UNIVERSIDADE FEDERAL DE SANTA CATARINA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

Almir dos Santos Albuquerque

Análise Comparativa dos Métodos de Sintetização de Voz

Dissertação de mestrado submetida à Universidade Federal de Santa Catarina como parte dos requisitos para a obtenção do grau de Mestre em Ciência da Computação

Dr. Roberto Willrich
Orientador

Florianópolis, dezembro de 2001.

Análise Comparativa dos Métodos de Sintetização de Voz

Almir dos Santos Albuquerque

Esta Dissertação foi julgada adequada para a obtenção do título de Mestre em Ciência da Computação Área de Concentração Sistemas de Computação e aprovada em sua forma final pelo Programa de Pós-Graduação em Ciência da Computação.

_____

Prof. Dr. Fernando Álvaro Ostuni Gauthier

Banca Examinadora

_____

Prof. Dr. Roberto Willrich (Orientador)

_____

Prof. Dr. Luís Fernando Jacintho Maia

_____

Prof. Dr. Vitório Bruno Mazzola

DEDICATÓRIA

Este trabalho é dedicado a todas as pessoas que de uma forma ou outra me auxiliaram no seu desenvolvimento e a todos pesquisadores que buscam fazer sua contribuição para o desenvolvimento, não só tecnológico, mas também de metodologias de melhor utilização da tecnologia.

AGRADECIMENTOS

À Deus, que me iluminou e me deu força para que eu pudesse alcançar meus objetivos e superar as barreiras durante todo caminho.

Aos meus Amparadores que sempre estão do meu lado mostrando o melhor caminho e suprindo-me de boas energias.

Aos meus pais: Raimundo Albuquerque e Rosália Régis dos Santos que sempre me apoiaram e sempre fizeram tudo que estava ao seu alcance para que eu tivesse uma boa educação e um dia pudesse me formar.

Aos meus irmãos e familiares que sempre torceram por mim, fazendo uma corrente positiva para que eu alcançasse o meu objetivo.

À minha esposa Sílvia Valéria dos Santos Albuquerque, que com toda sua paciência esteve sempre do meu lado me apoiando nas horas difíceis.

Aos meus queridos filhos: Carolina, Gabriela e Gabriel, que apesar da inocência, colaboram como fonte de inspiração na luta do dia-a-dia.

Ao professor Pedro A. Barbetta, por me auxiliar nos cálculos estatísticos.

À Fonoaudióloga Dra. Cecília Medeiros Oliveira, pelas indicações bibliográficas de fonoaudiologia.

Ao meu orientador Prof. Dr. Roberto Willrich, pelo seu profissionalismo e dedicação, que me guiou e deu todos os subsídios para que esse trabalho fosse desenvolvido.

E a todos que direta e indiretamente cooperam ou me apoiaram no decorrer desse projeto.

Lista de Figuras

Lista de Tabelas

Lista de Gráficos

Índice

Resumo

Na síntese de voz ***text-to-speech*** (TTS), o computador recebe como entrada, um texto digitado ou em memória e devolve, por meio de alto-falantes a leitura em voz alta do mesmo texto. As técnicas utilizadas para a síntese TTS são: concatenação, formantes, LPC e articulatória, sendo as duas primeiras as mais utilizadas. Esta dissertação tem como meta principal realizar uma análise comparativa destes dois métodos, por meio de análises teóricas e comparação de resultados de softwares TTS que seguem estas duas técnicas.

Abstract

In the voice synthesis text-to-speech (TTS), the computer receives as entered, a text typed or in memory and returns, by loudspeakers the reading in high voice of the same text. The techniques used for synthesis TTS are: concatenation, formants, LPC and articulatory. The two first ones are the most used. This thesis has, as main goal, to carry through a comparative analysis of the concatenative and formant techniques, using a theoretical analyses and comparison of results of these softwares TTS.

Capítulo 1 Introdução

A fala é a principal maneira de comunicação entre as pessoas. Ainda hoje, na nossa sociedade, as pessoas com deficiência vocal (pessoas mudas) enfrentam certas dificuldades no seu dia-a-dia. Essas mesmas dificuldades também fazem parte da vida dos deficientes visuais. A geração automática pelo computador de formas de onda da voz, conhecida como síntese de voz, tem recebido atenção da comunidade acadêmica e de profissionais em geral por várias décadas. A síntese de voz vem ajudar pessoas nessas dificuldades.

