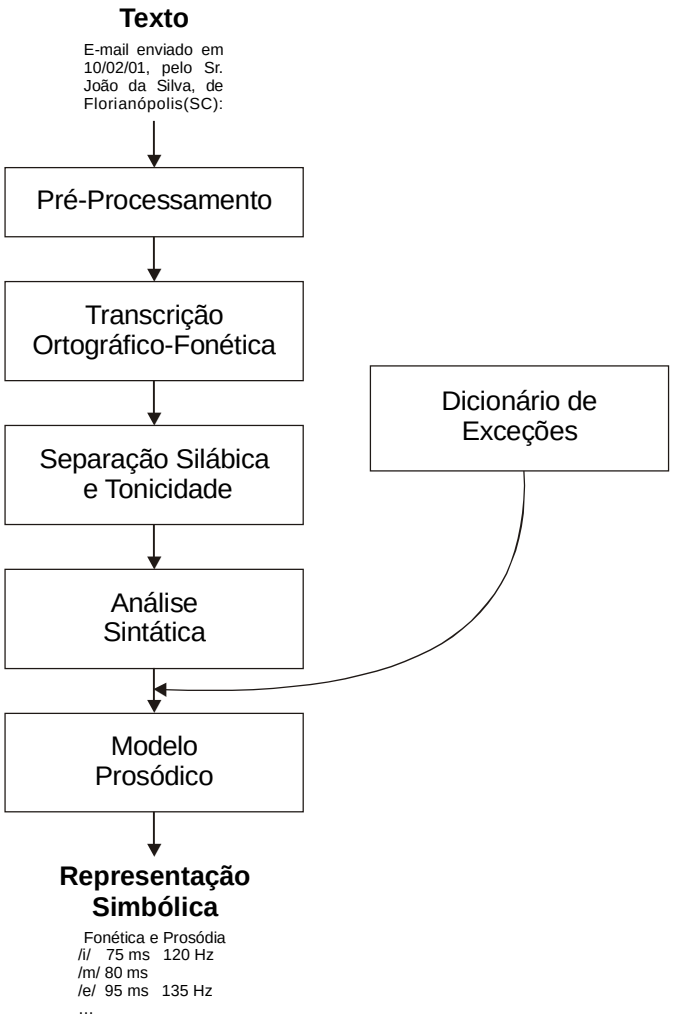

Fig. 3.1: Diagrama de blocos da etapa de análise do texto.

3.1 Pré-Processamento

A primeira função da etapa de pré-processamento é a separação do texto de entrada em grupos de palavras que facilitem o processo de análise. O grupo que parece mais evidente é a frase, e a maioria dos sistemas separa o texto em frases. Em alguns sistemas escritos, como o chinês, que possuem um símbolo exclusivo para assinalar o final de frases declarativas, não há dificuldades (em chinês, é usado um pequeno círculo) [33]. Já em línguas como o inglês e o português, o processo não é tão direto. Nessas línguas, o

mesmo sinal de ponto empregado para a marcação do final de frases declarativas é utilizado para assinalar, por exemplo, abreviaturas. Em português, o ponto em “Sr.” não marca (normalmente) um final de frase mas sim corresponde à abreviatura de “Senhor”. Assim, antes de definir um ponto como um marca de separação de frases é necessário eliminar outras possibilidades. No caso de abreviaturas, essas devem ser identificadas e expandidas.

O processo de expansão de abreviaturas também não é trivial, pois muitas delas são usadas com diferentes significados. Por exemplo, “v.” pode ser usado como abreviatura de “veja” ou de “verbo”. A letra “s” sem ponto abrevia “segundo” ou “segundos” (outro problema, determinar se será usado o plural ou não), e, seguida de ponto, pode significar “substantivo” ou “Sul”. “Sul” pode ser identificado se o texto foi escrito corretamente com o emprego de letras maiúsculas, pois a abreviatura correta é “S.”. Contudo, atualmente é muito comum, principalmente em mensagens de correio eletrônico, o “esquecimento” dos caracteres maiúsculos. Além das abreviaturas, siglas são empregadas no texto. Algumas são soletradas, como “FGTS” que deve ser expandida para “efe gê tê esse”. Outras podem ser lidas como se fossem palavras, pela identificação de padrões silábicos da língua, como é o caso de “CEF”.

Os algarismos também devem ser expandidos de forma adequada. Algarismos arábicos são empregados em diversas situações e cada uma delas deve ser tratada separadamente. Por exemplo, “331” pode ser lido como “trezentos e trinta e um” quando representa uma quantidade qualquer ou como “três três um” se for a primeira parte de um número telefônico. Os algarismos 1, 2 e as centenas de 200 até 900 apresentam um problema adicional: possuem uma variação feminina. Por exemplo, “542 éguas” deve ser expandido de forma diferente de “542 cavalos”. As formas de valores monetários, números cardinais, datas e horas têm suas peculiaridades próprias e devem ser tratadas de modo apropriado. A expansão dos algarismos romanos também é necessária. Para alguns casos,

não é tão simples. Por exemplo, “VI” pode representar o algarismo romano seis ou a primeira pessoa do singular do pretérito perfeito do indicativo do verbo “ver”.

3.2 Transcrição Ortográfico-Fonética

O objetivo da transcrição é a transformação da representação ortográfica em uma representação fonética. Se, para cada caracter, existisse um mapeamento único no domínio fonético, essa tarefa seria simples. Entretanto, algumas letras representam mais de um fonema, como a letra “x”, que, na língua portuguesa, descreve o fonema [ʃ] em “xale”, [z] em “exame”, [s] em “explicar” e os fonemas [ks] em “táxi”. Além disso, o processo de transcrição fonética deve ser robusto o suficiente para lidar com nomes próprios, derivados de diferentes idiomas.

Diferentes estratégias podem ser utilizadas para a transcrição fonética. A mais simples é a transcrição por regras, baseada no contexto em que está inserida a letra em análise. Em português, em que a correspondência entre letras e fonemas é razoavelmente estável, esta é a técnica mais empregada [1,3,34]. Um dicionário de exceções com um número relativamente pequeno de verbetes - da ordem de 1000 - cobre as eventuais falhas de transcrição. Como exemplo de regras de conversão, pode-se citar a análise realizada para a letra “c”, associada a dois fonemas, [k] e [s] [35]:

a) se a letra seguinte ao “c” for “a”, “o”, “u” ou consoante, o fonema associado será [k], como, por exemplo, nas palavras “caco”, “clube” e “cubo”;
b) se a letra seguinte ao “c” for “e” ou “i”, o fonema associado será [s], como nas palavras “certo” e “ciúme”.

Uma outra abordagem usa um grande dicionário de radicais de palavras, prefixos e sufixos com uma transcrição fonética associada. Os casos não cobertos pelo dicionário são resolvidos com algumas regras de conversão letra-fonema. Para o inglês, normalmente essa é a abordagem utilizada. Dicionários com um número de entradas da ordem de dezenas de milhares de palavras são empregados [13].

Para a língua portuguesa, uma das maiores dificuldades na transcrição ortográfico-fonética é determinar se as letras “e” e “o” sem acento ortográfico correspondem a vogais abertas ou fechadas [1]. Esse problema ocorre porque nesses casos apenas o contexto lexical não é suficiente para a determinação correta da abertura ou fechamento da vogal. Por exemplo, para as palavras “bolo” e “bola” não há como desenvolver uma regra que atue apenas pela avaliação do contexto anterior e posterior em que se insere a vogal. A solução é, para esses casos, a inclusão dessas palavras em um dicionário de exceções.

3.3 Separação Silábica e Determinação da Tonicidade

A sílaba desempenha um papel importante no estudo da prosódia. A implementação de um modelo prosódico pode ser facilitada se for efetuado um procedimento de separação silábica e determinação da tonicidade das sílabas.

Para o português, um algoritmo de separação silábica é apresentado em [1]. É implementado através de um diagrama de estados e realiza a separação no nível dos fones.

A posição da sílaba tônica é uma informação importante a ser considerada na formulação dos modelos prosódicos, pois a variação dos parâmetros suprassegmentais é fortemente dependente da tonicidade.

Em [1] é apresentado um procedimento de determinação da tonicidade das sílabas para o português baseado em regras. Apesar de não resolver todas as situações, as regras tentam abranger o maior número de casos possível. Citam-se, a título de exemplo, duas dessas regras:

a) palavras marcadas com diacríticos (acentos gráficos) já têm a sílaba tônica determinada, prevalecendo essa regra sobre todas as demais;

b) palavras terminadas em “im” ou “um” são oxítonas.

Além da tonicidade silábica, é muito importante considerar a tonicidade em níveis mais altos (entre palavras e dentro de uma frase). Por exemplo, nem todas as

palavras de uma frase têm a mesma proeminência. Numa frase como “a moça gosta de torta de banana e de maçã” é possível perceber que algumas palavras são mais importantes para a comunicação e são ditas com um destaque maior do que as outras. Palavras de conteúdo, isto é, nomes, verbos, adjetivos, tendem a ser mais salientadas do que as palavras funcionais, que incluem verbos auxiliares e preposições. Problemas podem ocorrer com os nomes compostos. Por exemplo, em inglês, “Madison Avenue” é acentuada foneticamente na última palavra, enquanto “Wall Street” na penúltima [33].

3.4 Análise Sintática

A análise sintática determina a estrutura da frase e identifica os elementos que a compõem. A estrutura frasal é uma informação indispensável para uma modelagem prosódica correta das pausas, entonação e ritmo. Alguns pontos da frase correspondem a limites prosódicos, onde ocorrem mudanças abruptas de *pitch*, duração e intensidade. As pausas, por exemplo, não podem ser colocadas em qualquer ponto da frase.

Além disso, a determinação da categoria sintática de cada palavra é usada para eliminar a ambigüidade na pronúncia de alguns vocábulos. Tome-se como exemplo as frases:

a) O almoço será servido logo.

b) Eu almoço sempre ao meio-dia.

Nas duas frases, a palavra “almoço” é escrita da mesma forma, mas pronunciada de maneiras diferentes. Só é possível determinar a pronúncia correta através do conhecimento da categoria gramatical. Se for verbo, a vogal “o” é aberta, se substantivo, a vogal é fechada.

Duas abordagens são comuns para a tarefa de análise sintática. A primeira utiliza um classificador estocástico [36]. Um modelo estatístico da linguagem é derivado de grandes conjuntos de texto classificado. Para cada palavra, é determinada uma categoria gramatical mais provável, dada a probabilidade de ocorrência das palavras dentro de um

certo contexto. A segunda abordagem é baseada em regras gramaticais, que descrevem uma seqüência válida de símbolos. Os símbolos correspondem a classes de palavras, grupos de palavras representando frases, orações ou mesmo frases inteiras.

3.5 Modelagem Prosódica

A incorporação de prosódia a um sistema de síntese de fala é um fator fundamental para que os requisitos de inteligibilidade e naturalidade sejam atendidos. A prosódia é imposta à fala através da variação temporal dos parâmetros prosódicos *pitch*, duração e intensidade. O objetivo de um modelo prosódico é a determinação da evolução temporal dos parâmetros prosódicos, de forma que seja possível identificar na fala sintetizada os atributos lingüísticos de acento, ritmo e entonação, que, em última análise, conferem uma avaliação de boa qualidade.

Uma estrutura comumente adotada é a separação do modelo prosódico em modelo de duração e modelo entonacional ou de *pitch*. Na literatura consultada, não foram encontradas referências ao desenvolvimento de um modelo específico de intensidade. É importante destacar que a modelagem prosódica é fortemente relacionada aos módulos precedentes de análise, principalmente os de determinação de tonicidade e de análise sintática.

Por modelo de duração aplicado à síntese de fala, entende-se qualquer tratamento automático pelo qual as durações dos fones de um enunciado a ser sintetizado possam ser determinadas [37]. Várias abordagens têm sido empregadas e uma revisão das técnicas é encontrada em [38]. Destaca-se, para o caso do português, o modelo desenvolvido em [3], que emprega um dicionário de contornos de duração obtido a partir de dados extraídos da fala de um locutor. O contorno mais adequado é selecionado através do cálculo de um índice que leva em consideração a classificação gramatical do grupo prosódico em análise. Um ajuste do contorno geral é realizado através de regras que modelam os efeitos locais da duração. Como exemplo, cita-se uma regra de efeito local:

a) segmentos da sílaba final de uma palavra têm suas durações aumentadas por um fator de 1,08 para vogais e de 1,05 para consoantes, com exceção de [p].

Em relação aos modelos entonacionais, diferentes abordagens têm sido propostas na literatura, baseadas nos três níveis de análise dos fenômenos prosódicos: nível acústico, perceptual e lingüístico. Cada modelo tem seu grau de complexidade e de relacionamento com os outros módulos de análise e um conseqüente grau de qualidade perceptual. Uma boa revisão das várias abordagens empregadas é encontrada em [11]. Para o português, [37] apresenta um modelo cuja principal característica é basear-se em uma estrutura hierárquica de sentença, composta pelos níveis de frase, constituinte prosódico, palavra, sílaba e fone. Nessa abordagem, cada nível obedece às regras do nível superior e gera outras regras para o nível inferior. No nível de frase, são estabelecidos limites de variação superior e inferior de *pitch* para a elocução. O contorno é aperfeiçoado dentro desses limites até chegar ao último nível, no qual estarão definidos os valores inicial e final de *pitch* para cada fone.

Capítulo 4

Síntese do Sinal de Fala

Após a etapa de processamento lingüístico, já são conhecidos os sons que devem ser sintetizados e os parâmetros prosódicos que devem ser aplicados. É realizada, então, a síntese do sinal acústico de fala. As abordagens mais utilizadas para a síntese propriamente dita são:

a) abordagem de sistema, também conhecida como síntese articulatória, em que o próprio mecanismo de produção da fala é modelado;
b) abordagem de sinal, em que o sinal de fala é o objeto de representação. A síntese por formantes e a síntese por concatenação figuram nessa abordagem.

4.1 Síntese Articulatória

A síntese articulatória tem por objetivo reproduzir o sinal de fala, modelando os mecanismos de sua produção natural [39]. É potencialmente o melhor método para a geração de fala sintética de alta qualidade. Ao mesmo tempo, o de implementação mais complexa, por depender de uma ampla compreensão do processo de produção da fala, e o mais custoso computacionalmente [15].

Nesta abordagem, a produção dos sons da fala, partindo da glote até os lábios, é modelada em diferentes passos. Inicialmente, é necessário criar um modelo para a fonte primária da voz humana, a vibração das cordas vocais. O formato do trato vocal é delineado em seguida, através da determinação da função área. Essa função é definida como a área instantânea da seção reta do trato vocal, da glote aos lábios, determinada pelo

posicionamento dos articuladores [40]. A estimação da função área pode ser realizada de duas formas:

a) diretamente pela observação da fala através de raios X ou ressonância magnética;
b) através de um mapeamento inverso ou acústico/articulatório, utilizando um processo analítico [40].

Na última etapa, é realizada a modelagem do movimento dos lábios. Esse modelo é essencial se a aplicação possibilitar também a síntese visual, que contribui para uma melhor compreensão da mensagem em situações ruidosas [39].

Como exemplo de parâmetros articulatórios de controle, pode-se citar o modelo descrito em [41] que utiliza: a área da abertura dos lábios, a constrição formada pela lâmina da língua, a abertura para as cavidades nasais, a área glotal média e a taxa de expansão ou contração ativa do volume do trato vocal na parte posterior de uma constrição.

Atualmente, a síntese articulatória deve ser considerada mais como uma ferramenta de pesquisa do que uma alternativa viável para aplicações comerciais [39]. Mesmo os sistemas no estado-da-arte não são capazes de gerar fala com a qualidade dos outros métodos baseados na abordagem de sinal.

4.2 Síntese por Formantes

A síntese por formantes, também conhecida por síntese paramétrica, é baseada na modelagem do processo de fala apresentada na Seção 2.3. O processo físico é descrito matematicamente pela combinação linear de três componentes: fontes de sinal, característica de filtragem do trato vocal e característica de radiação para o meio externo, conforme o diagrama de blocos mostrado na Fig. 4.1.

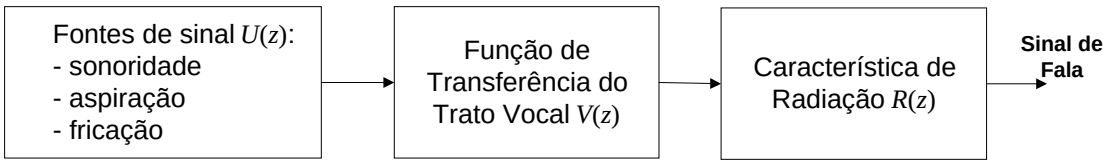

Fig. 4.1: Esquema simplificado do processo de produção da fala empregado na síntese por formantes.

A principal característica da síntese por formantes é a implementação da função de transferência do trato vocal $V(z)$ através da associação de seções de segunda ordem. Essas seções, conhecidas como ressonadores, têm a seguinte função de transferência:

$$R_n(z) = \frac{a_{1n}}{1 - a_{2n} z^{-1} - a_{3n} z^{-2}} \tag{4.1}$$

Os coeficientes a_{1n}, a_{2n} e a_{3n} estão relacionados à freqüência central de ressonância f_{rn} e à largura de banda do formante B_n por:

$$a_{3n} = -e^{-2\pi B_n T} \tag{4.2}$$

$$a_{2n} = 2e^{-2\pi B_n T} \cos(2\pi f_{rn} T) \tag{4.3}$$

$$a_{1n} = 1 - a_{3n} - a_{2n}, \tag{4.4}$$

onde

f_{rn} é a freqüência central de ressonância em Hz;

B_n é a largura de banda do ressonador em Hz;

T é o período de amostragem em segundos.

Na Fig. 4.2, é mostrada a magnitude da resposta em freqüência de um ressonador que implementa um formante com $f_r = 1000$ Hz e largura de banda $B = 150$ Hz. O período de amostragem é de 0,1 ms.

