Artigo disponibilizado *on-line*

Revista Ilha Digital

Endereço eletrônico:
http://ilhadigital.florianopolis.ifsc.edu.br/

ARTIGO DE REVISÃO: SISTEMAS DE SÍNTESE DE FALA

Fernando Santana Pacheco[1]

Resumo: Este artigo apresenta uma revisão sobre sistemas de síntese de fala. Inicia com uma contextualização histórica, partindo dos sistemas mecânicos do século XVIII e progredindo até os atuais programas computacionais de geração de fala sintética. Os sistemas de síntese de fala dividem-se em duas categorias: resposta vocal e conversão texto-fala. Foca-se, neste artigo, nos sistemas de conversão de texto para fala (TTS, do inglês *text-to-speech*). Discutem-se aplicações atuais e futuras de tais sistemas. Por fim, apresenta-se a estrutura dos modernos sistemas TTS, dividida em análise linguística e processamento de sinais.

Palavras-chave: Processamento de sinais. Conversão texto-fala. Síntese de fala.

***Abstract**: This paper presents a review about speech synthesis systems. We begin with an historic overview, from 18th-century mechanical systems through recent speech synthesis softwares. There are two categories of speech synthesis systems: vocal response and text-to-speech (TTS). In this paper, we focus on TTS systems. We discuss current and future applications in this area. Finally, we present the structure of modern TTS systems, formed by linguistic analysis and signal processing.*

***Keywords**: Speech processing. Speech synthesis. Text-to-speech systems.*

[1] Professor do DAELN do IF-SC <fspacheco@ifsc.edu.br>.

1. INTRODUÇÃO

O desejo humano de dar fala a um objeto ou máquina acompanha a civilização há muito tempo. 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 eletroeletrônicos e *softwares* de síntese de fala foram sendo desenvolvidos. Na década de 1960, foi possível gerar fala a partir de um texto. A ideia, que no início parecia uma brincadeira, foi tomando corpo e encontra um extenso campo de aplicações no mundo atual.

É inegável o papel fundamental que a escrita tem na forma de comunicação humana. Entretanto, isso não significa que a mensagem escrita seja sempre a forma mais conveniente de se obter acesso a informações (EGASHIRA, 1992). Em diversas circunstâncias, não se pode interromper uma dada atividade para se 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 são possivelmente mais eficientes do que respostas visuais. Em um telefone comum, a única forma de acesso a informações é a partir 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, assim, viável.

Com esses 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.

Para apresentar uma revisão desse tema, este artigo está assim organizado: na seção 2, apresenta-se uma contextualização histórica, desde os

sistemas mecânicos do século XVII até os sistemas computacionais atuais; a seção 3 apresenta uma classificação dos sistemas de síntese de fala; uma revisão sobre o funcionamento de sistemas de conversão texto-fala é discutida na seção 4; as conclusões e os comentários finais são apresentados na seção 5.

2. HISTÓRICO

A potencialidade de aplicações de sistemas de síntese de fala despertou, há longo tempo, um forte interesse nessa área. O histórico apresentado a seguir, baseado na literatura aberta (EGASHIRA, 1992; DUTOIT, 1997; LEMMETTY, 1999; RUBIN; VATIKIOTIS-BATESON, 2001; HUANG; ACERO; HON, 2001; KLATT, 1987), mostra a evolução dos sistemas de síntese de fala.

2.1. 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[1]. A estrutura básica desses ressoadores é mostrada na Figura 1. Esses dispositivos eram similares à configuração do trato vocal humano e emitiam sons pelo uso de palhetas, como em instrumentos musicais.

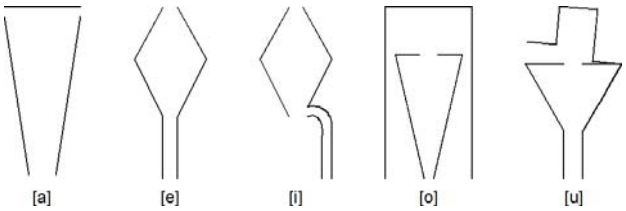

FIGURA 1 – Ressoadores de Kratzenstein.