"Os deficientes visuais agora também podem consultar seus extratos de conta-corrente, poupança e fazer aplicações pela Internet sem precisar da ajuda de outras pessoas. A novidade foi anunciada pelo Bradesco, que está distribuindo 20 mil cópias demo do CD-ROM Bradesco Net - Internet Banking para Deficientes Visuais..." [Net Estado, 1998].

Os progressos mais recentes na área têm sido motivados por diversos fatores, dentre eles, destacam-se:

- Rápido aumento da habilidade dos computadores para realizar tarefas velozmente e com baixo custo;
- Um grande aumento no número de textos disponíveis e banco de vozes;
- Melhoramento da tecnologia de reconhecimento de voz e de síntese.

Os estudos realizados na área de síntese de voz vêm se despontando cada vez mais, principalmente com o advento dos avanços tecnológicos da informática; prometendo resolver - pelo menos em parte - os problemas sociais de pessoas portadoras desse certo tipo de deficiência.

Os progressos recentes em síntese de voz permitiram a produção de sintetizadores com alta inteligibilidade, mas a qualidade do som e a naturalidade ainda continuam sendo o grande problema. Mesmo assim, a qualidade dos produtos alcança

um nível adequado para várias aplicações, tais como multimídia e telecomunicações. Com alguma informação audiovisual ou animação facial é possível aumentar a inteligibilidade da voz consideravelmente.

Um tipo de síntese de voz é a ***text-to-speech*** (TTS). Por ela, o computador recebe como entrada um arquivo de texto ou um texto digitado e produz, por meio de alto-falantes, a leitura em voz alta do mesmo texto. A síntese *text-to-speech* consiste basicamente de duas fases principais. A primeira é a análise do texto, na qual o texto de entrada é transcrito em fonemas ou outra representação lingüística, e a segunda fase é a geração de formas de onda da voz, na qual a saída acústica é a reprodução dessas representações e da informação prosódica (duração e entonação). Essas duas fases são usualmente chamadas de síntese de alto e baixo nível, respectivamente.

A maneira mais simples de produzir voz sintética é reproduzir amostras pré-gravadas da voz natural, tais como sentenças ou palavras simples, para depois organizá-las ordenadamente de acordo com o texto de entrada e assim serem formadas frases e sentenças inteligíveis. Este método de síntese de voz é conhecido como **Síntese de Concatenação**. Esta concatenação gera alta qualidade e naturalidade, não obstante um vocabulário limitado. O método é muito adaptável para alguns sistemas de informação. Entretanto, resta claro que não se pode criar um banco de dados composto por todas as palavras e nomes comuns no mundo. Para uma irrestrita síntese de voz é necessário dispor de pedaços mais curtos do sinal de voz, tais como sílabas, fonemas, ou mesmo segmentos mais curtos. É dessa forma que os sistemas de síntese que utilizam esse método, trabalham atualmente.

Outro método largamente aplicado na produção de síntese de voz é a **síntese de formantes**. O método é algumas vezes chamado de analogia terminal porque modela a fonte de som e as freqüências formantes e não qualquer característica física do trato vocal. O sinal de excitação poderia ser vozeado com freqüência fundamental ou ruído não vozeado. A excitação misturada destes dois pode, também, ser usada para consoantes vozeadas e alguns sons da aspiração.

Há também o método **LPC** (*Linear Predictive Coding*), que desponta como uma técnica de síntese de voz muito poderosa, capaz de codificar a voz com boa qualidade a uma baixa taxa de bit. Este método analisa o sinal de fala formantes,

removendo os efeitos do sinal e avaliando a intensidade e freqüência do sinal restante. O processo de remover os formantes é chamado de filtro inverso e o sinal restante é chamado de resíduo. O LPC sintetiza o sinal de voz invertendo o processo: usa o resíduo para criar um sinal de fonte e usa os formantes para criar um filtro (como se fosse uma espécie de tubo); a fonte (sinal) percorre pelo filtro, gerando como resultado a voz (fala).

Há ainda um método para gerar voz artificial que modela diretamente o sistema de produção da fala humana, chamado de **síntese articulatória**, o qual tipicamente envolve o processo de modelar os articuladores humanos e as cordas vocais. Os articuladores são usualmente modelados com um conjunto de funções de área de seções pequenas. O modelo de corda vocal é usado para gerar um sinal de excitação apropriado. A síntese articulatória se constitui em promissora tecnologia na produção de voz sintetizada de alta qualidade que, devido sua complexidade, não teve o potencial plenamente explorado.