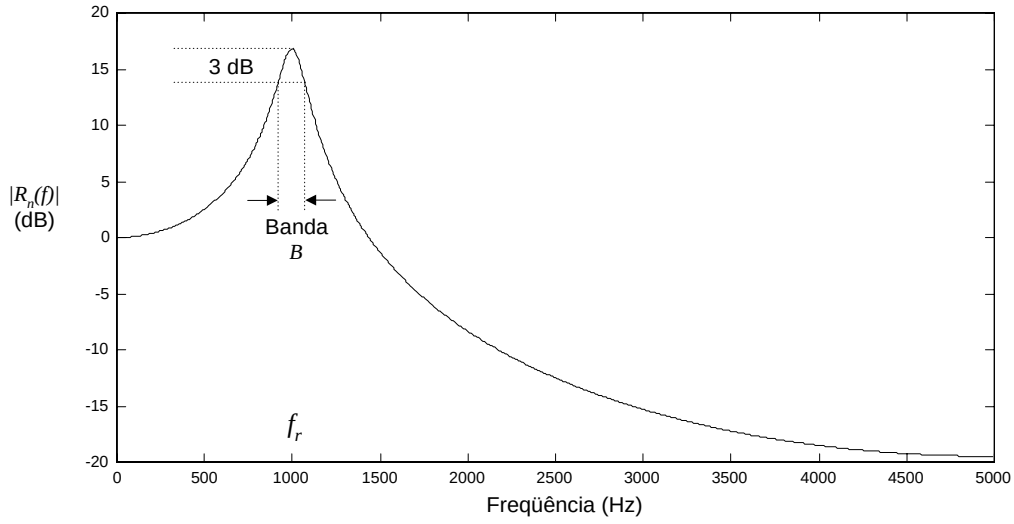

Fig. 4.2: Magnitude da resposta em freqüência de um ressonador digital com f_r =1000 Hz e largura de banda B =150 Hz.

A estrutura do ressonador digital de segunda ordem é ilustrada na Fig. 4.3.

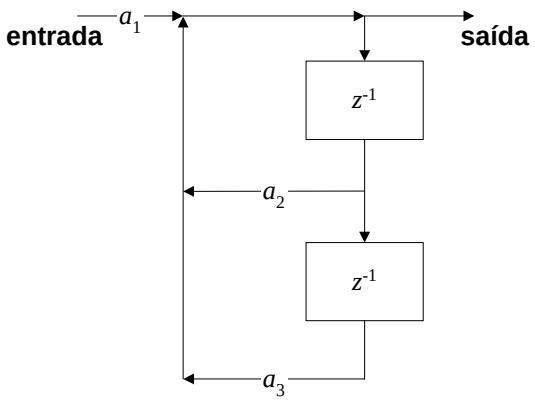

Fig. 4.3: Estrutura do ressonador digital de segunda ordem.

Os ressonadores podem ser associados em cascata ou paralelo. Na associação em cascata ou série, a saída de cada ressonador é conectada à entrada do seguinte, conforme mostrado na Fig. 4.4.

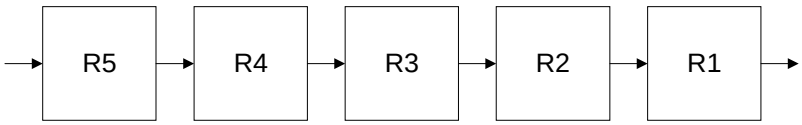

Fig. 4.4: Estrutura em cascata.

A principal vantagem desta estrutura é a manutenção das amplitudes relativas entre os formantes, sem necessidade de controle individual de amplitude para cada um deles, o que corresponde ao modo natural de ressonância do trato vocal [42]. Assim, a conexão em série é adequada para a síntese de vogais não-nasais [43], além de ser simples de implementar [15]. Entretanto, esta associação só consegue modelar adequadamente a existência de zeros no espectro empregando uma ordem bastante elevada [1].

Na associação em paralelo, mostrada na Fig. 4.5, é possível controlar individualmente a amplitude de cada formante. Essa não é uma boa aproximação do comportamento ressonante do trato vocal e, por isso, os sintetizadores em paralelo são mais adequados para a produção de consoantes do que de vogais [42]. São úteis também para a geração de sinais sintéticos utilizados em experimentos perceptuais, nos quais se deseja estudar a relação de amplitude entre os formantes ou a influência isolada de cada um deles [43].

Para utilizar as melhores características de cada uma das associações, algumas implementações híbridas foram propostas na literatura [3,43]. Uma das mais conhecidas é a mostrada na Fig. 4.6, o sintetizador de Klatt [43]. Nesse modelo, a função de transferência do trato vocal em cascata é implementada pelos ressonadores R1 a R5. A síntese de sons nasais é efetuada com um ressonador adicional RNP e por um anti-ressonador RNZ. Na configuração em paralelo, sete ressonadores estão disponíveis (R1,..., R6, RNP), cada um com um controle de ganho associado (A1,..., A6, AN). Uma conexão de *by-pass*, com um controle de ganho *AB*, permite a simulação de sons que não têm características de ressonância bem definidas [42]. A chave *SW* controla a mudança entre a estrutura em série e paralelo.

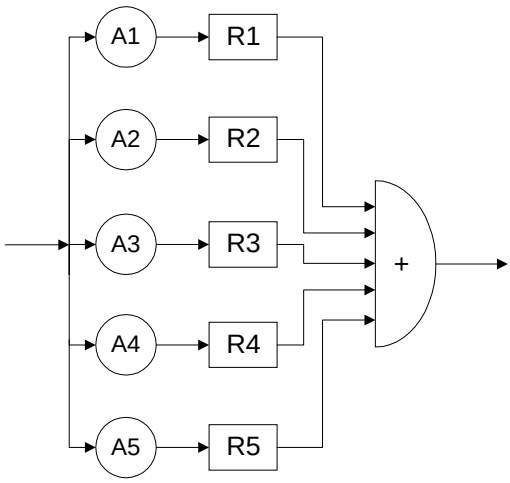

Fig. 4.5: Estrutura em paralelo.

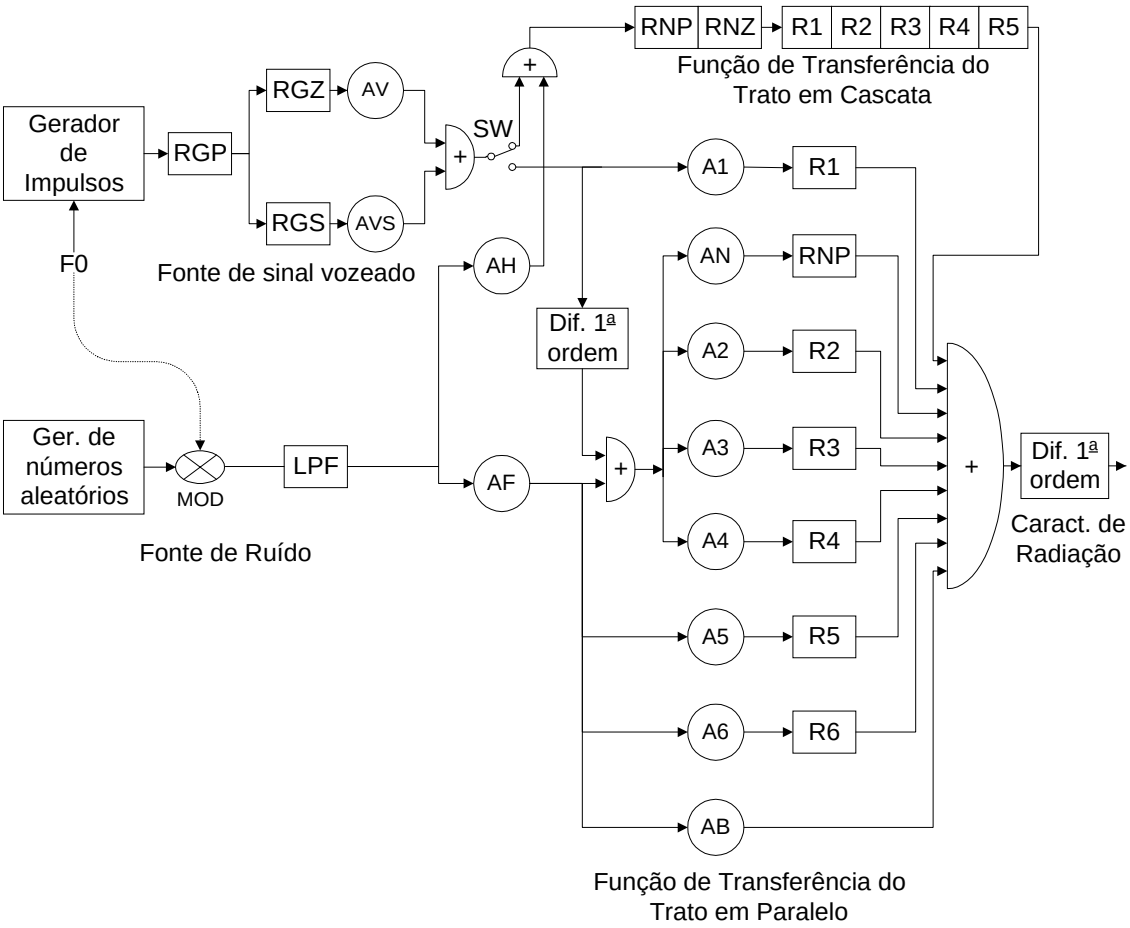

Fig. 4.6: Diagrama do sintetizador de Klatt [43].

A implementação das fontes pode produzir dois tipos de excitação: sonora e ruidosa. Para sons vozeados, ainda é possível gerar duas excitações. Na primeira, o modelo consiste em um trem de pulsos, conformado por um filtro passa-baixas RGP que impõe um decaimento espectral de –12 dB/oitava. O resultado é um sinal que se assemelha aos pulsos glotais naturais. O anti-ressonador opcional RGZ modifica alguns detalhes espectrais do sinal. A segunda alternativa de fonte vozeada gera um sinal quase-senoidal utilizado para a geração das fricativas vozeadas. Um decaimento de –24 dB/oitava é obtido com um segundo filtro RGS.

A fonte de ruído simula o ruído de turbulência produzido pela passagem do ar por uma constrição [1]. Se a constrição está localizada no nível das cordas vocais, o ruído é de aspiração, com ganho controlado por AH. Se a constrição está acima da laringe, o ruído é de fricação, com amplitude controlada por AF. A saída do gerador de números aleatórios, com espectro aproximadamente plano, é passada por um filtro passa-baixas LPF que cancela o efeito da radiação nos lábios. Uma modulação de amplitude do ruído é realizada pelo modulador MOD.

A característica de radiação nos lábios é implementada por um diferenciador de primeira ordem.

O controle do sintetizador é efetuado através de 39 parâmetros, atualizados a cada 5 ms. Na configuração padrão, o sistema opera com uma taxa de amostragem de 10 kHz. Os parâmetros de controle, com a respectiva descrição e valores mínimo e máximo, são mostrados na Tabela 4.1.

Tabela 4.1: Parâmetros de controle do sintetizador de Klatt

Parâmetro	Descrição	Mínimo	Máximo
AV	Amplitude de vozeamento normal (dB)	0	80
AVS	Amplitude de vozeamento quase senoidal (dB)	0	80
AF	Amplitude de ruído de fricação (dB)	0	80
AH	Amplitude de ruído de aspiração (dB)	0	80
F0	Freqüência fundamental de vozeamento (Hz)	0	500
F1	Freqüência do primeiro formante (Hz)	150	900
F2	Freqüência do segundo formante (Hz)	500	2500
F3	Freqüência do terceiro formante (Hz)	130	3500
F4	Freqüência do quarto formante (Hz)	250	4500
F5	Freqüência do quinto formante (Hz)	350	4900
F6	Freqüência do sexto formante (Hz)	400	4999
B1	Largura de banda do primeiro formante (Hz)	40	500
B2	Largura de banda do segundo formante (Hz)	40	500
B3	Largura de banda do terceiro formante (Hz)	40	500
B4	Largura de banda do quarto formante (Hz)	100	500
B5	Largura de banda do quinto formante (Hz)	150	700
B6	Largura de banda do sexto formante (Hz)	200	2000
AN	Amplitude do formante nasal (dB)	0	80
A1	Amplitude do primeiro formante (dB)	0	80
A2	Amplitude do segundo formante (dB)	0	80
A3	Amplitude do terceiro formante (dB)	0	80
A4	Amplitude do quarto formante (dB)	0	80
A5	Amplitude do quinto formante (dB)	0	80
A6	Amplitude do sexto formante (dB)	0	80
AB	Amplitude do trajeto de bypass (dB)	0	80
FGP	Freqüência do ressonador glotal 1 (Hz)	0	600
BGP	Largura de banda do ressonador glotal 1 (Hz)	100	2000
BGS	Largura de banda do ressonador glotal 2 (Hz)	100	1000
FNZ	Freqüência do zero nasal (Hz)	200	700
BNZ	Largura de banda do zero nasal (Hz)	50	500
FNP	Freqüência do pólo glotal (Hz)	200	500
BNP	Largura de banda do pólo nasal (Hz)	50	500
FGZ	Freqüência do zero glotal (Hz)	0	500
BGZ	Largura de banda do zero glotal (Hz)	100	9000
NFC	Número de formantes em cascata	4	6
G0	Controle geral de ganho (dB)	0	80
SR	Taxa de amostragem (Hz)	500	2000
NWS	Número de amostras geradas por conjunto de parâmetros	1	200
SW	Seleção da configuração em cascata ou paralela	0	1

A síntese baseada em regras é uma abordagem poderosa para síntese de fala. É possível gerar fala sintetizada de alta qualidade desde que os parâmetros de controle sejam

ajustados de forma correta. A flexibilidade também é um ponto forte. Novas vozes e diferentes efeitos podem ser criados facilmente. Entretanto, a grande dificuldade se encontra na obtenção dos parâmetros de controle, principalmente para a transição entre sons diferentes. A metodologia mais empregada é a de, tomando como referência frases produzidas naturalmente, obter e ajustar os parâmetros por tentativa e erro. O desenvolvimento torna-se lento, sendo comum o esforço de vários anos para obter uma boa qualidade [11]. Pesquisas recentes [17] buscam a obtenção dos parâmetros de forma automática, empregando técnicas inicialmente utilizadas em reconhecimento de fala, como os modelos ocultos de Markov (HMM).

4.3 Síntese Concatenativa

Em síntese concatenativa, fala sintética é produzida pela concatenação de segmentos. Esses segmentos são previamente gravados e armazenados formando um banco de unidades. A escolha dos segmentos necessários para a geração de uma dada elocução baseia-se nas informações obtidas através da etapa de processamento lingüístico. Com uma etapa de concatenação e alteração de parâmetros prosódicos, a fala sintetizada é gerada. O diagrama de blocos desse processo é mostrado na Fig. 4.7.

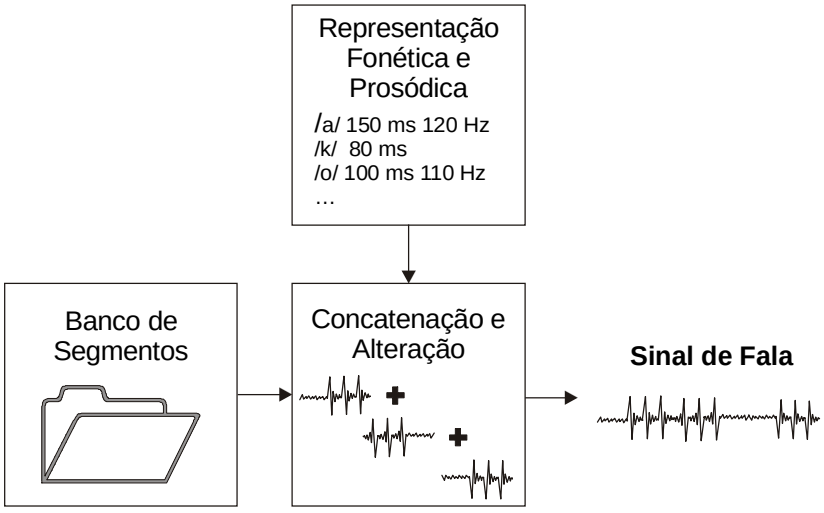

Fig. 4.7: Diagrama de blocos da síntese do sinal de fala pela técnica concatenativa.

Em oposição à síntese por formantes, aqui não há necessidade de definição de regras de transição entre sons, pois as transições podem estar incorporadas aos segmentos armazenados. Cada segmento é obtido de uma gravação de um locutor, e um resultado de alta qualidade poderia ser esperado. Contudo, problemas podem ocorrer, fazendo com que os sistemas concatenativos sofram de uma grande variação de qualidade: em uma sentença, o resultado é excelente, mas na seguinte, pode ser sofrível. Se a combinação das unidades em uma dada frase sintética é adequada, o resultado é tão bom quanto o obtido naturalmente em uma gravação. Mas, se ocorrem muitas descontinuidades espectrais entre os segmentos, a qualidade torna-se baixa.

As descontinuidades espectrais ocorrem quando os formantes de segmentos adjacentes não têm os mesmos valores e estão relacionadas, principalmente, à coarticulação. A coarticulação é a influência de um fonema sobre outro. Por exemplo, na Fig. 4.8, é mostrado um espectrograma da concatenação de dois segmentos: [nẽ] e [ẽb]. A vogal [ẽ] do primeiro segmento é influenciada pelos movimentos articulatórios da consoante [n], tendo assim a posição de seus formantes alterada. Quando é feita a concatenação com o segmento [ẽb], retirado da palavra “câmbio”, um descasamento espectral é gerado, pois a vogal [ẽ] desse segmento é influenciada pelos movimentos articulatórios da consoante [k].

Na etapa de modificação prosódica também podem ocorrer perdas de qualidade dependendo das técnicas que são utilizadas. Com esses problemas, os ouvintes avaliam a fala sintética de forma negativa, mesmo com os segmentos sendo obtidos de forma natural.

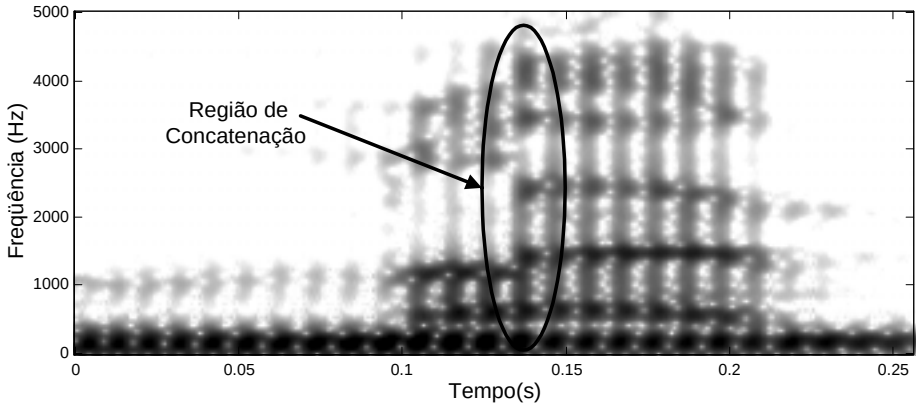

Fig. 4.8: Descasamento espectral em síntese concatenativa.

Assim, pode-se dizer que o resultado final em síntese por concatenação é fortemente dependente dos seguintes fatores:

a) "qualidade" do banco de segmentos;

b) técnicas de concatenação e alteração prosódica.