Fonte: RUBIN; VATIKIOTIS-BATESON, 2001.

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 Figura 2, é apresentado um esboço da máquina de von Kempelen.

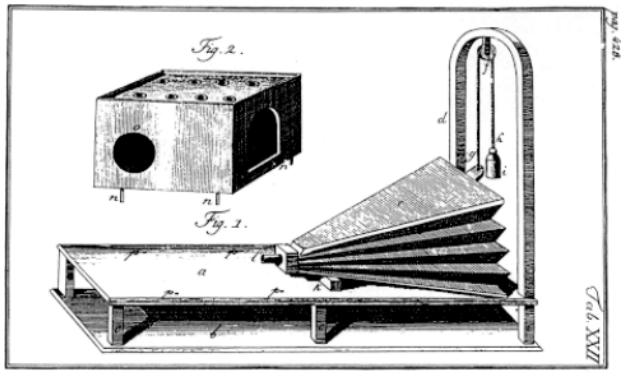

FIGURA 2 – Dispositivo mecânico de síntese de fala de von Kempelen.

Fonte: RUBIN; VATIKIOTIS-BATESON, 2001.

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 (ONDREJOVIC, 2000). Infelizmente, o mecanismo principal da máquina era um hábil jogador de xadrez colocado no interior, o que arrasou a reputação de von Kempelen.

Na metade do século XIX, Charles Wheatstone construiu uma versão da máquina falante de von Kempelen. Essa, 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*[2], 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 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 de 1960, mas sem muito sucesso.

2.2. Sistemas eletroeletrônicos

O primeiro sintetizador eletroeletrônico de fala foi desenvolvido por Stewart, em 1922. Dois

[1] Uma versão moderna dos ressoadores de Kratzenstein pode obtida no endereço eletrônico Vocal Vowels (VOCAL VOWELS, 2001).

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

circuitos ressonantes, excitados por uma cigarra elétrica, modelavam as duas frequências 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 eletroeletrônico 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!" (Science News Letter, 2000). O VODER consistia de chaves para seleção de uma fonte sonora ou de ruído, com controle da frequência fundamental a partir de um pedal. O sinal da fonte era transmitido por 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 Figura 3.

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 Figura 4, 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 de a 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.