Todos os métodos de síntese apresentam benefícios e problemas, de forma que é realmente difícil dizer qual deles é o melhor. Os métodos de Concatenação, Formantes e LPC alcançaram muitos resultados promissores, mas também a síntese Articulatória pode surgir como um método potencial no futuro.

1.1 Objetivo da Dissertação

O presente trabalho tem o objetivo de fazer uma análise comparativa entre os métodos de Concatenação e Formantes, utilizados na síntese TTS, através de análises teóricas e experimentação. Esta última foi realizada a partir de uma avaliação da qualidade de voz, utilizando softwares que implementam os métodos analisados. A análise de voz foi feita utilizando-se a pré-gravação das sentenças (frases) e fonemas, que são utilizados nas técnicas de fonoaudiologia para avaliação qualitativa da voz. Este trabalho espera ainda ser uma contribuição significativa sobre o assunto síntese de voz para a comunidade acadêmica e também um material que possa ser utilizado para o desenvolvimento de novas pesquisas, aplicações e dissertações na área.

Os objetivos específicos são os seguintes:

- Estudos das tecnologias TTS existentes.
- Testes e experimentação de softwares.
- Efetuar uma análise teórica dos métodos.
- Efetuar uma análise através da experimentação prática dos métodos.

1.2 Organização da Dissertação

O primeiro capítulo apresenta o contexto motivador do tema desta dissertação, seus principais objetivos e estrutura do texto dissertativo. O segundo capítulo, trata da fala humana, sua grafia e principais características de produção da voz pelo corpo humano. O terceiro capítulo trata da tecnologia TTS, seus elementos, características e funcionamento. O quarto capítulo, trata do corpo teórico específico da sintetização automatizada da voz descrevendo os seus métodos, técnicas e algoritmos. O quinto capítulo traz a análise comparativa dos métodos e técnicas estudadas. Finalmente, o sexto capítulo apresenta as conclusões deste estudo.

Capítulo 2 Fala e sua Grafia

Este capítulo apresenta o aparelho fonador. Sua compreensão é muito importante para o entendimento do funcionamento da produção da voz pelo corpo humano, os tipos de sons que compõem a fala e a freqüência do som da voz masculina e feminina. Este capítulo também apresenta as unidades básicas de uma língua, que são os fonemas e um alfabeto fonético, para nos dar o entendimento da transcrição e leitura de um som em qualquer idioma. Ele apresenta ainda os sons emitidos na língua portuguesa.

2.1 A Fala Humana

A compreensão do funcionamento do aparelho fonador é importante para entender os parâmetros envolvidos na produção da voz, e por esse motivo ainda hoje é um tópico de ativas pesquisas na área de fonética acústica e articulatória [KLATT e KLATT, 1987].

O aparelho fonador humano é constituído pelas seguintes partes:

- Os pulmões, os brônquios e a traquéia, que são os órgãos respiratórios responsáveis pelo fornecimento da corrente de ar, que corresponde à "matéria-prima" da fonação;
- A laringe, na qual se localizam as cordas vocais, que produzem a energia sonora utilizada na fala;
- As cavidades supra-laríngeas (faringe, boca e fossas nasais), que funcionam como uma caixa de ressonância.

A cavidade bucal pode variar profundamente de forma e volume, graças aos movimentos dos órgãos ativos, sobretudo da língua. Através da movimentação do palato mole (vélum), a cavidade nasal pode ser acoplada à cavidade bucal. Estas duas últimas partes, a laringe e as cavidades supra-laríngeas, são também conhecidas como trato vocálico.

O trato vocálico pode ser considerado como um tubo acústico de seção variável, com início nas cordas vocais e que termina nos lábios e narinas. Em um adulto do sexo masculino apresenta aproximadamente 17 cm de comprimento, sendo a área seccional determinada pela posição dos lábios, maxilares, língua e vélum, e pode variar de zero (no caso de lábios fechados) até aproximadamente 20 cm^2. A cavidade nasal tem em média 12 cm de comprimento e volume aproximado de 60 cm^3 [FLANAGAN, 1972].