Por sua vez, a montagem do banco de unidades, num estágio anterior à concatenação, envolve etapas de:

a) escolha dos segmentos;

b) definição do *corpus*[9] de extração das unidades;

c) gravação;

d) segmentação.

Essas etapas são mostradas na Fig. 4.9.

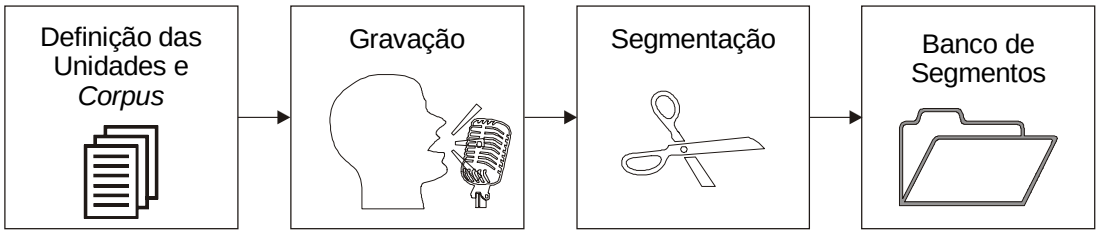

Fig. 4.9: Etapas envolvidas na criação de um banco de unidades para síntese concatenativa.

[9] Conjunto de enunciados, colhidos de forma controlada, que servem como material para extração das unidades.

4.3.1 Escolha de Unidades

Um dos aspectos mais importantes na síntese por concatenação é a escolha apropriada do tipo de segmento, isto é, da unidade base para a geração da fala. A seleção depende normalmente do compromisso entre tamanho do segmento e número de unidades. Segmentos maiores exigem menos pontos de concatenação e maior controle sobre a coarticulação, aumentando a qualidade da síntese. Mas, ao mesmo tempo, determinam um grande número de unidades, o que pode se tornar impraticável devido a limitações de memória dos sistemas. As unidades devem ser escolhidas de forma que permitam:

a) pequena distorção na concatenação;
b) captura da maior quantidade possível de efeitos coarticulatórios;
c) número e tamanho reduzidos.

Estes fatores são conflitantes e um compromisso entre eles deve ser buscado. Na Tabela 4.2, são mostrados os tipos de unidades mais comuns, juntamente com informações sobre tamanho do segmento, número de unidades necessárias e qualidade alcançada. Uma descrição mais detalhada de cada tipo de unidade será dada a seguir.

Tabela 4.2: Relação entre tamanho, número e qualidade para diferentes tipos de unidades (adaptado de [17])

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

4.3.1.1 Palavras

Se forem ignoradas unidades longas como orações ou frases, a primeira escolha como unidade de síntese talvez seja a palavra. É a primeira sugestão de pessoas não ligadas à área [44]. A concatenação é relativamente simples de ser realizada e os efeitos de

coarticulação são reduzidos. Entretanto, para um sistema de síntese de texto irrestrito, é praticamente impossível gravar todas as palavras de um idioma, incluindo nomes próprios e flexões. Além disso, há uma diferença muito grande de fluência em palavras ditas de forma isolada ou de forma contínua em uma frase [15]. A concatenação de palavras e frases é utilizada somente nos sistemas de resposta vocal com um domínio bem definido. Numa aplicação de saldo bancário, por exemplo, é possível gravar uma frase base “Seu saldo é de” completada por números gravados isoladamente. Para aplicações de domínio irrestrito, entretanto, faz-se necessário o uso de segmentos menores.

4.3.1.2 Fonemas Independentes de Contexto

No outro extremo em relação ao tamanho da unidade, encontra-se o fonema. Um fonema é o mínimo segmento distintivo da fala. Por exemplo, em português, [p] é um fonema e [t] é outro, pois diferenciam palavras como “pato” e “tato”. Um sistema bastante compacto e generalizável, capaz de produzir qualquer frase, pode ser obtido com o uso de fonemas. Considerando a independência do contexto fonético vizinho, para um idioma com N fonemas, seriam necessárias apenas N unidades. Para o inglês, aproximadamente 40 segmentos fariam parte do inventário de unidades e para o português, 33 a 34, incluindo as vogais nasais [24,32]. Infelizmente, fonemas independentes de contexto não produzem um resultado aceitável devido aos efeitos coarticulatórios e às muitas descontinuidades.

4.3.1.3 Díades (Difones e Demissílabas)

A díade, na forma de difone ou demissílaba, tem sido extensivamente utilizada em sistemas de síntese concatenativa. Uma díade é composta pela parte final de uma unidade fonética e pela porção inicial da unidade fonética seguinte [45], capturando a transição entre fonemas. Apesar do termo difone normalmente ser utilizado como um sinônimo para díade, os termos podem ser diferenciados [18]. O difone, assumindo a

unidade fonética da definição da díade como o fonema, é o segmento obtido do meio de um fonema até o meio do fonema seguinte [17,18].

Demissílabas representam a metade inicial e final de sílabas [44]. Em alguns casos, a demissílaba aproxima-se muito do difone, diferenciando-se apenas pelo ponto de corte da unidade. Em uma sílaba CVC (consoante-vogal-consoante), dois difones são obtidos, um CV (consoante-vogal) e outro VC (vogal-consoante). Nessa mesma sílaba, duas demissílabas também são obtidas. O ponto de corte do difone é feito na área estável, no meio da vogal. O corte da demissílaba é realizado no ponto em que a transição consoante-vogal termina e se inicia a porção estável da vogal [45]. Na Fig. 4.10, é mostrada a segmentação de uma palavra usando diferentes tipos de segmentos.

A abordagem por demissílabas assume que a coarticulação é minimizada nos limites da sílaba, enquanto que o modelo por difones assume que o efeito coarticulatório é minimizado no centro acústico do fonema [18,46]. A validade dessas suposições é dependente do idioma em questão. Alguns sistemas aproveitam as melhores características de cada um dos modelos usando um inventário misto de unidades [46].

Em uma língua com N fonemas, há aproximadamente N^2 difones potenciais. Mas na prática, algumas combinações de fonemas nunca ocorrem. O número médio de difones nos sistemas comerciais para diferentes idiomas está na faixa de 1200 a 1500.

Apesar dos difones capturarem a transição entre os fonemas, distorções podem ocorrer devido às diferenças espectrais entre unidades obtidas de diferentes contextos.

4.3.1.4 Trifones

Trifones são unidades usadas com sucesso em reconhecimento de fala que podem ser utilizadas também em síntese de fala. Em uma sílaba CVC, um trifone engloba a transição CV, a vogal V e a transição VC [45]. Pode ser considerado também como um fonema com um contexto específico anterior e posterior. Apesar do número teórico de

trifones ser N^3 para uma língua com N fonemas, muitas combinações não ocorrem na prática. Para o inglês, o número inicial de 74 mil (42^3) é reduzido para 25 mil [17].

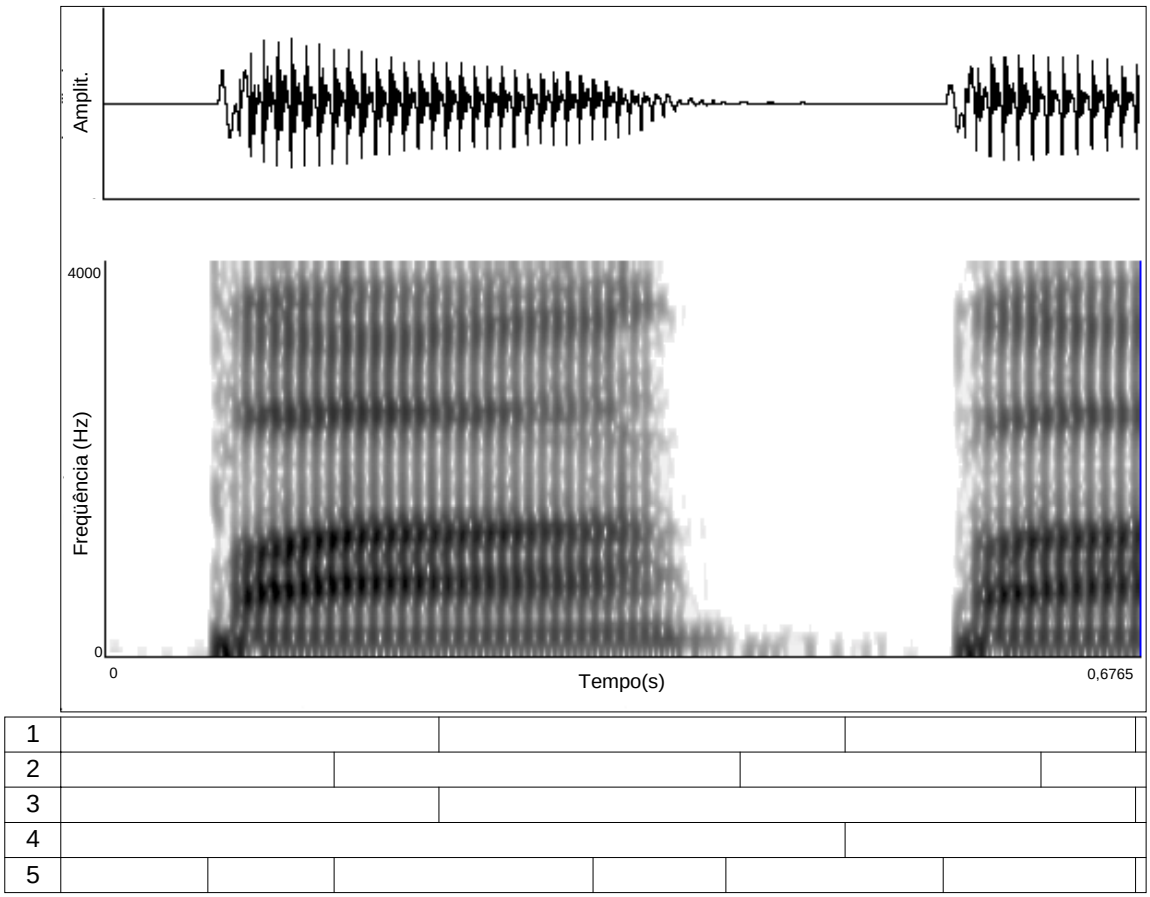

Fig. 4.10: Segmentação com diferentes abordagens: 1-Difones; 2- Demissílabas; 3- Segmentos VCV; 4- Segmentos CVC e 5- Segmentos sub-fonêmicos.

4.3.1.5 Segmentos Sub-Fonêmicos

Ao contrário das unidades que capturam a transição entre fonemas, mantendo-a intacta, pode-se segmentar cada porção estacionária ou transitória dos fonemas levando à geração de segmentos sub-fonêmicos. Esta abordagem, relativamente nova, tem levado a alguns resultados animadores. A grande vantagem é a possibidade de utilização em

sistemas com forte restrição de memória e processamento, já que o número de unidades é relativamente baixo (da ordem de 400) [45].

4.3.2 *Corpus* de Extração das Unidades

As unidades acústicas são normalmente extraídas de um *corpus* criado especialmente para a tarefa, por permitir um grau de controle adequado. Devido à grande variação de freqüência de ocorrência das unidades na fala natural, tentar obtê-las a partir da gravação de um texto qualquer é praticamente inviável. Algumas seriam repetidas várias vezes e outras não apareceriam. Como exemplo, em [11] é citado o caso do francês, em que um dado conjunto de 100 frases foneticamente balanceadas cobre apenas 43% dos 1200 difones requeridos, com uma redundância de aproximadamente 80%.

Assim como a obtenção das unidades a partir da gravação de um texto qualquer é díficil, a gravação dos segmentos isolados também não é adequada, pois os padrões de entonação, ritmo e intensidade da fala são muito diferentes da fala contínua. Desta forma, os segmentos devem ser inseridos em algum contexto que seja neutro e minimize os efeitos de coarticulação. Normalmente, é utilizada uma frase veículo, com o objetivo de prover um contexto neutro. Um exemplo de frase veículo é “digo (palavra com o segmento) pra ele”.

Em relação à palavra que contém o segmento desejado, pode-se usar palavras existentes na língua ou logatomas (vocábulos sem sentido). O difone [bi], por exemplo, poderia ser retirado do logatoma “pabipá” ou da palavra “bebida”. Algumas bibliografias indicam que logatomas facilitam a segmentação [11] e reduzem os efeitos de coarticulação [34]. Mas bons resultados são obtidos na literatura usando unidades retiradas de palavras comuns [11,47]. A tonicidade também é uma questão debatida. Sílabas tônicas são mais longas e estariam menos sujeitas à coarticulação. As átonas, por outro lado, são mais comuns na fala. A utilização de um inventário composto unicamente por átonas poderia aumentar a qualidade (porque são mais freqüentes e não seria necessário alterá-las) e reduzir as exigências de memória, já que são mais curtas [11]. A velocidade da fala na

gravação define um compromisso entre inteligibilidade e naturalidade. Se o locutor falar mais rápido, as unidades incorporarão uma boa fluência, mas com uma menor inteligibilidade. Se falar mais devagar, o resultado será mais inteligível, porém, menos natural, devido à articulação exagerada.

4.3.3 Processo de Segmentação

Após a escolha do tipo de unidade utilizado e da gravação do *corpus*, é realizada a segmentação das unidades. Este processo requer um conhecimento profundo de parâmetros fonéticos e fonológicos, deixando quase que como única alternativa a segmentação manual. Nesse caso, um especialista, através de ferramentas de visualização de sinais, deve escolher os pontos inicial e final de cada segmento levando em consideração aspectos como a estabilidade dos formantes e a inexistência de ruídos. É uma tarefa bastante complexa e sujeita a erros quando realizada de forma manual.

Técnicas automáticas poderiam acelerar o processo de segmentação e reduzir os erros. O desenvolvimento de técnicas de reconhecimento de fala está sendo aproveitado nesta etapa de processamento dos sistemas de síntese de fala. Artigos não muito atuais (1992-93) mostravam que era possível utilizar segmentação automática com 85% de sucesso [11], mas sem dispensar um ajuste manual. Um artigo mais recente [48] aponta que com segmentação automática baseada em modelos ocultos de Markov (HMM), utilizando modelos dependentes de locutor, é obtido um resultado melhor do que com uma cuidadosa segmentação manual.

Qualquer que seja o método de segmentação, manual ou automático, deve-se levar em consideração as peculiaridades da fala do locutor. Não é possível segmentar sem o estabelecimento de uma referência em relação à voz utilizada.

Com um inventário de unidades pronto, a última etapa para a geração de fala sintetizada é a de concatenação e alteração dos parâmetros prosódicos, a qual é discutida no próximo capítulo.

Capítulo 5

Técnicas de Alteração de Parâmetros Prosódicos

Em sistemas de conversão texto-fala baseados em síntese concatenativa, métodos de alteração de parâmetros prosódicos (em especial, *pitch* e duração) são essenciais para atender, principalmente, aos requisitos de naturalidade.

Técnicas de processamento de sinais, de diferentes abordagens, são empregadas para realizar estas alterações e serão apresentadas neste capítulo.

5.1 Considerações Iniciais

Nesta seção, as modificações na escala de tempo e de *pitch* serão analisadas tomando por base o modelo fonte-filtro. A análise enfoca o caso de sinais vozeados e é baseada nas literaturas [4,5].

No modelo fonte-filtro de produção da fala, o filtro variante no tempo $H(z)$ modela o efeito do pulso glotal e das características de transmissão do trato vocal. A relação entrada-saída deste sistema pode ser caracterizada pela resposta ao impulso unitário $h(n,m)$ variante no tempo ou pela resposta em freqüência associada

$$\sum_{m=-\infty}^{\infty} h(n,m)e^{-j\omega m} = H(n,\omega)e^{j\psi(n,\omega)} \tag{5.1}$$

onde $h(n,m)$ é definida como a resposta do sistema no instante n a um impulso aplicado na amostra m. $H(n,\omega)$ e $\psi(n,\omega)$ são, respectivamente, a magnitude e a fase variantes no tempo do sistema. Para pequenos intervalos de tempo, entretanto, $h(n,m)$ pode ser

considerada como aproximadamente constante, representando um sistema quase-estacionário.

Para sinais vozeados, a excitação $u(n)$ é formada pela soma de exponenciais complexas harmonicamente relacionadas com amplitude unitária, fase inicial zero e uma freqüência fundamental $f_k(n) = \omega_k / 2\pi = 1/P(n)$.

Então, o sinal de excitação pode ser expresso como:

$$u(n) = \sum_{k=0}^{P(n)-1} e^{j\Phi_k(n)} \tag{5.2}$$

$$\Phi_k(n) = \sum_{m=0}^{n} \omega_k(m) = \sum_{m=0}^{n} \frac{2\pi k}{P(m)} \tag{5.3}$$

onde $P(n)$ é o período de *pitch* do sinal de fala. Como $P(n)$ é aproximadamente constante em torno de um instante de tempo qualquer n, a fase da excitação $\Phi_k(m)$ pode ser aproximada por

$$\Phi_k(m) \approx \Phi_k(n) + \frac{2\pi k}{P(n)}(m-n), \quad \text{para } |m-n| \text{ pequeno.} \tag{5.4}$$

Da teoria de sistemas lineares, o sinal vozeado $x(n)$ obtido na saída do sistema $h(n,m)$ é dado por:

$$x(n) = \sum_{m=-\infty}^{\infty} h(n,m)u(n-m) \tag{5.5}$$

Assumindo que o período de *pitch* $P(n)$ é aproximadamente constante em relação à duração de $h(n,m)$, o sinal de excitação pode ser aproximado por sua representação harmônica local, obtendo-se o sinal $x(n)$ como:

$$x(n) = \sum_{k=0}^{P(n)-1} H(n,\omega_k(n))e^{j(\Phi_k(n)+\psi(n,\omega_k(n)))} \tag{5.6}$$

$$x(n) = \sum_{k=0}^{P(n)-1} A_k e^{j\theta_k(n)}, \tag{5.7}$$

onde $A_k(n)$ é a amplitude da *k*-ésima harmônica, sendo igual à amplitude do sistema na freqüência harmônica $\omega_k(n)$:

$$A_k(n) = H(n, \omega_k(n)) \tag{5.8}$$

A fase $\theta_k(n)$ é definida como a fase instantânea da *k*-ésima harmônica, sendo formada pela soma da fase da excitação $\Phi_k(m)$ e da fase do sistema $\psi(n, \omega_k(n))$:

$$\theta_k(n) = \Phi_k(n) + \psi(n, \omega_k(n)) \tag{5.9}$$

Considerando a fase do sistema $\psi(n, \omega_k(n))$ como uma função de variação lenta em n, a fase instantânea pode ser aproximada por:

$$\theta_k(m) = \theta_k(n) + \omega_k(n)(m-n) \text{ para } |m-n| \text{ pequeno.} \tag{5.10}$$

5.1.1 Modificação na Escala de Tempo

O objetivo da modificação na escala de tempo é alterar a taxa aparente de articulação da fala sem variar o conteúdo espectral do sinal. Esta definição é importante, pois uma mudança na escala de tempo de um sinal $x(n)$ de duração original Δ pode levar a um deslocamento de freqüência correspondente. Numa experiência simples, em que um sinal de fala $x(n)$ é reproduzido em uma velocidade maior do que a de gravação, o som resultante é distorcido por um aumento de *pitch*. Se o sinal for reproduzido em uma velocidade mais lenta, o *pitch* é reduzido. Não só o *pitch* é alterado, mas também o timbre da voz. A simples alteração da velocidade de reprodução prejudica a compreensão da mensagem, não sendo uma alternativa aceitável de escalamento de tempo.