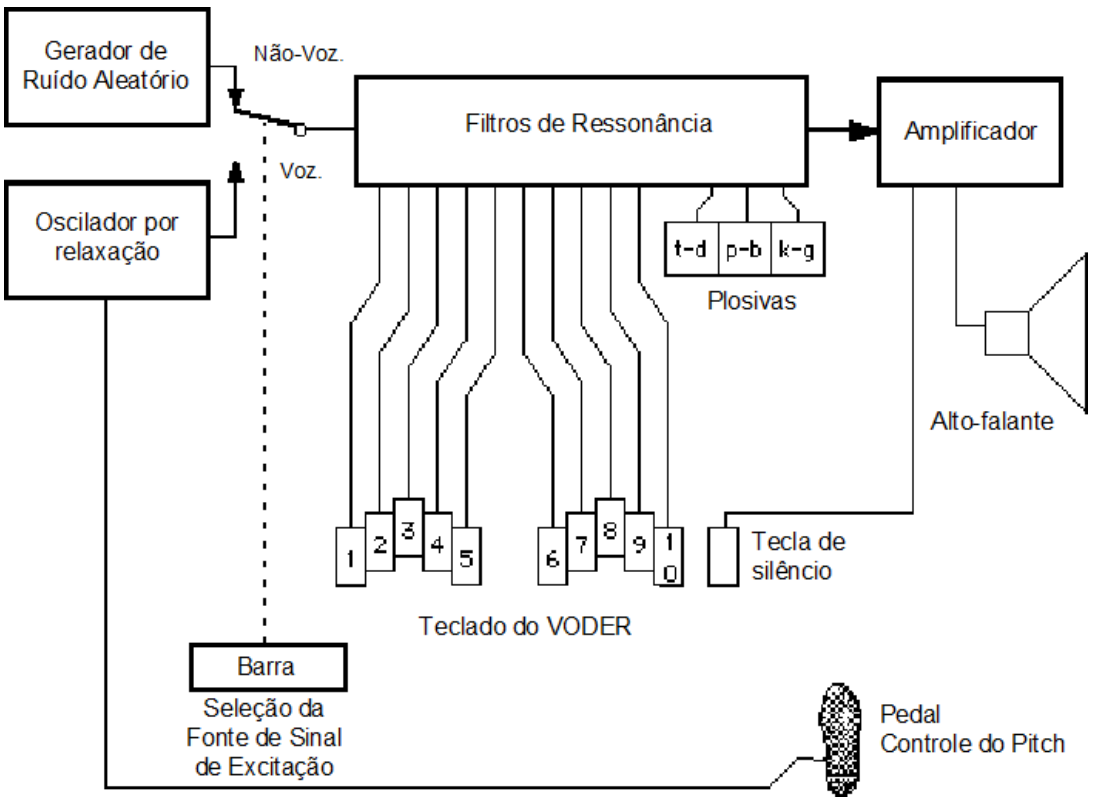

FIGURA 3 – Sintetizador VODER de 1939.

Fonte: RUBIN; VATIKIOTIS-BATESON, 2001.

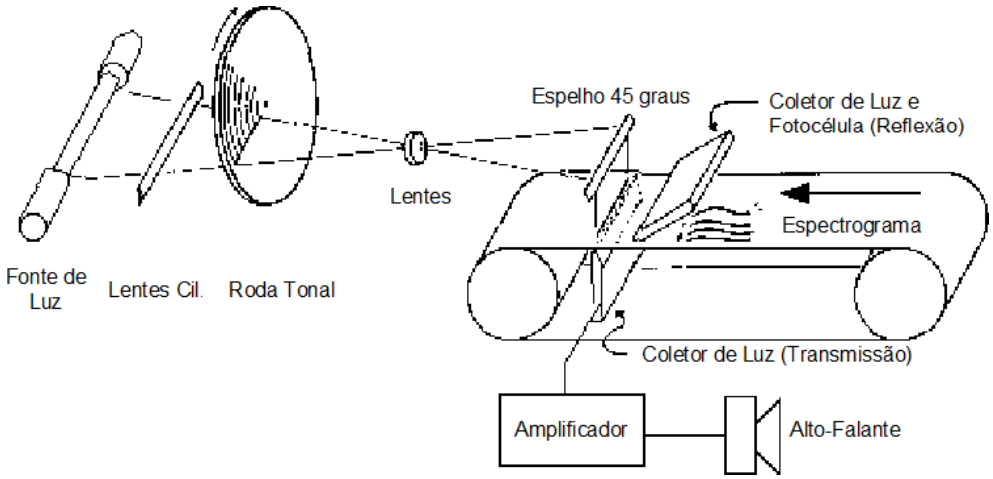

FIGURA 4 – Sistema *Pattern Playback* de 1951.

Fonte: RUBIN; VATIKIOTIS-BATESON, 2001.

O VODER e o *Pattern Playback* funcionavam a partir 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, em 1960, e o consequente 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 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 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[3] eram parametrizados por frequências 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.

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 (TELECOM-MUNICATIONS INDUSTRY PRODUCT BACKGROUND, 2001) e NextGen, da AT&T (SYRDAL *et al*., 2000).

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

Um resumo das etapas do desenvolvimento histórico de síntese da fala é apresentado na Figura 5.