Um órgão essencial na fonação é a laringe, que corresponde a um tubo de paredes cartilaginosas semi-rígidas, contendo dois pares sobrepostos de membranas, denominadas cordas vocais, que delimitam uma fenda chamada glote. Quando pretende-se emitir um som, utilizando-se as cordas vocais, a glote é fechada, e sob a ação de um esforço expiratório, o ar afasta ligeiramente as bordas das cordas vocais e escoa pela glote. Simultaneamente, as cordas vocais começam a vibrar, permitindo a passagem de pulsos de ar, que excitam o sistema acústico localizado imediatamente acima das cordas vocais [SANCHES, 1989].

2.2 Sons que compõem a Fala

A voz, produzida pela passagem do ar fornecido pelos pulmões no trato vocálico, pode ser gerada de três maneiras distintas originando *sons sonoros* ou *vocálicos, sons fricativos* e *sons plosivo*s. O modo como esses sons são produzidos foi descrito detalhadamente por vários autores como FLANAGAN (1972), CAMPOS (1980), SANCHES (1989), CASAES (1990) e O'MALLEY (1990).

Os *sons sonoros* ou *vocálicos* são produzidos pela elevação da pressão de ar nos pulmões, forçando a sua passagem através do orifício das cordas vocais (glote) e causando sua vibração. Essa vibração obstrui a passagem de ar de maneira periódica, causando a interrupção do fluxo de ar, que excita o trato vocálico. O período dessa interrupção é chamado de "pitch" e seu inverso é a "freqüência fundamental ($F\ 0$)".

Os *sons fricativos* são gerados pela formação de uma constrição em algum ponto do trato vocálico, normalmente nos lábios, forçando a passagem de ar através dessa constrição com velocidade suficiente para produzir turbulência, criando assim, uma fonte de "ruído branco". Podem ser produzidos com ou sem vibração das cordas

vocais, condição em que serão chamados respectivamente de *fricativos sonoros* ou *fricativos surdo*s.

Os *sons plosivos* resultam da constrição completa do trato vocálico em alguma parte, com acumulação de pressão e liberação abrupta em seguida. O ponto de completo fechamento pode ser efetuado em várias zonas de articulação e a excitação pode ou não causar vibração das cordas vocais, como no caso dos sons fricativos.

2.3 Freqüências da Fala

De acordo com CAMPOS (1980), à medida que os sons, gerados por qualquer uma das formas descritas anteriormente, propagam-se pelo trato vocálico; apresentam alteração em seu espectro de freqüências e com ressonância em determinadas freqüências. Essas freqüências são denominadas *freqüências formantes* do som, ou simplesmente *formante*s, sendo o número de formantes variável conforme o som. Um som pode ser caracterizado pelas suas três freqüências formantes mais baixas, que são comumente designadas por *F 1 , F 2* e *F 3* .

As freqüências formantes dependem da forma do trato vocálico e conseqüentemente as propriedades espectrais do som produzido variam em decorrência da geometria do trato vocálico.

Juntamente com a freqüência fundamental, as formantes constituem os principais parâmetros acústicos da voz. Tipicamente, para uma voz masculina a freqüência fundamental varia entre 60 e 240 Hz, enquanto que as três formantes variam em torno de 500 Hz, 1500 Hz e 2500 Hz. Para uma voz feminina, a freqüência fundamental tem valores entre 100 e 400 Hz, enquanto que as demais formantes estão aproximadamente 10% acima das formantes masculinas [O'MALLEY, 1991].

2.4 Fonemas

Os *fonemas* são as unidades básicas de uma Língua, e têm a propriedade de mudar o sentido de uma palavra quando uma unidade é substituída por outra (FLANAGAN, 1972). Por exemplo, na série de palavras *dia, fia, mia, pia, tia* e *via,* a distinção entre as palavras ocorre apenas pelo elemento consonântico inicial, que

caracterizam unidades sonoras distintas, correspondendo cada uma delas a um fonema diferente.

Entendidos como uma unidade de som no início do século XIX, os fonemas são hoje considerados como unidades mentais, abstratas, das quais o som é a sua realização física. O fonema é uma unidade da Língua e os sons ou fones são unidades da fala [CALLOU e LEITE, 1990].

Os fonemas são comuns a todos os indivíduos que falam a mesma Língua, enquanto que os sons que o representam variam não apenas de um indivíduo para outro, como também, para um mesmo indivíduo de um ato para outro [PAIS, 1986].