Para compreender a dualidade entre o escalamento no tempo e o deslocamento em freqüência, considera-se um sinal $y(n) = x(\alpha n)$, que corresponde a um sinal original

$x(n)$ reproduzido a uma velocidade α vezes maior ou menor do que a de gravação. Assim, a duração original Δ passa a Δ/α e da equação (5.7), obtemos

$$y(n) = \sum_{k=0}^{P(\alpha n)-1} A_k(\alpha n) e^{j\theta_k(\alpha n)} \tag{5.11}$$

$$\theta_k(\alpha m) \approx \theta_k(\alpha n) + \alpha \frac{2\pi k}{P(\alpha n)}(m-n) \quad \text{para } |m-n| \text{ pequeno} \tag{5.12}$$

Para o caso de $\alpha < 1$, a compressão da escala de tempo, além de realizar a compressão da evolução temporal do *pitch* $P(n)$ e das amplitudes harmônicas $A_k(n)$, comprime a evolução temporal da fase instantânea $\theta_k(n)$, o que implica na expansão da escala de freqüência por α. Isso significa que a modificação da escala de tempo de um sinal por um dado fator implica na modificação das características freqüenciais pelo inverso desse fator. Esse escalamento duplo não é conveniente. Outra forma de escalamento no tempo é necessária, de forma que a versão modificada do sinal seja percebida como a mesma seqüência de eventos acústicos apenas reproduzida em outra velocidade. As modificações na escala de tempo devem alterar apenas a estrutura temporal do sinal sem variar suas características freqüenciais.

Formalmente, as modificações na escala de tempo são especificadas definindo um mapeamento $n \to n' = D(n)$ entre a escala original e a alterada. $D(n)$ é definida como a função de deformação temporal. É conveniente especificar uma taxa de modificação de tempo contínua e variante no tempo $\beta(t)$, de onde a função de deformação temporal possa ser derivada como

$$n \to n' = D(n) = \frac{1}{T}\int_0^{nT} \beta(\tau) d\tau \tag{5.13}$$

T representa o período de amostragem e $\beta(t) > 0 \forall t$ corresponde a uma função de deformação estritamente monotônica. Se $\beta(t) = \beta$ é constante, a função de deformação é linear.

Um escalamento temporal com $\beta(t) > 1$ corresponde a um decréscimo na taxa de articulação, enquanto com $0 < \beta(t) < 1$, a um acréscimo. Em relação ao modelo de fala vozeada, os parâmetros devem ser modificados da seguinte forma:

$$P'(n') = P(D^{-1}(n')) \tag{5.14}$$

$$A'_k(n') = H'(n', \omega_k(n')) = H'(D^{-1}(n'), \omega_k(D^{-1}(n'))) \quad 0 \le k \le P'(n') - 1 \tag{5.15}$$

$$\theta'_k(n') = \Phi'_k(n') + \psi\left(D^{-1}(n'), \frac{2\pi k}{P(D^{-1}(n'))} \right) \tag{5.16}$$

$$\Phi'_k(n') = \sum_{m=0}^{n'} \frac{2\pi k}{P(D^{-1}(m))} \tag{5.17}$$

$D^{-1}(.)$ representa o mapeamento inverso. Estas equações expressam que:

a) o contorno de *pitch* $P'(n')$ do sinal modificado é uma versão escalada em tempo do contorno de *pitch* original;
b) a função do sistema de magnitude $H'(n', \omega)$ e fase $\psi'(n', \omega)$ é uma versão escalada em tempo da função original;
c) as freqüências instantâneas no instante n' correspondem às freqüências instantâneas do sinal original no instante de tempo $D^{-1}(n')$.

5.1.2 Modificação na Escala de *Pitch*

Em analogia à modificação temporal, o objetivo da modificação da escala de *pitch* é alterar a freqüência fundamental da fala sem afetar o envelope espectral variante no tempo. Para a especificação de uma transformação na escala de *pitch*, é conveniente definir um fator de modificação de *pitch* variante no tempo $\alpha(n) > 0$ que transforma o contorno de *pitch* $P(n)$ em $P'(n) = P(n)/\alpha(n)$.

Para o fator $\alpha(n) > 1$, o *pitch* local é aumentado (o período de *pitch* é diminuído), enquanto, para $\alpha(n) < 1$, o *pitch* é reduzido. Em relação ao modelo de fala, os parâmetros devem ser modificados da seguinte forma:

$$P'(n') = \frac{P(n')}{\alpha(n')} \tag{5.18}$$

$$A'_k(n') = H'_k(n') = H(n', \alpha(n')\omega_k(n')) \tag{5.19}$$

$$\theta'_k(n') = \Phi'_k(n') + \psi(n', \alpha(n')\omega_k(n')) \tag{5.20}$$

$$\Phi'_k(n') = \sum_{m=0}^{n'} \alpha(m)\omega_k(m) \tag{5.21}$$

Estas equações expressam que:

a) o contorno de *pitch* é escalado pelo fator $\alpha(n')$;

b) as magnitudes das harmônicas modificadas em *pitch* são amostras da função de magnitude do sistema tomadas nas freqüências harmônicas modificadas.

De forma oposta à modificação na escala de tempo, a modificação na escala de *pitch* requer a estimação das amplitudes do sistema $H(n, \alpha(n)\omega_k(n))$ e fases $\psi(n, \alpha(n)\omega_k(n))$ nas freqüências $\alpha(n)\omega_k(n)$, que não são necessariamente as freqüências harmônicas de *pitch* no sinal original. Algumas técnicas de alteração de *pitch* decompõem o sinal de fala em um sinal de excitação com envelope espectral plano e em um envelope espectral variante no tempo.

5.1.3 Abordagens para Modificações Prosódicas

Para a alteração de parâmetros prosódicos, uma abordagem comum consiste na análise do sinal de fala para a obtenção de parâmetros de entrada, na aplicação da transformação desejada sobre os parâmetros e na síntese do sinal correspondente.

Este processo pode ser realizado através de:

a) modelos que realizam a decomposição fonte-filtro, como a análise LPC;

b) utilização de técnicas não paramétricas, baseadas apenas na representação tempo-freqüência do sinal de fala. Um conjunto dessas técnicas está classificado sob o nome de síntese por recobrimento e soma (*Overlap-and-Add*). Uma variação da técnica *Overlap-and-Add* (OLA) é apresentada a seguir.

5.2 PSOLA (Pitch-Synchronous Overlap-and-Add)

A técnica PSOLA (*Pitch-Synchronous Overlap-and-Add*) pertence à classe de métodos baseados no processo de análise e síntese por recobrimento e soma [4], permitindo alterações na escala de tempo (duração) e de *pitch*. A técnica consiste de três etapas: (a) análise, com a decomposição do sinal em segmentos de curta duração, (b) modificação opcional destes segmentos e (c) síntese através da sobreposição e soma dos segmentos [4,49].

5.2.1 Etapa de Análise

No processo de análise, é realizada a segmentação do sinal original de fala. Considerando-se $x(n)$ o sinal de fala amostrado e $h_a(n)$ a janela de análise, assume-se que $h_a(n)$ é:

a) centrada em torno da amostra zero;

b) de duração finita T_a e simétrica;

c) a resposta ao impulso de um filtro passa-baixas.

Os pontos de localização das janelas de análise são chamados de instantes de análise $t_a(n)$. O sinal de análise de tempo finito $x_a(n)$, associado aos instantes de análise

$t_a(n)$, é definido como o produto do sinal de fala $x(n)$ e da janela de análise centrada em $t_a(n)$:

$$x_a(n) = h_a(t_a - n)x(n) \tag{5.22}$$

O sinal de tempo finito pode ser associado a uma representação freqüencial $X_a(\omega)$, definida como a transformada discreta de Fourier de $x_a(n)$:

$$X_a(\omega) = \sum_{n=-\infty}^{\infty} h_a(t_a - n)x(n)e^{-j\omega n} \tag{5.23}$$

Na técnica PSOLA, os instantes de análise $t_a(n)$ são dispostos em uma taxa síncrona com o *pitch* nas regiões vozeadas do sinal - daí a primeira parte do nome da técnica - e em uma taxa constante nas áreas não-vozeadas.

Na Fig. 5.1, é mostrado o processo de análise em um segmento vozeado.

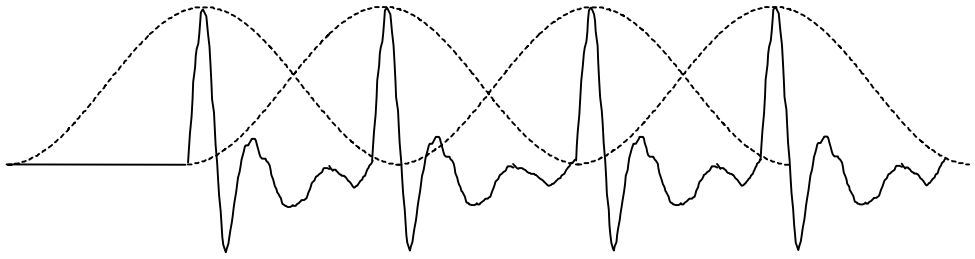

Fig. 5.1: Janelamento efetuado na etapa de análise da técnica PSOLA.

5.2.2 Etapa de Modificação

Nesta etapa, é realizada a transformação da seqüência de sinais de análise em uma seqüência de sinais de síntese. Os sinais de síntese $x_s(n)$ são sincronizados com um novo conjunto de marcas de *pitch*, t_s, as marcas de síntese. Os instantes de síntese t_s dependem dos fatores de modificações nas escalas de *pitch* e duração, respectivamente β e γ. O intervalo $t_s - t_{s-1}$ entre duas marcas de *pitch* sucessivas deve ser igual ao período de

pitch de síntese. Um mapeamento $t_s \rightarrow t_a$ é realizado, especificando quais segmentos de análise $x_a(n)$ devem ser selecionados para gerar cada sinal de síntese $x_s(n)$. A variante *Time-Domain Pitch-Synchronous Overlap-and-Add* (TD-PSOLA) obtém os sinais de síntese $x_s(n)$ através de uma cópia dos sinais de análise $x_a(n)$ correspondentes. O algoritmo consiste na escolha de um número de sinais de análise transladados por uma seqüência de atrasos $\delta_s = t_s - t_a$. Assim,

$$x_s(n) = x_a(n - \delta_s) = x_a(n + t_a - t_s) \tag{5.24}$$

Na variante *Frequency-Domain Pitch-Synchronous Overlap-and-Add* (FD-PSOLA), a representação freqüencial $X_a(\omega)$ dos sinais de análise $x_a(n)$ é utilizada para a geração dos sinais de síntese. Devido à maior complexidade computacional, a técnica FD-PSOLA usualmente não é empregada.

5.2.3 Síntese

Na última etapa, o sinal sintetizado $\tilde{x}(n)$ é obtido pela combinação dos sinais de síntese sincronizados com as marcas de síntese t_s. O processo geral de síntese por sobreposição e soma [4] descrito pela equação (5.25) é utilizado.

$$\tilde{x}(n) = \frac{\sum_s \alpha_s x_s(n) h_s(t_s - n)}{\sum_s h_s^2(t_s - n)} \tag{5.25}$$

O fator α_s é introduzido para compensar as modificações de energia resultantes das alterações de *pitch e* h_s representa as janelas de síntese. O uso desta equação de síntese leva à minimização do erro quadrático entre os espectros dos sinais de análise $x_a(n)$ e os sinais de síntese $x_s(n)$ [50].

Como a técnica no domínio do tempo TD-PSOLA não modifica nenhum sinal de análise $x_a(n)$, é possível utilizar a mesma janela de análise e síntese $h_a = h_s$.

Assumindo-se $\alpha_s = 1$, a síntese se reduz à equação simplificada de sobreposição e soma:

$$\tilde{x}(n) = \sum_s x_s(n) \tag{5.26}$$

O sinal sintetizado é uma simples combinação linear de segmentos janelados e transladados do sinal original. Com exceção da etapa de janelamento, todas as operações são lineares.

5.2.4 Modificação de Duração

Para o caso de alteração de duração com um fator constante β, o mapeamento $t_s \rightarrow t_a$ associa cada instante de síntese t_s ao instante de análise t_a mais próximo de $(\beta \cdot t_s)$. Se o objetivo é reduzir a duração, sinais de análise $x_a(n)$ são eliminados. Para o aumento de duração, sinais de análise são repetidos, como mostrado na Fig. 5.2.

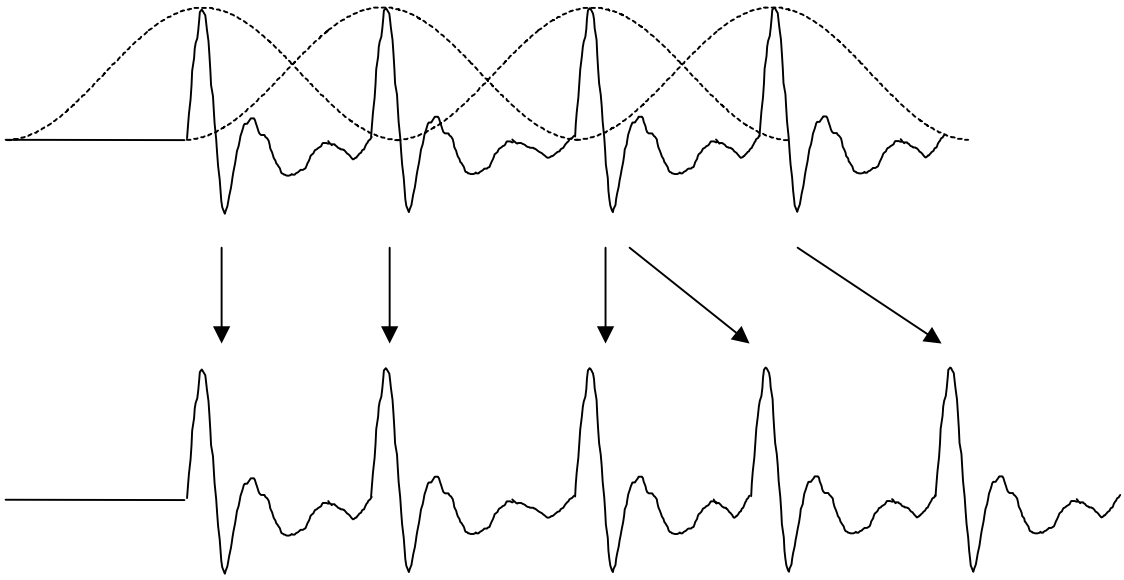

Fig. 5.2: Modificação de duração com *PSOLA*.

Para fatores moderados de alteração de duração, as distorções acústicas são desprezíveis. Porém, em regiões não-vozeadas, um aumento de duração com fator igual ou maior do que dois, cria um efeito de ruído. Isto se deve ao fato da repetição dos sinais de curta duração introduzir uma forte correlação de curta duração no sinal sintetizado [6]. Em síntese de fala, entretanto, dificilmente são utilizados fatores desta ordem.

5.2.5 Modificação de *Pitch*

A modificação do *pitch* por um fator α é obtida pela alteração dos intervalos de tempo entre sinais de síntese sucessivos, como é mostrado na equação (5.27).

$$t_{s+1} - t_s = \frac{(t_{a+1} - t_a)}{\alpha} \tag{5.27}$$

As modificações de *pitch* são um pouco mais complexas, pois interferem nas modificações de duração. O caso mais simples é o de modificações simultâneas de duração e *pitch* pelos mesmos fatores $\beta = \alpha$. Nesse caso, o mapeamento entre as marcas de síntese e de análise é de um para um e não há necessidade de ajustes de duração. Para quaisquer outros valores, períodos devem ser repetidos ou excluídos para manter a duração requerida. Na Fig. 5.3, é mostrado um exemplo de aumento de *pitch* usando a técnica *PSOLA*.

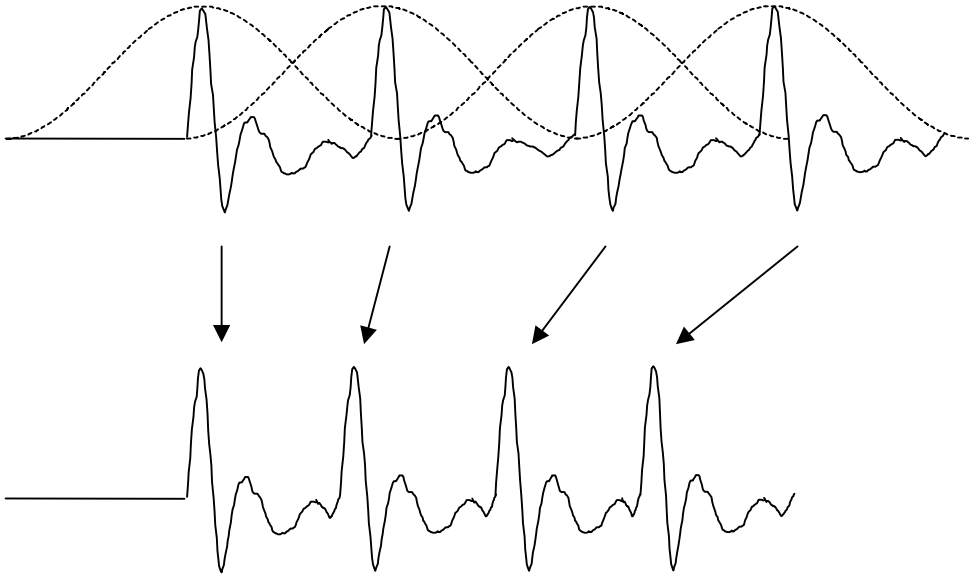

Fig. 5.3: Modificação de *pitch* usando *PSOLA*.