3. CLASSIFICAÇÃO DOS SISTEMAS DE SINTESE 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 (EGASHIRA, 1992) 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 (VIEIRA; PACHECO, 2010). Entretanto, o domínio é restrito e bem definido, e a capacidade de armazenamento das mensagens é limitada pela memória disponível do sistema.

Já 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 a partir 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 a partir de três parâmetros básicos (RABINER, 1994):

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

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 a 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 texto irrestrito em uma representação falada.

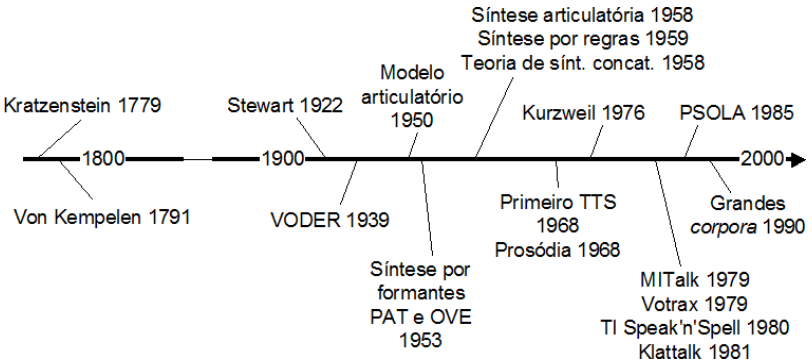

FIGURA 5 – Etapas do desenvolvimento histórico dos sistemas de síntese de fala.

Fonte: LEMMETTY, 1999 (adaptado).

4. SISTEMA TTS

O processo de conversão de um texto irrestrito em fala 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 Figura 6, 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 linguísticos, 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 frequência fundamental, duração e intensidade do sinal.

Ao final do processamento linguístico, os sons que devem ser sintetizados estão definidos. 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 linguística. Os modelos que realizam a síntese podem ser classificados em dois paradigmas, de acordo com o domínio em que atuam (DUTOIT, 1997):

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 a 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 (DUTOIT,1997). 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.

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 ideia 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 associados 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 a partir de técnicas de processamento de sinais.

FIGURA 6 – Diagrama básico do processo de conversão de texto em fala.

4.1. 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 (PAGE; BREEN, 1996):

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 a partir 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 (EGASHIRA, 1992; RABINER, 1994; DUTOIT, 1997; LEVINSON; OLIVE; TSCHIRGI, 1993; COX *et al*., 2000):

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 a partir 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 linguistas, 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.

4.2. Processamento linguístico

Na primeira etapa de um sistema texto-fala, 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 Figura 7, é mostrado um diagrama de blocos das etapas envolvidas na análise de texto para conversão texto-fala.

4.2.1. 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) (SPROAT; OLIVE, 1995). 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 uma marca de separação de frases é necessário eliminar outras possibilidades. No caso de abreviaturas, essas devem ser identificadas e expandidas.

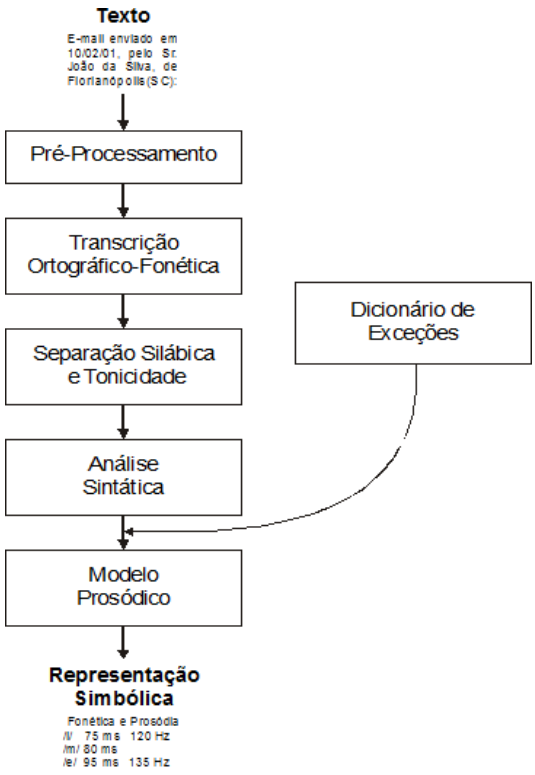

FIGURA 7 – Diagrama de blocos da etapa de análise do texto.