Para distingui-los dos sons realmente produzidos, os fonemas são normalmente representados entre barras oblíquas (/ /), enquanto que os sons são representados entre colchetes ([ ]). No caso da representação entre barras, a transcrição é dita fonológica e no caso da representação entre colchetes, a transcrição é fonética. A palavra *dia* por exemplo, é representada pelos fonemas /dia/ e pode ser pronunciada como [diia] [EGASHIRA, 1992].

Aos vários sons que realizam o mesmo fonema dá-se a denominação *variantes* ou *alofone*s. Por exemplo, os fonemas /d/ e /t/ apresentam em determinados dialetos do Português uma realização palatal diante do /i/, como nas palavras *tia* e *dia* e uma realização alveolar ou dental diante das outras vogais como nas palavras *dad*o, *doca*s, *tel*a, *tua* [CALLOU e LEITE, 1990].

Cada idioma tem seus próprios fonemas, que são elementos fônicos dotados de função representativa no sistema. A nossa Língua Portuguesa possui 26 (vinte e seis) fonemas segmentais, sendo 19 (dezenove) consoantes e 7 (sete) vogais, e um fonema supra-segmental, o *acent*o, que não é um segmento e sim uma qualidade que se superpõe a certos segmentos [CALLOU e LEITE, 1990]. Por exemplo as palavras como *dívida* e *divid*a; *sábia, sabia* e *sabiá* são diferentes entre si apenas pela posição do acento tônico. Para que as seqüências fônicas de uma Língua sejam reproduzidas na escrita, utilizam-se sinais gráficos representativos desses sons, que são as *letras* ou *grafema*s. De acordo com CEGALLA (1977), não existe uma correspondência exata entre número de letras e o número de fonemas nos idiomas. Abaixo mostramos alguns exemplos fornecidos por ele:

- Na Língua Portuguesa pode-se observar que uma mesma letra pode representar mais de um fonema, como por exemplo na seqüência de palavras *exame, xale* e *próximo;*
- Um mesmo fonema pode ser figurado por mais de uma letra, como nas palavras *casa, exílio, cozinha* ou representado por um grupo de duas letras, os dígrafos, como nas palavras *machado, mulher, unha, missa* e *carro;*
- Há ainda letras que por vezes não representam fonemas, funcionando somente como notações léxicas, como nas palavras *campo [cãpo]* e *regu*e, na qual o ***u*** é insonoro, para que não seja proferido *reje;*
- E também são utilizadas letras simplesmente decorativas, na medida em que não representam fonemas e não funcionam como notações léxicas, como em *discípulo [dicipulo], hotel [otél]* e *exceção [esesão*]; além de fonemas que, em certos casos, não são representados graficamente como em *eram [érãu], falam [fálãu]*.

Existe um sistema ortográfico que regulamenta essa representação na linguagem escrita, sendo a ortografia vigente até hoje no Brasil, a oficialmente adotada nas normas do Vocabulário Ortográfico de 1943, com as alterações determinadas pela Lei nr. 5.765 de 18 de dezembro de 1971 [FERREIRA, 1986].

Existem ainda conforme CALLOU e LEITE (1990), muitas discussões e propostas no sentido de se ter a possibilidade de uma reforma ortográfica que leve em consideração as relações entre a pronúncia e a ortografia portuguesa do Brasil e de Portugal e que também procure aproximar o sistema de fonemas ao sistema de letras, como a substituição da letra "s" por "z" em palavras nas quais a letra "s" representa o som [z] *(casa, mes*a) e de "ss", "c", "ç" e "x" por "s" para representarem o som [s] *(posso, cedo, laço, próxim*o).

Segundo esses estudiosos ainda, esse sistema integrado letra-fonema parece ser inviável, pois em um País com o tamanho do Brasil qualquer tentativa de aproximação seria precária e deixaria a desejar, já que teriam de ser levados em consideração todas as diferenças regionais, sócio-culturais e até mesmo individuais.

Dizem ainda que, quanto mais um idioma desenvolve-se, mais o sistema ortográfico afasta-se do sistema fonológico, o que tem ocorrido com os idiomas Inglês e Francês.

A aproximação de fonemas citada anteriormente, enfrentaria sérios problemas por causa das palavras homófonas, como por exemplo a representação do som [s] sempre pela letra "s" e do som [z] sempre pela letra "z", nas palavras *coser/cozer, expiar/espiar, cessão/sessão/seçã*o, além de palavras como *aterrisar* e *subsídio*s, para as quais existem normalmente duas pronúncias, *aterri[s]ar* e *aterri[z]a*r, *sub[s]ídios* e *sub[z]ídio*s.