5.2.6 Limitações

A técnica PSOLA permite realizar modificações prosódicas em um segmento de fala praticamente sem perda de qualidade. As limitações relacionadas a alterações dentro do segmento são as seguintes [34]:

a) as modificações de duração só ocorrem em passos do tamanho do período de *pitch*;

b) alterações de *pitch* conduzem a mudanças na duração que devem ser compensadas de forma adequada;

c) em porções não-vozeadas dos segmentos, o aumento de duração efetuado com a repetição de quadros de análise introduz uma periodicidade que gera um efeito “metálico” à fala sintetizada;

d) para grandes alterações de *pitch*, a envoltória resultante da sobreposição das janelas se afasta de um nível constante, causando distorções.

Outros problemas ocorrem na concatenação de segmentos. Nos limites das unidades, extraídas de palavras diferentes, podem ocorrer descasamentos relacionados à fase, *pitch*, envelope espectral e amplitude [11,17].

- Descasamento de Fase

Descontinuidades podem surgir se as janelas de análise não estiverem posicionadas nos mesmos pontos dentro do período de *pitch* em diferentes unidades, como mostrado na Fig. 5.4.

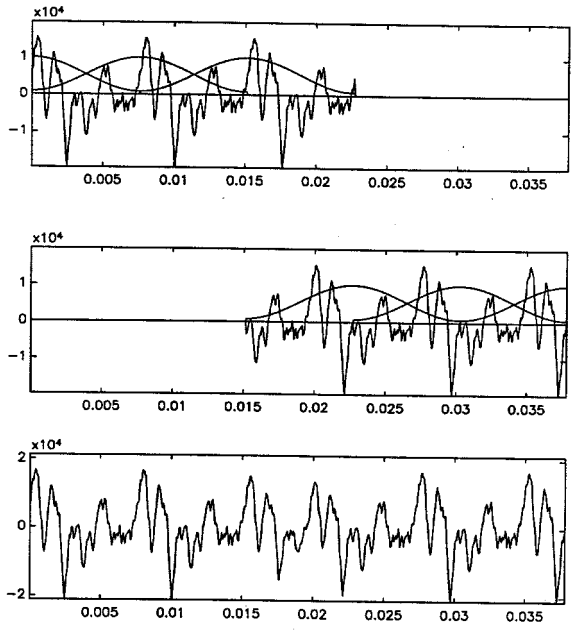

Fig. 5.4: Descasamento de fase [11].

Uma marcação de *pitch* extremamente precisa, buscando o instante de fechamento da glote como centro da janela, poderia resolver o problema. Entretanto, uma marcação deste tipo não é conseguida com métodos automáticos e consome muito tempo se realizada manualmente. Além disso, mesmo manualmente não há garantias de que não ocorrerão erros, pois, uma vez que a primeira marca do primeiro período tenha sido

determinada, todas as outras estarão impostas por propagação. Em [11], foram obtidos bons resultados com o auxílio de um sistema de marcação de *pitch* combinando técnicas automáticas e correção manual. A coerência de fase é difícil de ser obtida, mas possível com o emprego de uma metodologia bem definida.

- Descasamento de *Pitch*

Segmentos com mesma envoltória espectral e "janelados" em posições relativas coerentes, mas com períodos de *pitch* diferentes podem trazer problemas na concatenação. Na Fig. 5.5, é mostrado esse tipo de descasamento. Em um banco de unidades com milhares de segmentos é muito difícil evitar problemas dessa natureza. Uma minimização pode ser tentada, gravando o *corpus* de extração das unidades com um *pitch* constante, o que exige muito treinamento do locutor.

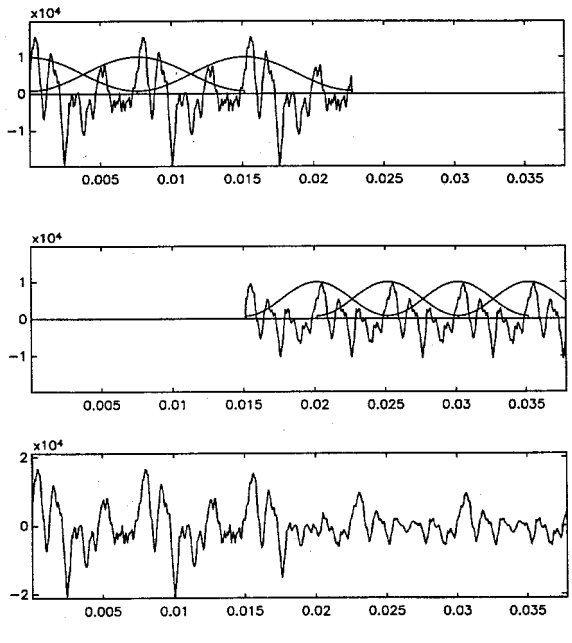

Fig. 5.5: Descasamento de *pitch* [11].

- Descasamento de Envoltória Espectral

Na implementação mais comum, no domínio do tempo, a técnica não provê meios de adaptar as envoltórias espectrais dos segmentos. Essa incoerência espectral, mostrada na Fig. 5.6, está relacionada aos efeitos de coarticulação e à variabilidade natural da fala.

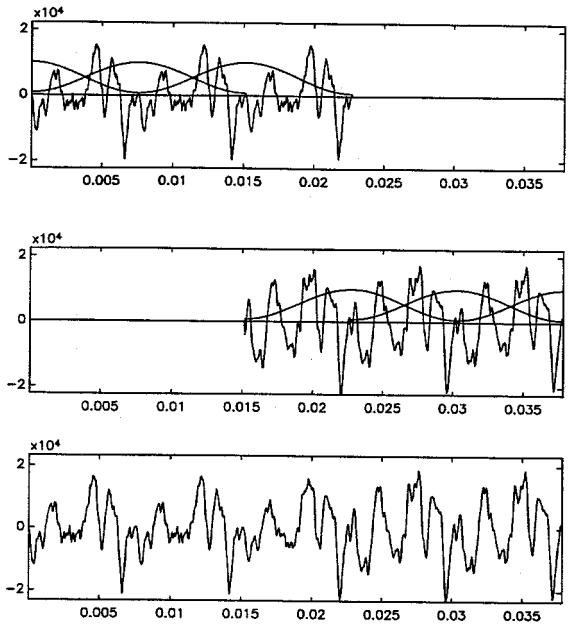

Fig. 5.6: Descasamento espectral [11].

A utilização da técnica FD-PSOLA, no domínio da freqüência, pode minimizar o problema, mas com um custo computacional bem maior. Com TD-PSOLA, a alternativa é testar a concatenação das unidades uma a uma e rejeitar as que tiverem um descasamento muito grande.

- Descasamento de Amplitude

Uma incoerência de amplitude entre diferentes segmentos pode ser corrigida com uma simples equalização. A dificuldade está na escolha dos parâmetros (potência,

energia, pico do sinal) que devem ser considerados para o cálculo do fator de equalização. Além disso, o timbre provavelmente se altera com diferentes níveis de volume [17].

5.3 Análise/Ressíntese usando LPC (Linear Predictive Coding)

5.3.1 Teoria Básica

A análise/ressíntese LPC (*Linear Predictive Coding*) é uma técnica originalmente desenvolvida para codificação de fala baseada no modelo fonte-filtro de produção da fala (apresentado na Seção 2.3). Em codificação, o processo de análise/transmissão/ressíntese tem como objetivo sintetizar um sinal de saída o mais próximo possível do sinal de entrada através da transmissão dos parâmetros da fonte e dos coeficientes do filtro que representa o trato. Em síntese de fala, o objetivo da utilização da técnica LPC é permitir as alterações nas escalas de duração e *pitch* sem degradar o sinal.

O modelo fonte-filtro de produção da fala considerado para a técnica LPC, reapresentado na Fig. 5.7, é composto por uma fonte de excitação para sinais vozeados e não-vozeados, uma chave de decisão vozeado/não-vozeado, um fator de ganho e filtro digital variante no tempo.

O sinal de excitação é modelado por um trem de impulsos para sinais vozeados e um ruído aleatório para sinais não-vozeados. O filtro digital representa as características do trato vocal, a forma do pulso glotal e os efeitos de radiação nos lábios [26].

O conceito em que se baseia a técnica LPC é de que cada amostra do sinal de fala pode ser aproximada por uma combinação linear das últimas P amostras [51]. Minimizando a diferença quadrática entre as amostras originais e as amostras linearmente preditas, os coeficientes de um filtro preditor $H(z)$ podem ser determinados.

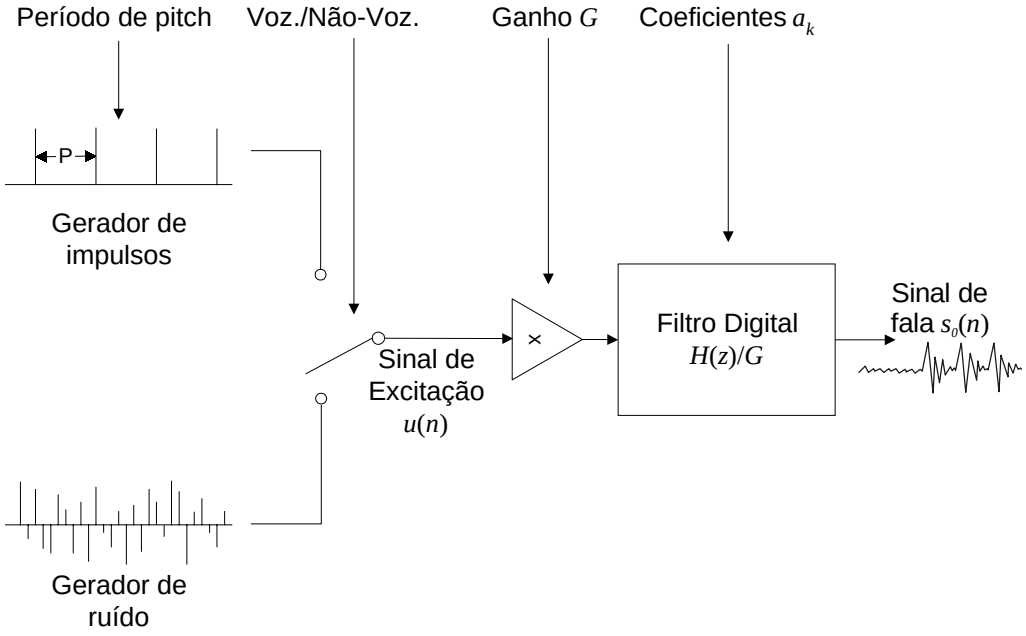

Fig. 5.7: Modelo fonte-filtro usado na técnica LPC.

Em relação ao modelo fonte-filtro, seja $H(z)$ a transformada Z da resposta ao impulso do filtro digital, $U(z)$ a transformada Z da excitação $u(n)$, $S_0(z)$ a transformada Z do sinal de fala sintetizado $s_0(n)$ e G, o ganho da excitação. Então:

$$S_0(z) = U(z) \cdot H(z) \tag{5.28}$$

Como $H(z)$ apresenta somente pólos (modelagem simplificada do comportamento do trato), tem-se:

$$H(z) = \frac{G}{1 - \sum_{k=1}^{P} a_k z^{-k}}, \tag{5.29}$$

onde a_k são os coeficientes do filtro.

O sinal de fala sintetizado é dado por:

$$s_0(n) = \sum_{k=1}^{P} a_k s_0(n-k) + Gu(n) \tag{5.30}$$

Definindo-se $s(n)$ como o sinal original e $\hat{s}(n)$ como uma estimativa desse sinal, obtida com um preditor linear de ordem P, tem-se:

$$\hat{s}(n) = \sum_{k=1}^{P} a_k s(n-k) \tag{5.31}$$

O erro de predição ou resíduo é definido como a diferença entre o sinal $s(n)$ e o seu valor estimado $\hat{s}(n)$:

$$e(n) = s(n) - \hat{s}(n) \tag{5.32}$$

$$e(n) = s(n) - \sum_{k=1}^{P} a_k s(n-k) \tag{5.33}$$

Definindo-se $E(z)$ como a transformada Z de $e(n)$ e $S(z)$ a transformada Z de $s(n)$, pode-se escrever:

$$A(z) = \frac{E(z)}{S(z)} \tag{5.34}$$

$$A(z) = 1 - \sum_{k=1}^{P} a_k z^{-k} \tag{5.35}$$

$A(z)$ é a função de transferência do filtro de erro de predição, chamado filtro inverso.

Se o modelo de produção da fala fosse exato, $s_0(n)$ seria igual a $s(n)$ e o erro de predição seria igual ao sinal de excitação:

$$e(n) = Gu(n) \tag{5.36}$$

Como o modelo não é exato, existe uma diferença entre o erro $e(n)$ e a excitação $Gu(n)$. Se o sinal de resíduo $e(n)$ for utilizado como excitação do filtro $H(z)/G$, o sinal sintetizado $s_0(n)$ será igual ao sinal original $s(n)$.

No método de mínimo erro quadrático, os coeficientes a_k são obtidos pela minimização do sinal de erro $e(n)$ em relação a cada um dos coeficientes. Existem duas

técnicas principais: o método da autocorrelação, em que é realizado um janelamento do sinal de fala e o erro é minimizado no intervalo $0 \le n \le (N + P - 1)$, e o método da covariância, em que é feito um janelamento do erro, calculado no intervalo $0 \le n \le (N - 1)$, onde N é o comprimento da janela e P, a ordem do preditor. O método da autocorrelação é mais eficiente em termos computacionais do que o método da covariância, pois o último requer um número maior de multiplicações para a resolução das equações matriciais. Além disso, no método da covariância, a estabilidade do filtro preditor não é garantida [26].

A ordem do preditor P necessária para representar de forma adequada um segmento de fala é determinada pelo número de formantes presentes no sinal. Como regra prática, considera-se que a cada 1 kHz existe um formante em sinais vozeados, e esse formante é representado por um par de pólos complexos conjugados. Para modelar os efeitos do fluxo do volume glotal e da irradiação sonora a partir da boca, precisa-se de mais 2 pólos [51]. Como exemplo, para um sinal amostrado a 10 kHz, deve-se empregar 10 pólos para modelar o trato e 2 pólos para os efeitos do pulso glotal e da irradiação. A ordem resultante do filtro preditor é 12. Lembra-se que essa é uma estimativa e deve ser adaptada às características da voz do locutor.

5.3.2 Aplicação em Síntese de Fala

A parametrização efetuada pela análise LPC, além de encontrar importantes aplicações em codificação de fala, permite a modificação dos parâmetros prosódicos em sistemas de síntese de fala. No modelo LPC, pode-se controlar, de forma independente, o envelope espectral, o *pitch*, o ganho e as durações dos segmentos de fala. O *pitch* é alterado variando-se o intervalo entre os pulsos da excitação. A intensidade do sinal pode ser modificada através do parâmetro ganho, com certo cuidado. Em segmentos vozeados, se a análise LPC tiver sido feita de forma síncrona com o *pitch*, a duração pode ser variada através da repetição ou exclusão de quadros LPC. Uma alternativa para evitar a simples repetição é a criação de novos períodos pela interpolação de períodos vizinhos. Além das

alterações prosódicas, a análise LPC pode ser usada, por exemplo, para a criação de novas vozes em um sistema de síntese concatenativa. Modificações de *pitch* e das posições dos formantes permitem que uma voz feminina seja obtida a partir de uma voz masculina original [51].

Em síntese concatenativa, a etapa de análise da técnica LPC normalmente é efetuada *off-line*, durante o processo de criação do banco de unidades. As unidades são armazenadas em uma representação que emprega os parâmetros LPC. A vantagem dessa representação parametrizada é a redução da quantidade de memória necessária para armazenar as unidades. As etapas de modificação e síntese são efetuadas, por razões evidentes, em tempo real, no ambiente de aplicação do sistema.

O grande problema da técnica LPC padrão está no modelo da excitação vozeada, na forma de impulsos espaçados pelo período de *pitch*. O trato vocal é excitado em uma única amostra por período fundamental, o que corresponde uma simplificação muito extrema dada a variabilidade das ondas glotais na fala natural. Na Fig. 5.8, são mostrados dois sinais que ilustram o resultado da simplificação empregada na técnica padrão. O primeiro sinal é o original, da frase “souberam que algo aconteceu” dita por um locutor masculino. O segundo é o sintetizado usando um preditor linear de ordem 10, com quadros de análise síncronos com o período de *pitch* e empregando o método da autocorrelação. A taxa de amostragem é de 8 kHz. Não foi realizado qualquer tipo de alteração prosódica. Na Fig. 5.9, o intervalo de tempo de 0,4 a 0,6 s é ampliado, sendo possível perceber o efeito da excitação com um único pulso por período de *pitch*. Como a memória do filtro preditor é muito menor do que o tamanho do período de *pitch*, o sinal sintetizado torna-se zero em grande parte do período. Com isso, perceptualmente avalia-se o sinal como “metálico” [51], mesmo para o caso de não se realizar alteração prosódica de qualquer tipo.

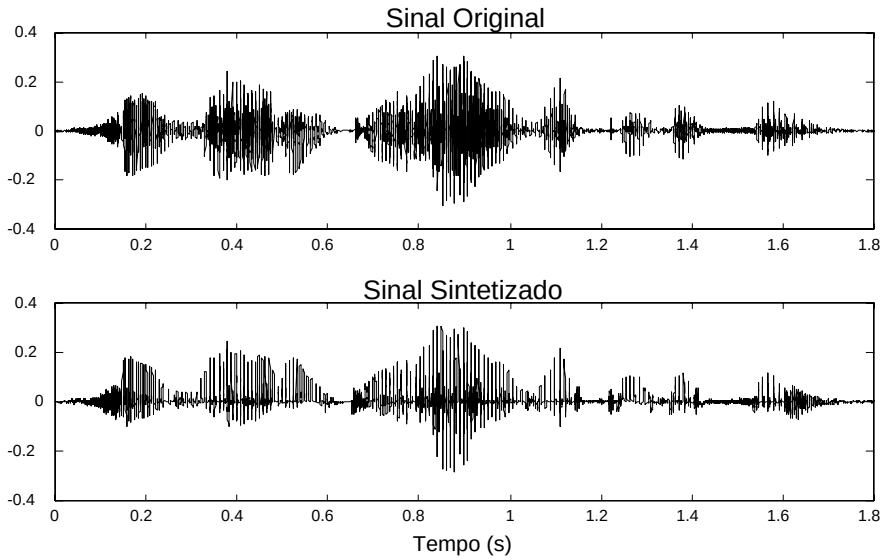

Fig. 5.8: Sinal original e sintetizado usando a técnica LPC padrão.