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 é 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".

4.2.2. Transcrição ortográfico-fonética

O objetivo da transcrição é a transformação da representação ortográfica em uma representação fonética. Se, para cada caractere, 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 (EGASHIRA, 1992; GOMES, 1998; COSTA NETO, 2000). 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] (FIGUEIREDO; NAVINER; AGUIAR NETO, 1997):

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 (LEVINSON; OLIVE; TSCHIRGI, 1993).

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 (EGASHIRA, 1992). 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.

4.2.3. 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 (EGASHIRA, 1992). 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 muito dependente da tonicidade.

Em Egashira (1992), 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 (SPROAT; OLIVE, 1995).

4.2.4. 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 ambiguidade 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 (EDGINGTON *et al*., 1996). 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 sequência válida de símbolos. Os símbolos correspondem a classes de palavras, grupos de palavras representando frases, orações ou mesmo frases inteiras.

4.2.5. 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 a partir 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 linguísticos 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 (SILVA; VIOLARO, 1995). Várias abordagens têm sido empregadas e uma revisão das técnicas é encontrada em Santen (1995). Destaca-se, para o caso do português, o modelo desenvolvido em Gomes (1998), 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 a partir do cálculo de um índice que leva em consideração a classificação gramatical do grupo prosódico em análise. Um ajuste do contorno geral é realizado a partir 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 linguístico. Cada modelo tem seu grau de complexidade e de relacionamento com os outros módulos de análise e um consequente grau de qualidade perceptual. Uma boa revisão das várias abordagens empregadas é encontrada em Dutoit (1997). Para o português, Silva e Violaro (1995) apresentam 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.

4.3. Processamento de sinais

Após a etapa de processamento linguístico, 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.

4.3.1. Síntese articulatória

A síntese articulatória tem por objetivo reproduzir o sinal de fala, modelando os mecanismos de sua produção natural (GABIOUD, 1994). 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 (LEMMETTY, 1999).

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, a partir 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 (PRADO, 1993). A estimação da função área pode ser realizada de duas formas:

a) diretamente pela observação da fala a partir de raios X ou ressonância magnética;
b) a partir de um mapeamento inverso ou acústico/articulatório, utilizando um processo analítico (PRADO, 1993).

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 (GABIOUD, 1994).

Como exemplo de parâmetros articulatórios de controle, pode-se citar o modelo descrito em Bickley, Stevens e Williams (1996) 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 (GABIOUD, 1994). Mesmo os sistemas no estado-da-arte não são capazes de gerar fala com a qualidade dos outros métodos baseados na abordagem de sinal.

4.3.2. Síntese por formantes

A síntese por formantes, também conhecida por síntese paramétrica, é baseada no modelo fonte-filtro de produção da fala. 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 Figura 8.

A principal característica da síntese por formantes é a implementação da função de transferência do trato vocal a partir da associação de seções de segunda ordem. Essas seções são conhecidas como ressonadores. A estrutura do ressonador digital de segunda ordem é ilustrada na Figura 9.

Os ressonadores podem ser associados em cascata ou paralelo. Para utilizar as melhores características de cada uma das associações, algumas implementações híbridas foram propostas na literatura Gomes (1998) e Klatt (1980). Uma das mais conhecidas é a mostrada na Figura 10, o sintetizador de Klatt (1980). 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.