Então, levando-se todos esses argumentos, a convivência com o sistema ortográfico atual parece ainda inevitável, pelo menos a curto e médio prazo.

2.5 O Alfabeto Fonético

Para simbolizar na escrita a pronúncia real de um som utiliza-se de um alfabeto especial, conhecido como alfabeto fonético. A finalidade da transcrição fonética e portanto, do alfabeto fonético é justamente a transcrição e a leitura de um som em qualquer Idioma por uma pessoa treinada. Assim, esse alfabeto deve apresentar convenções inequívocas e de maneira explícita. Algumas dessas convenções tornaram-se bastante difundidas, como por exemplo, as propostas no "International Phonetic Alphabet - IPA" pela Sociedade Internacional de Fonética. Esse alfabeto, no entanto, emprega caracteres pouco comuns em máquinas de escrever e computadores, o que dificulta sua utilização [CALLOU e LEITE, 1990].

A Tabela 1 apresenta o alfabeto fonético baseado nos símbolos IPA, e outros dois possíveis alfabetos para a Língua Portuguesa, sendo um deles baseado em letras maiúsculas, utilizando até dois caracteres e outro, utilizando apenas um único caractere, proposto por Dimas Trevizan Chbane [CHBANE, 1994].

Símbolos IPA (CUNHA e CINTRA,1985)	Símbolos com 1 ou 2 caracteres (CAMPOS, 1980)	Símbolos com 1 Caracter (CHBANE,1994)	Exemplos
a	A	a	pá, gato
e	E	e	vê, medo
ε	EH	é	pé, ferro
I	I	i	vir, bico
o	O	o	avô, morro
⊃	OH	ó	avó, cola
u	U	u	utu, bambu
ã	AN	ã	Lã
m	M	m	mar, amigo
n	N	n	nada, cano
η	NH	ñ	vinha, caminho
b	B	b	bravo, ambos
p	P	p	pai, caprino
d	D	d	dar, andar
t	T	t	tu, canto
g	G	g	frango, agrado
k	C	k	casa, que
f	F	f	filho, afiar
v	V	v	vinho, uva
s	S	s	saber, posso
z	Z	z	bazar, casa
∫	X	x	chover, xarope
¥	J	j	já, jarra
l	L	l	lado, veludo
l	L	|	alto, fuzil
λ	LH	L	filho, pilha
r	R	r	caro, cores
r	R	h	mar, carta
R	RR	R	carro, roda

TABELA 1. Alfabetos fonéticos para a Língua Portuguesa [CHBANE, 1994].

Na Língua Portuguesa, os fonemas /i/ e /u/, quando formam sílaba com outra vogal, são chamados *semi-vogais* e normalmente transcritos como [j] e [w], como em [rej] e [mew] [CUNHA e CINTRA, 1985].

Conforme mostrado por CAMPOS (1980), um ditongo pode ser considerado como junção de duas vogais de menor duração com transições suaves entre as suas freqüências formantes.

2.6 Sons da Língua Portuguesa

Vários autores já discutiram exaustivamente detalhes sobre a classificação dos sons da Língua Portuguesa, tais como: CUNHA e CINTRA (1985), PAIS (1986), CALLOU e LEITE (1990) e CASAES (1990). É de consenso entre eles que existe duas classes de sons em nossa língua, o som das vogais e o som das consoantes.

2.6.1 Classificação das Vogais

As vogais são normalmente classificadas segundo quatro critérios: quanto à região de articulação, quanto ao timbre, quanto ao papel das cavidades bucal e nasal e quanto à intensidade. Os três primeiros critérios são fundamentalmente de base articulatória e o último de base acústica.

Região de Articulação

Diz respeito ao ponto ou parte deste, em que se dá o contato ou aproximação dos órgãos que cooperam para a produção dos fonemas, no caso das vogais, a língua e o palato. Produz-se a *vogal média* [a] mantendo-se a língua baixa, quase em posição de descanso, e a boca entreaberta.

Para passar da vogal ***a*** para as *vogais anteriores* ([e], Zé], [i]) levanta-se gradualmente a parte anterior da língua em direção ao palato duro, ao mesmo tempo em que diminui-se a abertura da boca. Para emitir as *vogais posteriores* ([o], [ó], [u]), eleva-se a parte posterior da língua em direção ao véu palatino, arredondando progressivamente os lábios.