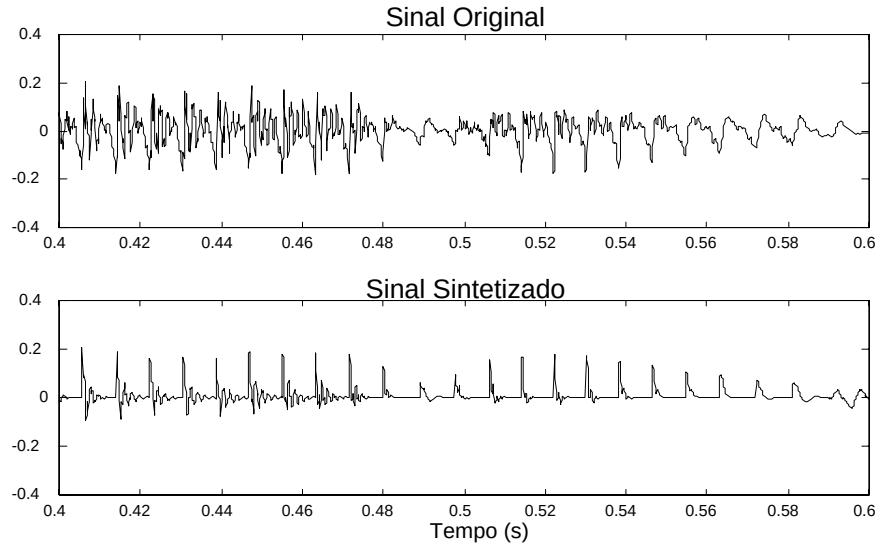

Fig. 5.9: Detalhe mostrando o efeito da excitação simplificada.

5.3.3 Modelo com Excitação Residual

Como visto anteriormente, se o sinal de resíduo $e(n)$ for utilizado como excitação do filtro $H(z)/G$, o sinal sintetizado $s_0(n)$ será igual ao sinal original $s(n)$. Uma variação da técnica LPC padrão emprega o sinal de resíduo $e(n)$ como excitação. Aplicada à codificação de fala, essa técnica é conhecida com *Residual-Excited Linear Predictive Coding* (RELP). Apesar da taxa de bits requerida ser maior do que na técnica LPC padrão, a qualidade obtida é muito superior.

Em síntese de fala, além de permitir a geração de um sinal sintetizado sem degradações, a utilização desse esquema deve prover meios de alteração dos parâmetros prosódicos. Em vista disso, várias alternativas têm sido propostas na literatura [4,5].

Uma técnica chamada *Linear-Predictive Pitch-Synchronous Overlap-and-Add* (LP-PSOLA) aplica o método PSOLA sobre o sinal de resíduo [4]. Esse esquema permite a suavização entre quadros pela interpolação dos parâmetros LPC, além da compressão do banco de unidades. Entretanto, a complexidade computacional é maior do que a da técnica PSOLA padrão.

Em [8,52] é apresentada outra forma de alteração do sinal de resíduo. Nesse método, que emprega múltiplas janelas, o quadro de análise LPC, síncrono com o período de *pitch*, é dividido em subquadros. Esses subquadros são deslocados por pequenos fatores, alterando o comprimento do sinal de resíduo e, conseqüentemente, modificando o *pitch*. O processo básico, para o caso de aumento de *pitch*, é ilustrado na Fig. 5.10. Com o objetivo de avaliar diferentes métodos, foram integradas ao sistema *Laureate*, da *British Telecom*, as seguintes técnicas: TD-PSOLA, LP-PSOLA e o método de múltiplas janelas [8]. Testes informais mostraram uma preferência pelas duas técnicas baseadas no sinal de resíduo, com os ouvintes classificando a fala sintetizada como de “maior brilho” [8].

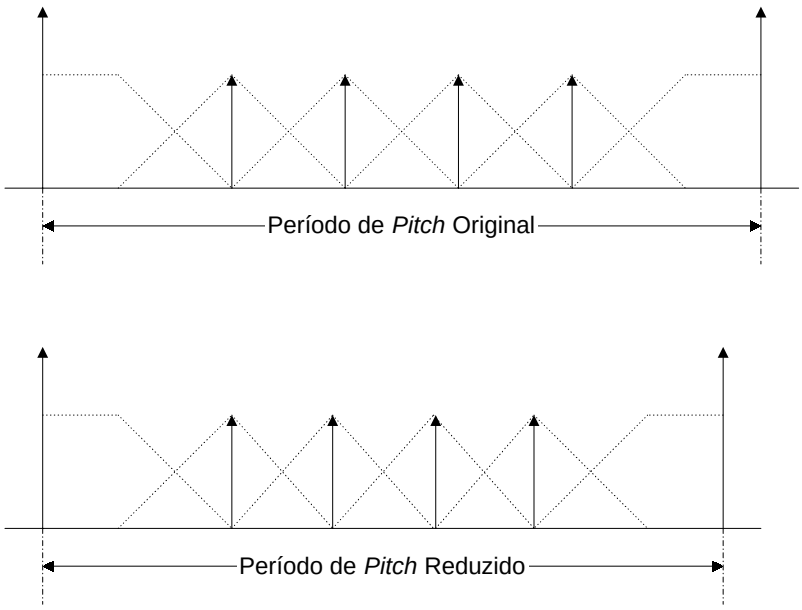

Fig. 5.10: Modificação de *pitch* através da divisão de um período em subquadros.

Em [7], é apresentado um método híbrido, que emprega PSOLA para a alteração de duração em segmentos não-vozeados e LPC baseado em excitação residual para a mudança de *pitch* em segmentos vozeados. Na análise LPC, o tamanho do quadro utilizado inclui um pequeno intervalo de sobreposição, da ordem de 2 ms, e um período de *pitch* completo. A modificação do comprimento do sinal de resíduo (e conseqüentemente do *pitch*) é executada através da aplicação de uma janela, para o aumento de *pitch*, ou da inserção de zeros, para a redução. A janela aplicada é da forma:

$$h(n) = \begin{cases} 1,0 & , 0 < n < 0,75T \\ \frac{1}{2}\left(1+\cos\left(\frac{\pi\left(i-0,75\mathrm{T}\right)}{0,25\mathrm{T}}\right)\right) & , 0,75T \le n < T \end{cases} \tag{5.37}$$

onde T é o comprimento do período de *pitch* desejado ou do período de pitch original, o que for menor.

Após essa modificação, uma interpolação entre quadros do sinal de resíduo e dos coeficientes do filtro é aplicada, com o objetivo de reduzir descontinuidades nos limites dos quadros.

Para avaliar esse método de modificação de *pitch*, ainda em [7], um sintetizador de fala foi implementado, dispondo de duas estratégias de síntese: com a técnica PSOLA padrão e com esse método híbrido. Em testes informais, o método híbrido forneceu a impressão subjetiva de ter menos distorções, especialmente quando aplicado a vozes com *pitch* alto. Esse resultado motivou a montagem de um procedimento de avaliação perceptual que medisse a inteligibilidade formalmente [7], em que foram comparadas três variáveis:

a) método: PSOLA padrão ou híbrido;
b) direção de mudança de *pitch*: incremento ou decremento;
c) locutor: adulto ou criança (ambos do sexo feminino).

Quarenta e nove palavras foram gravadas pelos locutores e os fatores de alteração de *pitch* foram 1,6 e 0,7. Os estímulos foram apresentados a 20 ouvintes, que deveriam soletrar o que foi escutado. Os resultados mostraram que:

a) em relação ao método e à direção, PSOLA foi melhor para a redução de *pitch* e o método híbrido, melhor para aumento de *pitch*;
b) em relação ao método e ao locutor, PSOLA foi melhor para a voz do adulto e o método híbrido, para a voz da criança.

Notou-se também que a técnica PSOLA teve o melhor e o pior resultado na experiência. Essa variação parece explicar a pior avaliação subjetiva da técnica PSOLA em relação ao método híbrido. Ao estimar a qualidade de uma técnica, os ouvintes tendem a dar um maior peso às falhas percebidas do que aos sucessos (que não são percebidos) [7].

Ainda em [7], sugere-se que o processo de decremento de *pitch* do método híbrido seja aperfeiçoado, de modo que não se utilize apenas a simples inserção de zeros.

5.4 Síntese Híbrida: Componentes Harmônico e Estocástico

Os problemas associados às técnicas PSOLA e LPC têm motivado o desenvolvimento de novos métodos de alteração de parâmetros prosódicos. Uma técnica que tem demonstrado grande potencial é a de síntese híbrida [9-11], em que o sinal de fala $s(n)$ é decomposto em um componente harmônico $s_h(n)$ e um componente de ruído $r(n)$. Assim:

$$s(n) = s_h(n) + r(n) \tag{5.38}$$

$$s_h(n) = \sum_{k=0}^{K} A_k \cos(k\omega_0 n + \theta_k) \tag{5.39}$$

onde K é o número de harmônicas, ω_0 é a freqüência fundamental, e A_k e θ_k são a amplitude e a fase da k-ésima harmônica. Nessa decomposição, assume-se que os componentes harmônico e de ruído ocupam diferentes bandas de freqüência [9]. O componente harmônico ocupa a banda de 0 a $K\omega_0$ rad, enquanto o componente de ruído ocupa de $K\omega_0$ a π rad. A freqüência máxima $K\omega_0$ é variável para permitir uma correta separação entre os dois componentes, evitando periodicidades no componente de ruído e ruído no componente harmônico. Em segmentos não-vozeados, o componente harmônico é zero. Nos segmentos vozeados, os dois componentes são levadas em consideração, para modelar por exemplo, as fricativas vozeadas, as transições vozeado/não-vozeado ou qualquer outro tipo de ruído presente no sinal de fala.

Em relação à técnica TD-PSOLA, a síntese híbrida apresenta algumas vantagens [9,10]:

a) permite uma suavização entre as unidades de forma simples e flexível;

b) evita os artefatos “metálicos” produzidos com TD-PSOLA;

c) suporta modificações de *pitch* e duração maiores;

d) permite modificações contínuas de *pitch* e duração;

e) mudanças de *pitch* podem ser efetuadas sem compensação de duração;

f) o controle de *pitch* e duração pode ser feito sobre cada período de *pitch*, de forma mais flexível do que na técnica TD-PSOLA, em que normalmente apenas um parâmetro de duração e dois ou três de *pitch* estão disponíveis por fone;

g) permite a compressão do banco em taxas de até 6 kb/s.

Entretanto, dois fortes inconvenientes estão associados a este modelo [9,10]:

a) é extremamente sensível a erros de classificação vozeado/não-vozeado;

b) tem um elevado custo computacional.

As técnicas apresentadas nesse capítulo têm diferentes graus de flexibilidade, qualidade e complexidade. Pretendendo buscar um compromisso entre esses três fatores, uma nova proposta será discutida no capítulo seguinte.

Capítulo 6

Método Proposto

6.1 Motivação

A técnica LPC, aplicada à síntese de fala, tem a vantagem de, com a separação fonte-filtro, permitir um fácil controle dos parâmetros prosódicos. O problema com o modelo de excitação simplificado da técnica padrão pode ser contornado com o uso do sinal residual. A técnica LPC aplicada a sistemas TTS envolve três etapas: análise, modificação e síntese. A etapa de análise tem um certo custo computacional, mas, em sistemas de conversão texto-fala, pode ser realizada *off-line*. As duas etapas seguintes são, obrigatoriamente, realizadas durante a execução do aplicativo. Enquanto a síntese envolve apenas um procedimento de filtragem digital, na etapa de modificação pode haver um custo computacional elevado. Técnicas de diferentes graus de complexidade são descritas na literatura [4,5,7,8].

Um sistema de conversão texto-fala para a língua portuguesa falada no Brasil encontra-se em desenvolvimento no Laboratório de Instrumentação Eletrônica: Circuitos e Processamento de Sinais (LINSE/UFSC). Nesse sistema, que emprega síntese concatenativa, a etapa de modificação prosódica faz-se fundamental e deve cumprir algumas exigências. O principal requisito que deve ser cumprido no desenvolvimento da técnica de alteração prosódica está relacionado à complexidade computacional. Para possibilitar a implementação do maior número possível de módulos de síntese em tempo real, a complexidade deve ser extremamente reduzida. Além da complexidade, o requisito de qualidade não pode ser deixado de lado. Avaliações perceptuais [7,8] mostram que a

técnica LPC com excitação residual tem dado bons resultados, superiores aos da técnica PSOLA. Tomando por base um método desenvolvido em [7], utilizando LPC, propõe-se neste capítulo uma técnica de modificação dos parâmetros prosódicos *pitch* e duração baseada em análise/ressíntese LPC com excitação residual.

6.2 Teoria de Operação

A técnica desenvolvida é utilizada em sistemas de síntese concatenativa, em que a carga computacional seja um fator limitante. Assume-se que os segmentos que compõem o banco de unidades possuem marcas de *pitch* previamente assinaladas. As marcas são feitas no cruzamento por zero anterior ao maior pico de um período de *pitch*, o que corresponde aproximadamente ao fechamento da glote.

Seja $x(n)$ o sinal de fala original de uma dada unidade do banco de unidades. A esse sinal estão associadas marcas de *pitch* $P(k)$, onde $k = 1, 2, ..., M$ define o número da marca de *pitch*. Um intervalo fixo de sobreposição é definido, com um número de amostras S. Um quadro de análise $x_k(n)$ é obtido das amostras $(P(k) - S)$ até $(P(k+1) - 1)$ de $x(n)$. Assim, um quadro de análise relacionado ao período de *pitch* k engloba um intervalo de sobreposição que inclui S amostras do período $(k-1)$ e mais todo o período de *pitch* definido por $t_k = (P(k+1) - 1) - P(k))$. Se para $k = 1$ não existirem S amostras anteriores a $P(1)$, nesse caso são inseridos zeros. O sinal $x_k(n)$ é passado por uma janela da forma:

$$h_s(n) = \begin{cases} \frac{1}{2}\left(1 - \cos\left(\pi \frac{n}{S+1}\right)\right) & , 1 \le \mathrm{n} < \mathrm{S} \\ 1,0 & , \mathrm{n} \ge \mathrm{S} \end{cases} \tag{6.1}$$

gerando o sinal $x'_k(n)$. Uma análise LPC é aplicada sobre o sinal $x'_k(n)$ (que corresponde a um período inteiro de *pitch*) e são obtidos os coeficientes LPC. Uma filtragem inversa é, então, realizada e os sinais estimado $\hat{x}_k(n)$ e de resíduo $e_k(n)$ são determinados. Sobre o

sinal $e_k(n)$ é aplicado um procedimento que depende da direção de mudança do *pitch*. Seja T o comprimento do sinal de resíduo desejado:

a) para um incremento de *pitch*, o sinal $e_k(n)$ é passado por uma janela

$$h_i(n) = \begin{cases} 1,0 & , 0 < n < 0,75T \\ \frac{1}{2}\left(1 + \cos\left(\frac{\pi\left(i - 0,75\mathrm{T}\right)}{0,25\mathrm{T}}\right)\right) & , 0,75T \le n < T \end{cases} \quad (6.2);$$

b) para um decremento de *pitch*, são copiadas as últimas $(T - t_k)$ amostras do sinal $e_k(n)$, gerando o sinal $\tilde{e}_k(n)$.

O sinal de resíduo modificado $\tilde{e}_k(n)$ é passado pelo filtro LPC, sendo gerado o sinal com *pitch* modificado $\tilde{x}_k(n)$.

Por fim, cada um dos sinais $\tilde{x}_k(n)$ é combinado de modo que o intervalo fixo de sobreposição seja considerado.

Alterações de duração são obtidas de forma simples pela cópia ou exclusão dos quadros de análise.

6.3 Aplicação e Resultados

Seja o sinal original $x(n)$ mostrado na Fig. 6.1, amostrado a f_s=10 kHz e com marcas de *pitch* associadas. O intervalo de sobreposição utilizado é de 1 ms, e o número de amostras de sobreposição $S = 0,001 / f_s = 10$.

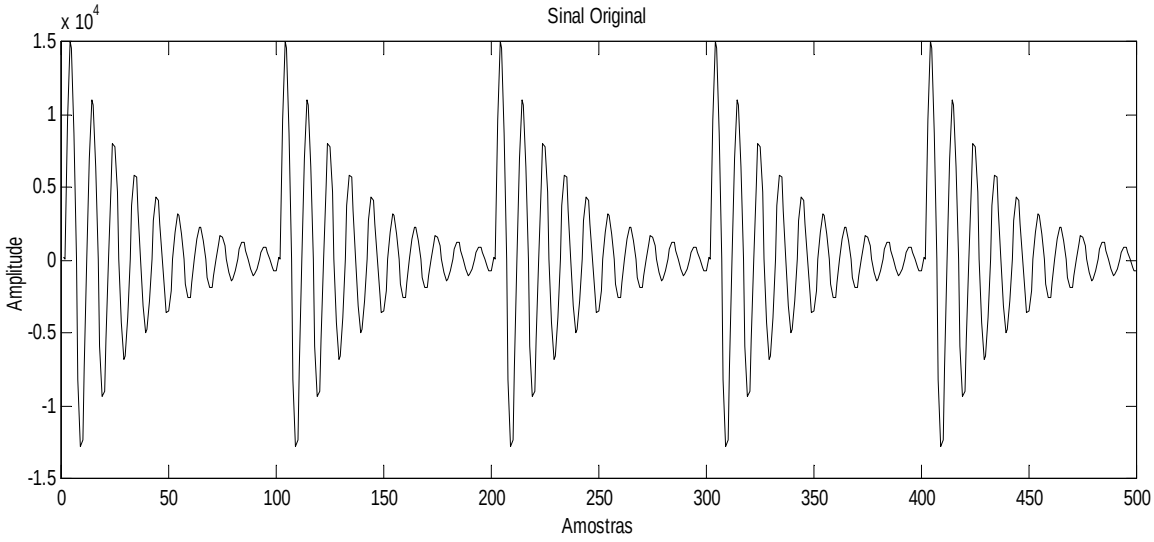

Fig. 6.1: Sinal original.

Na Fig. 6.2 é mostrado um quadro de análise $x_k(n)$ do sinal original, juntamente com o sinal "janelado" $x'_k(n)$. Na Fig. 6.3, mostra-se uma ampliação da região de sobreposição do sinal, mostrando o efeito da aplicação da janela.