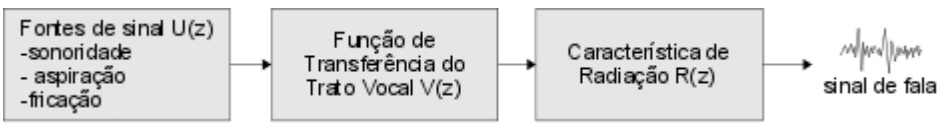

FIGURA 8 – Esquema simplificado do processo de produção da fala empregado na síntese por formantes.

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 (STYGER; KELLER, 1994). A chave SW controla a mudança entre a estrutura em série e paralelo.

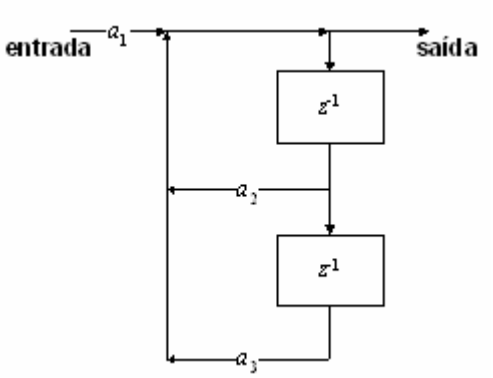

FIGURA 9 – Estrutura do ressonador digital de segunda ordem.

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 (EGASHIRA, 1992). 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 a partir de 39 parâmetros, atualizados a cada 5ms. Na configuração padrão, o sistema opera com uma taxa de amostragem de 10 kHz.