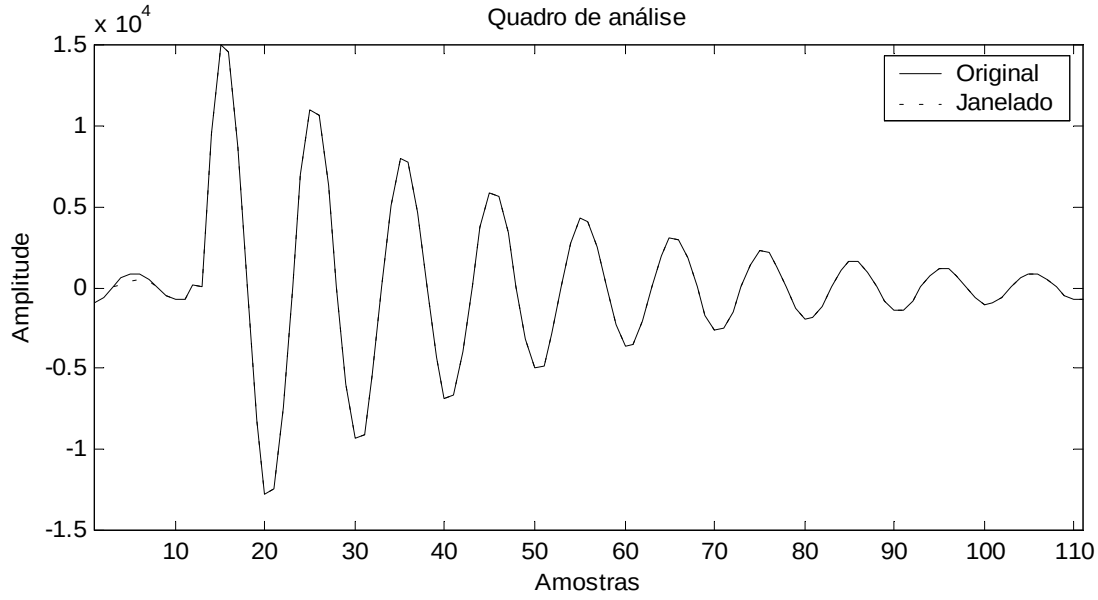

Fig. 6.2: Quadro de análise com o sinal original e janelado.

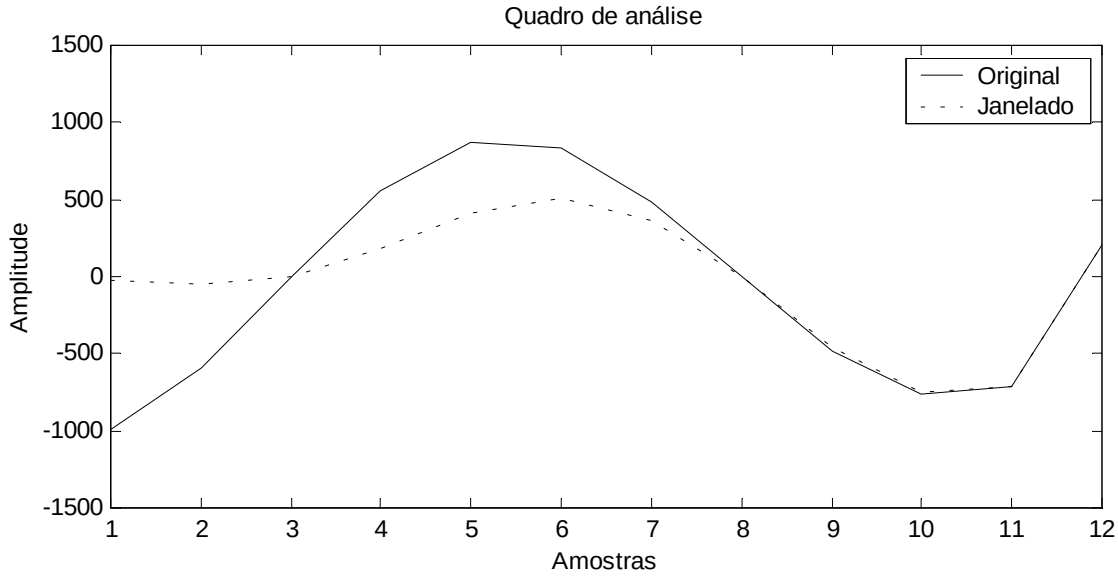

Fig. 6.3: Vista ampliada do quadro de análise.

Uma análise LPC de ordem 12, usando o método da autocorrelação, é realizada sobre o sinal janelado $x'_k(n)$ e são obtidos o sinal estimado $\hat{x}_k(n)$ e o resíduo $e_k(n)$, mostrados na Fig. 6.4. Na Fig. 6.5, o sinal de resíduo é visto de forma isolada.

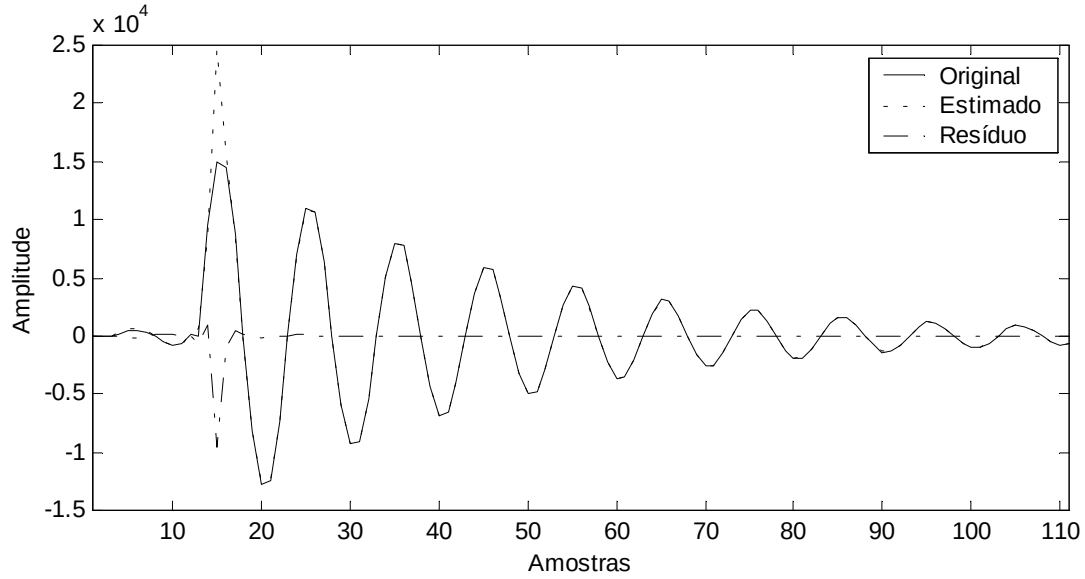

Fig. 6.4: Sinais original, estimado e resíduo para um quadro de análise.

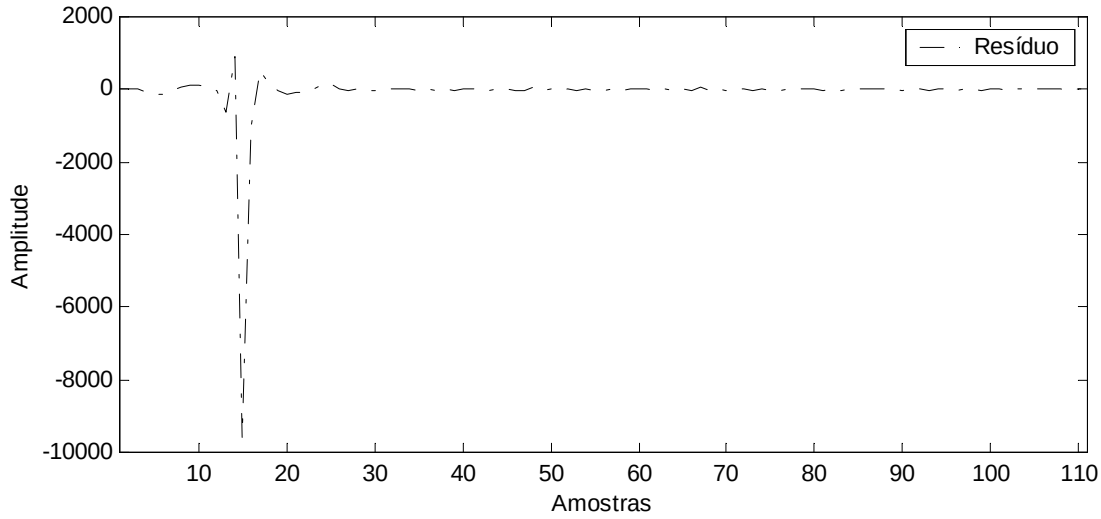

Fig. 6.5: Sinal residual para um quadro de análise.

Nesse quadro, o *pitch* é de 100 Hz. Deseja-se incrementá-lo para 120 Hz. Para tanto, aplica-se a equação (6.2), com $T = \frac{120}{f_s} + S = 83 + 10 = 93$. Obtém-se o sinal $\tilde{e}_k(n)$ mostrado na Fig. 6.6.

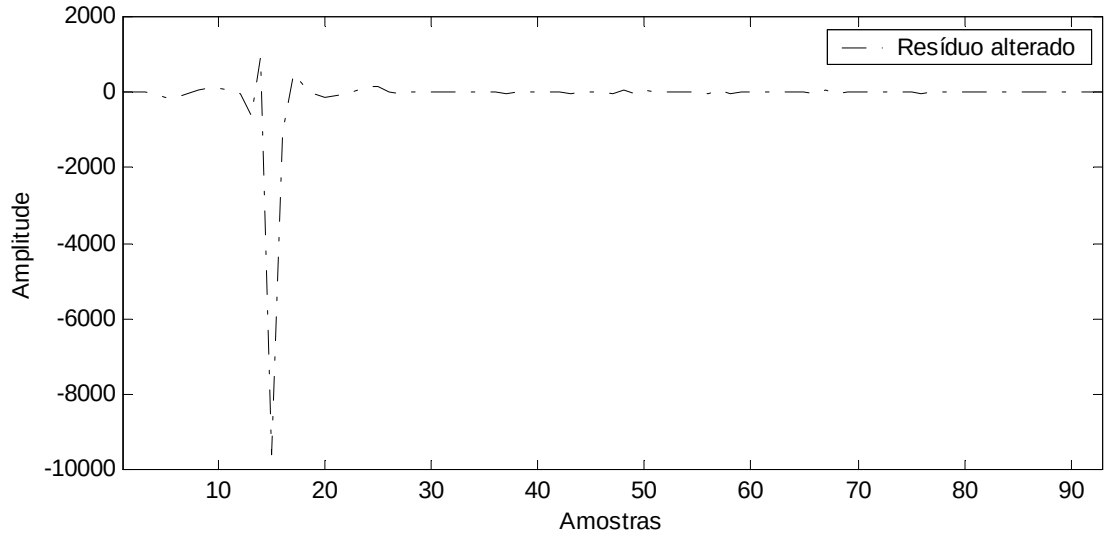

Fig. 6.6: Sinal residual modificado.

Procede-se à ressíntese com o sinal de resíduo modificado. O quadro sintetizado $\tilde{x}_k(n)$, com a modificação de *pitch*, mostrado na Fig. 6.7, é gerado.

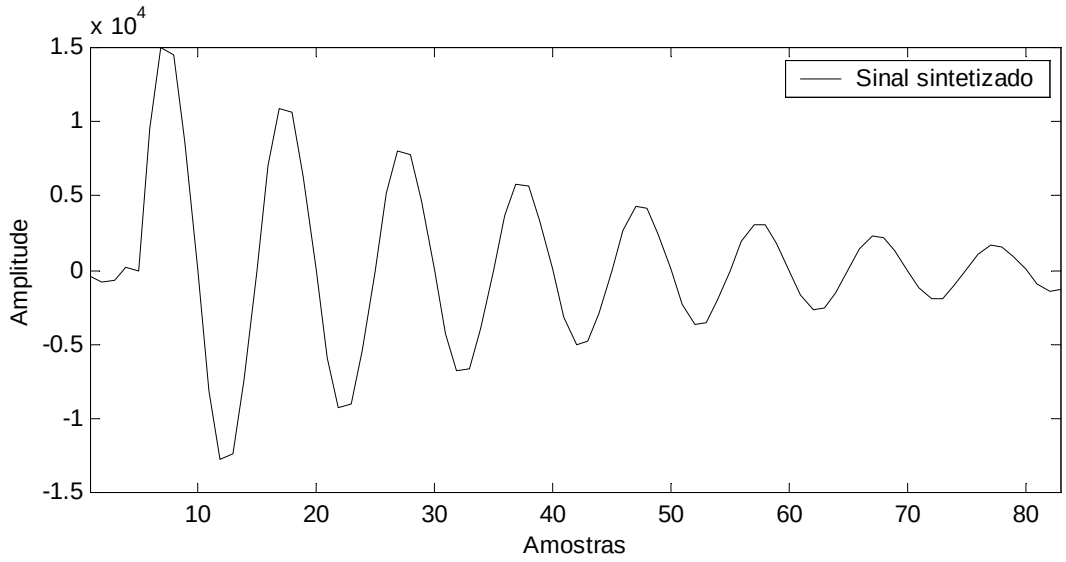

Fig. 6.7: Sinal modificado $\tilde{x}_k(n)$ sintetizado.

Para demonstrar a aplicação da técnica na alteração da entonação de uma frase, considera-se que a frase afirmativa “é suficiente” deva ser transformada na interrogativa “é suficiente?”. Nesse tipo de frase, um contorno interrogativo básico pode ser imposto se o *pitch* da penúltima sílaba for aumentado, seguindo um padrão baixo-alto. Na Fig. 6.8, são mostrados o segmento [iẽ] da palavra “suficiente” e o contorno de *pitch* associado a esse segmento. Percebe-se que o contorno segue um padrão alto-baixo, iniciando em 105 Hz e terminando em 75 Hz. Para dar um efeito interrogativo, o contorno de *pitch* desse segmento deveria iniciar aproximadamente em 90 Hz e terminar em 140 Hz. Na Fig. 6.9, são mostrados esse contorno de *pitch* (imposto) e o correspondente sinal de fala sintetizado através da técnica apresentada[10]. Com essa mudança de *pitch*, a frase que era afirmativa é percebida como interrogativa. Espectrogramas dos segmentos original e sintetizado são

[10] Exemplos de síntese empregando a técnica desenvolvida podem ser ouvidos no CD anexo ou obtidos no endereço eletrônico do LINSE (www.linse.ufsc.br).

mostrados na Fig. 6.10. Nota-se que não há deslocamento em freqüência dos formantes, ou seja, o timbre da voz não é alterado.

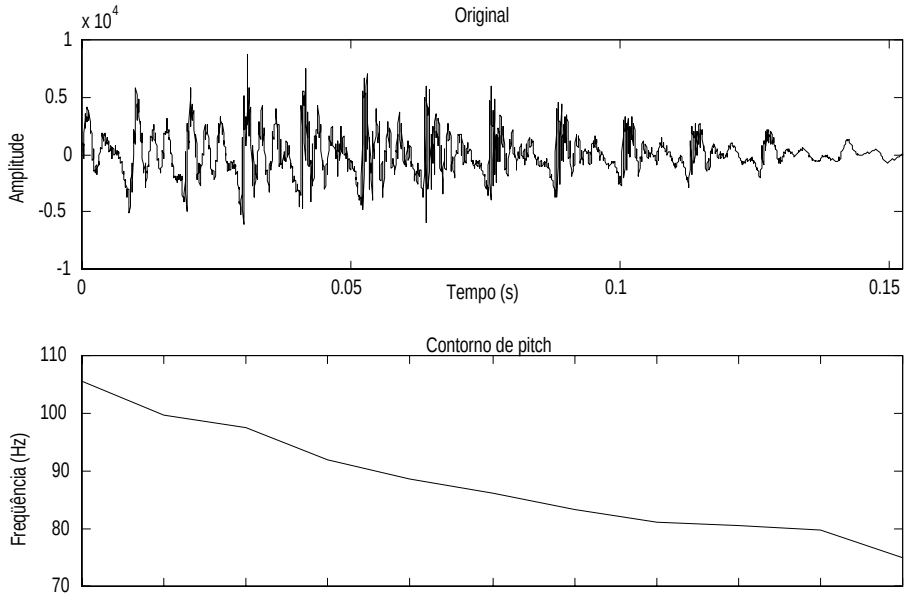

Fig. 6.8: Sinal original e contorno de *pitch*.

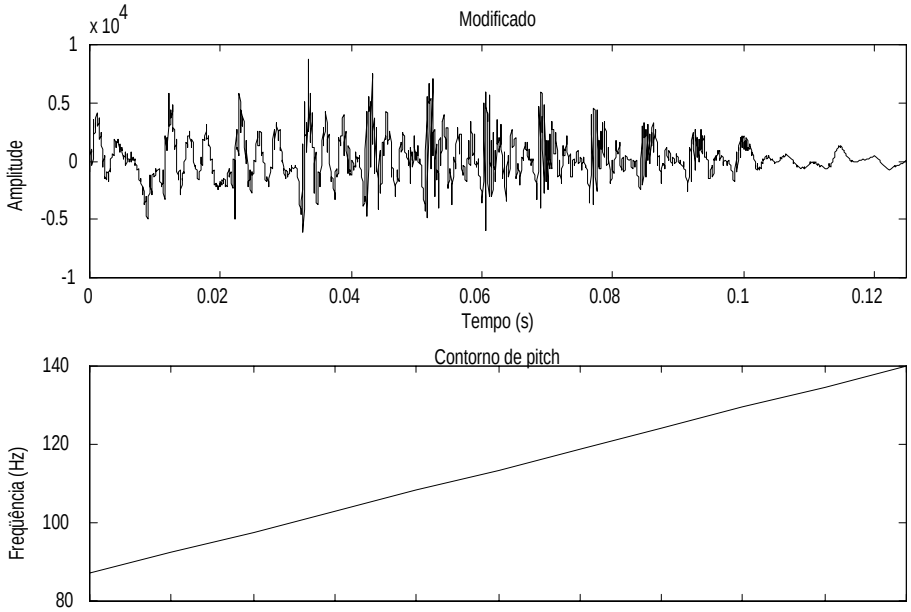

Fig. 6.9: Sinal sintetizado e contorno de *pitch* imposto.

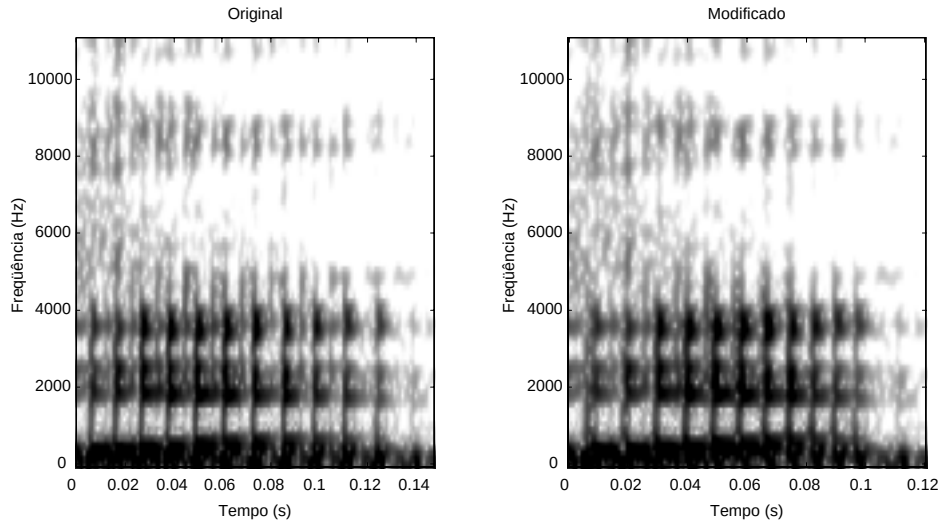

Fig. 6.10: Espectrogramas dos sinais original e sintetizado.

A grande vantagem da técnica desenvolvida é permitir que as alterações prosódicas sejam realizadas de forma extremamente simples, possibilitando, inclusive, a implementação de múltiplos módulos de síntese. Assim, o requisito de complexidade computacional é atendido. Além disso, o timbre da voz original é preservado, sem que haja deslocamento indesejado de formantes. Com essa técnica, também é possível realizar procedimentos de suavização espectral entre as unidades. Uma forma bastante simples de realizar essa suavização é através da interpolação dos coeficientes do filtro preditor entre os quadros de síntese. Testes efetuados com esse procedimento mostraram que há uma melhoria na qualidade subjetiva da fala, percebida como de maior “fluidez”.

Em testes formais, a técnica desenvolvida em [7] não apresentou desempenho muito bom na operação de redução de *pitch*, realizada pela inserção de zeros no sinal de resíduo. Mesmo assim, a avaliação subjetiva geral indicava que o resultado era melhor do que com PSOLA, que havia tido o melhor e o pior resultado. Na técnica desenvolvida aqui, o problema relacionado à redução de *pitch* foi contornado pela simples cópia das últimas amostras do sinal residual. Testes preliminares mostraram que mesmo para variações de *pitch* relativamente elevadas (da ordem de 40%), não são percebidas degradações no sinal sintetizado.

Capítulo 7

Conclusões

Sistemas de conversão texto-fala (TTS) possibilitam que qualquer texto seja transformado em fala. Esse processo de mapeamento da representação escrita para a falada só pode ser resolvido de forma multidisciplinar, envolvendo basicamente técnicas de processamento lingüístico e de processamento de sinais. Inicialmente o texto é submetido a uma análise e são derivados parâmetros fonéticos e prosódicos (*pitch* e duração, em especial). A síntese propriamente dita do sinal de fala é efetuada através de técnicas de processamento de sinais, as quais devem gerar os sons com os parâmetros prescritos anteriormente.

Os sistemas TTS possibilitam uma grande gama de aplicações, do auxílio a portadores de deficiências a serviços em telecomunicações. Entretanto, o número de aplicações práticas só pode crescer se a qualidade final também avançar. Em relação à etapa de processamento de sinais, uma das abordagens que tem obtido melhores resultados é a de síntese concatenativa. Nessa abordagem, um conjunto de segmentos de fala é previamente gravado e armazenado. A síntese é realizada pela concatenação desses segmentos. Contudo, para que os requisitos de inteligibilidade e, principalmente, de naturalidade sejam atendidos, torna-se necessário modificar os parâmetros da fala associados à prosódia. As alterações dos parâmetros prosódicos são efetuadas através de técnicas de processamento de sinais.

Neste trabalho, diferentes métodos de modificação dos parâmetros prosódicos *pitch* e duração foram discutidos e uma técnica baseada em análise LPC com excitação residual foi desenvolvida. Pode-se dizer que três requisitos básicos devem ser atendidos

pelas técnicas de alteração prosódica: qualidade, flexibilidade e custo computacional. A qualidade está ligada à inteligibilidade e naturalidade da fala sintetizada. A flexibilidade está relacionada à aplicação da técnica para diferentes vozes e à capacidade de efetivar as variações prosódicas definidas pelo modelo lingüístico. Manter um reduzido custo computacional também é indispensável para que a aplicação prática (operação em tempo real) seja possível. A definição de um compromisso entre os três fatores torna-se essencial.

A técnica de alteração de parâmetros prosódicos desenvolvida neste trabalho, além de ter uma reduzida carga computacional, permite que variações prosódicas relativamente elevadas sejam alcançadas. A base da técnica, formada por operações de cópia e corte, é extremamente simples, podendo ser implementada de forma fácil em qualquer ambiente. O procedimento mais custoso computacionalmente, de análise e determinação dos parâmetros, pode ser realizado em uma etapa prévia, de modo que o desempenho do sistema não seja afetado. Em relação à qualidade, testes preliminares mostraram que bons resultados têm sido alcançados, sem a presença de ruídos ou artefatos no sinal sintetizado. Além disso, o timbre da voz não é afetado pelas mudanças de *pitch*.

Ao optar por uma técnica baseada em análise LPC, outras possibilidades ficam abertas. Uma é a utilização de procedimentos de suavização entre segmentos sucessivos, para amenizar os efeitos de eventuais descasamentos espectrais. Outra é a alteração controlada do envelope espectral e do *pitch* para que novas vozes sejam criadas sem que seja preciso regravar todo o banco de unidades.

Uma sugestão para trabalhos futuros reside na exploração da características do envelope espectral para a suavização entre diferentes unidades e na conversão de vozes. Outra sugestão é o estudo de estratégias para equalização de amplitudes das unidades. Uma etapa importante da síntese concatenativa é a gravação do *corpus*. Normalmente a gravação não é finalizada em em um único dia. Assim, características como o timbre e a intensidade da voz acabam sofrendo uma certa variação. As variações de amplitude entre unidades acabam prejudicando a qualidade da fala sintetizada. Procedimentos de

equalização, efetuados *off-line* ou durante a operação do sistema, seriam particularmente interessantes. Estudos para a compressão do banco também são uma sugestão para trabalhos futuros. Quando o número de unidades é elevado e há limitações de memória do sistema, a compressão do banco faz-se necessária. Outra sugestão é o desenvolvimento de técnicas mais elaboradas para a alteração de duração, especialmente em segmentos não-vozeados. Para grandes variações na escala de duração, essas técnicas poderiam levar a um aumento de qualidade. Técnicas de modificações prosódicas de custo computacional mais elevado, porém de maior qualidade têm sido propostas na literatura. Como sugestão final, incentiva-se o estudo e implementação dessas técnicas, já que a capacidade de processamento computacional avança muito rapidamente.

Referências Bibliográficas

[1] EGASHIRA, F. **Síntese de voz a partir de texto para a língua portuguesa**. Campinas, 1992. 113 f. Dissertação (Mestrado em Engenharia Elétrica) – Faculdade de Engenharia Elétrica, Universidade Estadual de Campinas.

[2] RABINER, L. R. Applications of Voice Processing to Telecommunications. **Proceedings of the IEEE**, v. 82, n. 2, p. 199-228, Feb. 1994.

[3] GOMES, L. de C. T. **Sistema de conversão texto-fala para a língua portuguesa utilizando a abordagem de síntese por regras**. Campinas, 1998. 107 f. Dissertação (Mestrado em Engenharia Elétrica) – Faculdade de Engenharia Elétrica e de Computação, Universidade Estadual de Campinas.

[4] MOULINES, E.; VERHELST, W. Time-Domain and Frequency-Domain Techniques for Prosodic Modification of Speech. In: KLEIJN, W. B.; PALIWAL, K. K. (Ed.). **Speech Coding and Synthesis**. Amsterdam: Elsevier, 1995. p. 519-555.

[5] MOULINES, E.; LAROCHE, J. Non-parametric techniques for pitch-scale and time-scale modification of speech. **Speech Communication**, v. 16, p. 175-205, 1995.

[6] CHARPENTIER, F.; MOULINES, E. Pitch-Synchronous Waveform Processing Techniques for Text-to-Speech Synthesis Using Diphones. In: EUROSPEECH, 1989, Paris. **Proceedings**… v. II, p. 13-19.

[7] BUNNELL, H. T.; YARRINGTON, D.; BARNER, K. E. Pitch Control in Diphone Synthesis. In: ESCA/IEEE WORKSHOP ON SPEECH SYNTHESIS, 2., 1994, New Paltz. **Proceedings**... Disponível em <http://www.asel.udel.edu/speech/reports/mohonk/main.ps> Acesso em 07 fev. 2000.

[8] EDGINTON, M.; LOWRY, A. Residual-Based Speech Modification Algorithms for Text-to-Speech Synthesis. In: INTERNATIONAL CONFERENCE ON SPOKEN LANGUAGE PROCESSING, 4., 1996, Philadelphia. **Proceedings**… v. 3, p. 1425-1428. Disponível em <http://www.asel.udel.edu/icslp/cdrom/vol3/078/a078.pdf> Acesso em 05 mai. 2000.

[9] VIOLARO, F.; BOËFFARD, O. A Hybrid Model for Text-to-Speech Synthesis. **IEEE Transactions on Speech and Audio Processing**, v. 6, n. 5, p. 426-434, Sept. 1998.

[10] SYRDAL, A. et al. TD-PSOLA versus Harmonic Plus Noise Model in Diphone Based Speech Synthesis. In: INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH AND SIGNAL PROCESSING, 1998, Seattle. **Proceedings**… v. 1, p. 273-276. 1 CD-ROM.

[11] DUTOIT, T. **An Introduction to Text-to-Speech Synthesis**. Dordrecht: Kluwer, 1997. (Text, Speech and Language Technology, v. 3).

[12] PAGE, J. H.; BREEN, A. P. The Laureate text-to-speech system:-architecture and applications. **BT Technology Journal**, v. 14, n. 1, p. 57-67, Jan. 1996.

[13] LEVINSON, S. E.; OLIVE, J. P.; TSCHIRGI, J. S. Speech Synthesis in Telecommunications. **IEEE Communications Magazine**, v. 31, n. 11, p. 46-53, Nov. 1993.

[14] COX, R. V. et al. Speech and Language Processing for Next-Millennium Communications Services. **Proceedings of the IEEE**, v. 88, n. 8, p. 1314-1337, Aug. 2000.

[15] LEMMETTY, S. **Review of Speech Synthesis Technology**. Espoo, 1999. 104 f. Master Thesis (Master of Science) – Department of Electrical and Communications Engineering, Helsinki University of Technology.

[16] RUBIN, P.; VATIKIOTIS-BATESON, E. **Talking Heads**. Disponível em <http://www.haskins.yale.edu/Haskins/HEADS/contents.html> Acesso em 15 fev. 2001.

[17] HUANG, X.; ACERO, A.; HON, H. **Spoken Language Processing**: A Guide to Theory, Algorithm, and System Development. Upper Saddle River: Prentice Hall, 2001.

[18] KLATT, D. H. Review of text-to-speech conversion for English. **Journal of the Acoustical Society of America**, v. 82, n. 3, p. 737-793, Sept. 1987.

[19] VOCAL Vowels: Exploratorium Exhibit. Disponível em <http://www.exploratorium.edu/exhibits/vocal_vowels/vocal_vowels.html> Acesso em 15 fev. 2001.

[20] ONDREJOVIC, S. **Wolfgang Von Kempelen and his "Mechanism of Human Speech"**. Disponível em <http://www.slovakradio.sk/kultura/expstudio/kempe.html> Acesso em 10 mar. 2000.

[21] NOW a Machine That Talks With the Voice of Man. **Science News Letter**, 14 Jan. 1939. p.19. Disponível em: <http://www.americanhistory.si.edu/scienceservice/newsletters/39019p.htm> Acesso em: 05 mar. 2000.

[22] TELECOMMUNICATIONS Industry Product Backgrounder. Disponível em: <ftp://ftp.lhsl.com/uk/pr/backgrounders/telecom_bg.pdf> Acesso em: 16 fev. 2001.

[23] SYRDAL, A. K. et al. **Corpus-Based Techniques in the AT&T NextGen Synthesis System**. In: INTERNATIONAL CONFERENCE ON SPOKEN LANGUAGE PROCESSING, 6., 2000, Beijing. Disponível em: <http://www.research.att.com/projects/tts/papers/2000_ICSLP/corpus.pdf> Acesso em 20 dez. 2000.

[24] SILVA, T. C. **Fonética e fonologia do português**: roteiro de estudos e guia de exercícios. São Paulo: Contexto, 1999.

[25] DUBOIS, J. et al. **Dicionário de Lingüística**. São Paulo: Cultrix, 1993.

[26] MARTINS, J. A. **Vocoder LPC com Quantização Vetorial**. Campinas, 1991. 118 f. Dissertação (Mestrado em Engenharia Elétrica) – Faculdade de Engenharia Elétrica, Universidade Estadual de Campinas.

[27] LAVER, J. **Principles of Phonetics**. Cambridge: Cambridge University Press, 1994. (Cambridge Textbooks in Linguistics).

[28] LADEFOGED, P. **Vowels and Consonants**: An Introduction to the Sounds of Language. Oxford: Blackwell, 2001.

[29] CALLOU, D.; LEITE, Y. **Iniciação à Fonética e à Fonologia**. 5. ed. Rio de Janeiro: Jorge Zahar, 1995.

[30] NEPOMUCENO, L. de A. **Elementos de Acústica Física e Psicoacústica**. São Paulo: Edgard Blücher, 1994.

[31] CRYSTAL. D. **The Cambridge Encyclopedia of Language**. 2. ed. Cambridge: Cambridge University Press, 1997.

[32] CABRAL, L. S. **Introdução à lingüística**. 7. ed. Rio de Janeiro: Globo, 1988.

[33] SPROAT, R.; OLIVE, J. An Approach to Text-to-Speech Synthesis. In: KLEIJN, W. B.; PALIWAL, K. K. (Ed.). **Speech Coding and Synthesis**. Amsterdam: Elsevier, 1995. p. 611-632.

[34] COSTA NETO, M. L. da. **Conversor Texto-Fala de Alta Qualidade para a Língua Portuguesa**. Campina Grande, 2000. 112 f. Exame de Qualificação

(Doutorado em Engenharia Elétrica) – Centro de Ciências e Tecnologia, Universidade Federal da Paraíba.

[35] FIGUEIREDO, F. A.; NAVINER, L. A. B.; AGUIAR NETO, B. G. Uma Nova Abordagem para o Sistema de Conversão Texto-Fala para a Língua Portuguesa. In: SIMPÓSIO BRASILEIRO DE TELECOMUNICAÇÕES, 15., 1997, Recife. **Anais**... p. 328-331.

[36] EDGINGTON, M. et al. Overview of current text-to-speech techniques: Part I - text and linguistic analysis. **BT Technology Journal**, v. 14, n. 1, p. 68-83, Jan. 1996.

[37] SILVA, C. H.; VIOLARO, F. Modelamento Prosódico para Conversão Texto-Fala do Português Falado no Brasil. **Revista Brasileira de Telecomunicações**, Campinas, v. 10, n. 1, p. 15-24, dez. 1995.

[38] SANTEN, J. P. H. van. Computation of Timing in Text-to-Speech Synthesis. In: KLEIJN, W. B.; PALIWAL, K. K. (Ed.). **Speech Coding and Synthesis**. Amsterdam: Elsevier, 1995. p. 663-684.

[39] GABIOUD, B. Articulatory Models in Speech Synthesis. In: KELLER, E. (Ed.). **Fundamentals of speech synthesis and speech recognition**: basic concepts, state of the art and future challenges. Chichester: J. Wiley, 1994. p. 215-230.

[40] PRADO, P. P. L. do. Sintetizador Articulatório de Voz: Mapeamento Acústico/Articulatório. In: SIMPÓSIO BRASILEIRO DE TELECOMUNICAÇÕES, 11., 1993, Natal. **Anais**... v. I, p. 708-712.

[41] BICKLEY, C. A.; STEVENS, K. N.; WILLIAMS, D. R. A Framework for Synthesis of Segments Based on Pseudoarticulatory Parameters. In: SANTEN, J. P. H. van et al. (Ed.). **Progress in Speech Synthesis**. New York: Springer-Verlag, 1996. p. 211-220.

[42] STYGER, T.; KELLER, E. Formant Synthesis. In: KELLER, E. (Ed.). **Fundamentals of speech synthesis and speech recognition**: basic concepts, state of the art and future challenges. Chichester: J. Wiley, 1994. p. 109-128.

[43] KLATT, D. H. Software for a cascade/parallel formant synthesizer. **Journal of the Acoustical Society of America**, v. 67, n. 3, p. 971-995, Mar. 1980.

[44] DONOVAN, R. E. **Trainable Speech Synthesis**. Cambridge, 1996. 156 f. Thesis (Doctor of Philosophy) –Engineering Department, Cambridge University.

[45] BHASKARARAO, P. Subphonemic Segment Inventories for Concatenative Speech Synthesis. In: KELLER, E. (Ed.). **Fundamentals of speech synthesis and speech recognition**: basic concepts, state of the art and future challenges. Chichester: J. Wiley, 1994. p. 69-84.

[46] PORTELE, T.; HÖFER, F.; HESS, W. J. A Mixed Inventory Structure for German Concatenative Synthesis. In: SANTEN, J. P. H. van et al. (Ed.). **Progress in Speech Synthesis**. New York: Springer-Verlag, 1996. p. 263-277.

[47] SHUKLA, S. R.; BARNWELL III, T. P. **Implementation of High Quality Text-to-Speech Synthesis Using Words and Diphones**. Disponível em: <http://users.ece.gatech.edu/~shukla/docs/icassp20.ps> Acesso em 10 fev. 2001.

[48] MAKASHAY, M. J. et al. **Perceptual Evaluation of Automatic Segmentation in Text-to-Speech Synthesis**. In: In: INTERNATIONAL CONFERENCE ON SPOKEN LANGUAGE PROCESSING, 6., 2000, Beijing. Disponível em: <http://www.research.att.com/projects/tts/papers/2000_ICSLP/autoseg.pdf> Acesso em 20 dez. 2000.

[49] KORTEKAAS, R. W. L.; KOHLRAUSCH, A. Psychoacoustical evaluation of the pitch-synchronous overlap-and-add speech-waveform manipulation technique using single-formant stimuli. **Journal of the Acoustical Society of America**, v. 101, n. 4, p. 2202-2213, Apr. 1997.

[50] HAMON, C.; MOULINES, E.; CHARPENTIER, F. A Diphone Synthesis System Based on Time-Domain Prosodic Modification of Speech. In: INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH AND SIGNAL PROCESSING, 1989, Glasgow. **Proceedings**… p. 238-242.

[51] WEIHMAN, T. **Processamento Digital de Sinais Aplicado à Transmissão de Voz**. Porto Alegre, 1992. 156 f. Dissertação (Mestrado em Engenharia) – Escola de Engenharia, Universidade Federal do Rio Grande do Sul.

[52] BRITISH TELECOMMUNICATIONS PLC, Andrew Lowry. **Apparatus for Synthesizing Speech by Varying Pitch**. Int. Cl^6. G10L 9/00. U.S. n. 5,787,398. 26 Aug. 1996, 28 Jul. 1998.