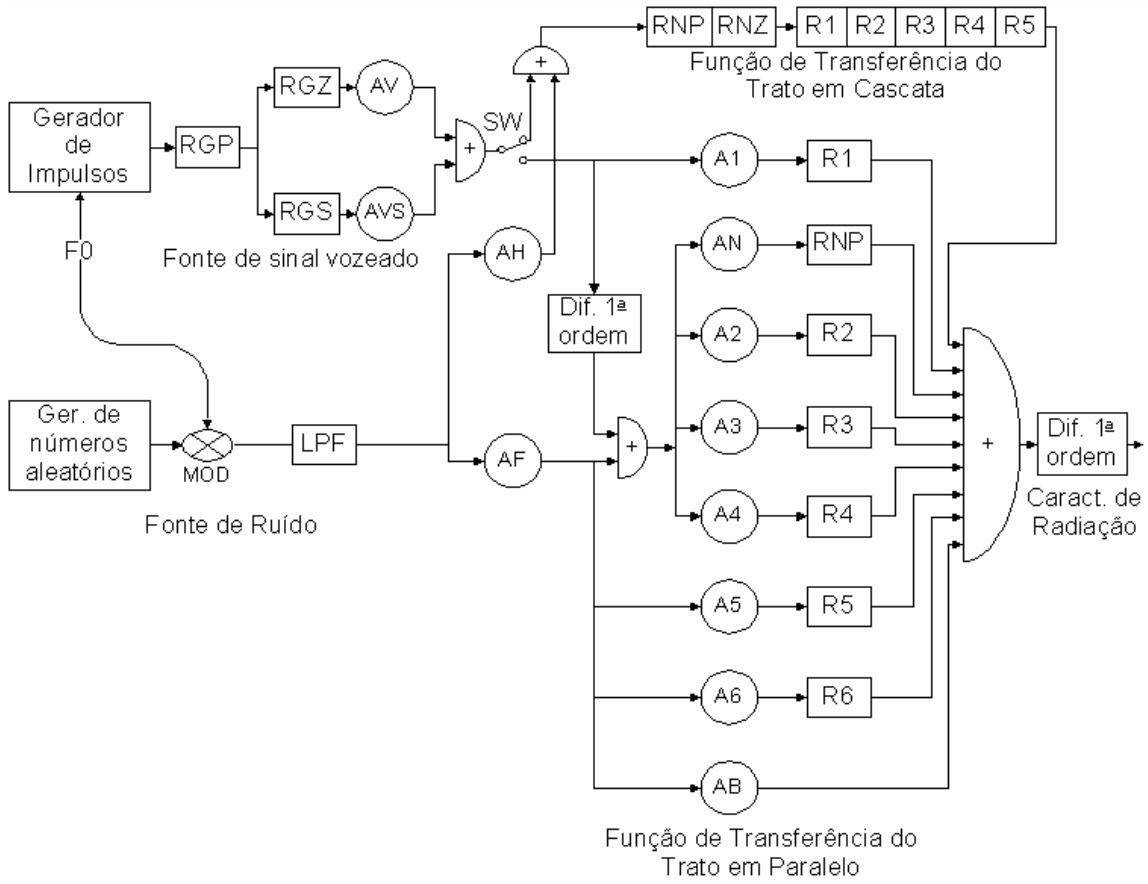

FIGURA 10 – Diagrama do sintetizador de Klatt (1980).

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 (DUTOIT, 1997). Alguns trabalhos, como os de Huang, Acero e Hon (2001) buscam a obtenção dos parâmetros de forma automática, empregando técnicas inicialmente utilizadas em reconhecimento de fala, como os modelos ocultos de Markov (HMM).

4.3.3. 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 a partir da etapa de processamento linguístico. Com uma etapa de concatenação e alteração de parâmetros prosódicos, a fala sintetizada é gerada. O diagrama de blocos desse processo é mostrado na Figura 11.

Em oposição à síntese por formantes, aqui não há necessidade de definição de regras de transição entre sons, pois essas 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 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.

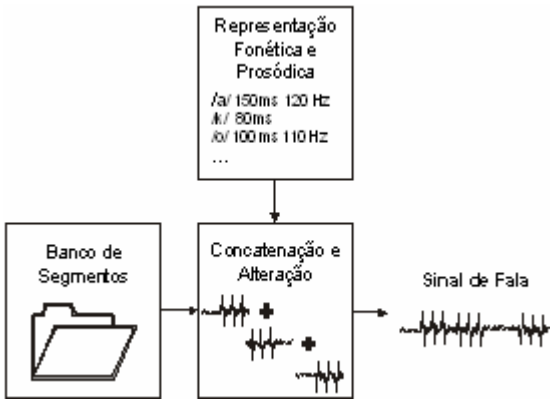

FIGURA 11 – Diagrama de blocos da síntese do sinal de fala pela técnica concatenativa.

As descontinuidades espectrais ocorrem quando os formantes de segmentos adjacentes não têm os mesmos valores e estão relacionadas, principalmente, à coarticulação, que pode ser entendida como a influência de um fonema sobre outro.

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.

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 corpus1 de extração das unidades;
c) gravação;
d) segmentação.

Essas etapas são mostradas na Figura 12.

FIGURA 12 – Etapas envolvidas na criação de um banco de unidades para síntese concatenativa.

5. CONSIDERAÇÕES FINAIS

Este artigo apresentou uma revisão dos sistemas de síntese de fala, com ênfase nos sistemas de conversão texto-fala. Uma contextualização histórica foi apresentada, além de uma revisão detalhada das aplicações e módulos necessários para uma adequada conversão.

REFERÊNCIAS

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.

COSTA NETO, M. L. da. **Conversor Texto-Fala de Alta Qualidade para a Língua Portuguesa**. Campina Grande, 2000. 112 p. Exame de Qualificação (Doutorado em Engenharia Elétrica) – Centro de Ciências e Tecnologia, Universidade Federal da Paraíba.

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.

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

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.

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

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.

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.

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

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

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.

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.

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

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.

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.

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.

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.

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.

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

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

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.

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.

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.

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.

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.

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

VIEIRA, J.; PACHECO, F. S. **Desenvolvimento de um módulo de resposta vocal para a plataforma microcontrolada Arduino**. In: I CONGRESSO DE INICIAÇÃO CIENTÍFICA E PÓS-GRADUAÇÃO SUL BRASIL, 2010, Florianópolis. Anais....

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