Timbre

Refere-se ao maior ou menor grau de abertura dos lábios. Essa abertura é máxima para a vogal [a] e mínima para as vogais [i] e [u].

Papel das Cavidades Bucal e Nasal

Depende da posição da úvula durante a passagem de ar pelo trato vocálico. Quando a corrente sonora é impedida de ressonar na cavidade nasal devido à posição levantada da úvula, tem-se a produção das *vogais orais* ([a], [e], [é], [i], [o], [ó], [u]).

Quando as fossas nasais são acopladas à cavidade bucal através do abaixamento da úvula, parte da corrente sonora ressoa na cavidade nasal, produzindo as *vogais nasais* ([ã], [.], [0], [õ], [ß]).

Intensidade

É uma qualidade física da vogal que depende da força expiratória e da amplitude da vibração das cordas vocais.

As vogais que se encontram nas sílabas pronunciadas com maior intensidade chamam-se *tônicas* e caracterizam-se no idioma Português por um reforço da energia expiratória.

As vogais que se encontram em sílabas não acentuadas denominam-se *átona*s. No idioma Português normal do Brasil, as vogais [é] e [ó] não aparecem em posição átona, assim como as vogais nasais.

No Brasil, nas sílabas átonas ocorre a chamada "neutralização", na qual as vogais anteriores "e" e "i", quando em posição final absoluta, são reduzidas a uma única vogal [i], como na palavra *tarde* → *[tardi]* e as vogais posteriores "o" e "u", quando nessa situação também são reduzidas a uma única vogal [u], como no caso da palavra *povo* → *[povu]*.

2.6.2 Classificação das Consoantes

São 19 (dezenove) as consoantes da Língua Portuguesa e tradicionalmente classificadas em função de quatro critérios de base articulatória, ou seja, quanto ao modo de articulação, quanto ao ponto de articulação, quanto à função das cordas vocais e quanto ao papel das cavidades bucal e nasal.

Modo de Articulação

Refere-se à maneira pela qual os fonemas consonantais são articulados. Vindo da laringe, a corrente de ar chega à boca, onde encontra obstáculos totais ou parciais da parte dos órgãos bucais. Se o fechamento dos lábios ou a interrupção da corrente de ar é total, tem-se as *consoantes oclusivas* ([p], [t], [k], [b], [d], [g]); se o fechamento for parcial, produz-se as *consoantes constritiva*s.

As consoantes constritivas dependendo de como a corrente expiratória escapa, podem ser:

- *fricativas:* são produzidas quando o trato vocálico é excitado por um fluxo de ar turbulento, que se forma quando a corrente expiratória passa pela constrição ([f], [s], [x], [v], [z], [j]).
- *vibrantes:* são caracterizadas pelo movimento vibratório rápido da língua ([r]) ou da úvula ([R]), que provocam breves interrupções da passagem da corrente expiratória.
- *laterais:* caracterizam-se pela passagem da corrente expiratória pelos dois lados da cavidade bucal, em virtude de um obstáculo formado no centro desta pelo contato da língua com os alvéolos dos dentes ([l]) ou com o palato ([L]).

Ponto de Articulação

Diz respeito ao lugar onde os órgãos fonadores entram em contato para a emissão do som, podendo ser bilabiais ([p], [b], [m]), labiodentais ([f], [v]), lingüo-dentais ([t], [d], [s], [z]), alveolares ([l], [r], [n]), palatais ([x], [j], [L], [ñ]) ou velares ([k], [g], [R]).

Função das Cordas Vocais

Se durante a produção das consoantes a corrente de ar produzir vibração das cordas vocais tem-se uma *consoante sonor*a; caso contrário, a consoante será *surd*a.

Papel das Cavidades Bucal e Nasal

Quando o ar sai exclusivamente pela boca, as consoantes são ditas *orais.* Quando o ar penetra nas fossas nasais pelo abaixamento da úvula, as consoantes são ditas *nasais* ([m], [n], [ñ]).

2.7 Conclusão

A fala humana é um sistema complexo. Apresentamos neste capítulo apenas as bases para que possamos entender o processo de produção da fala pelo nosso corpo. Produzir voz sintética também é complexo, conforme apresentaremos no próximo capítulo.

Capítulo 3 Sintetização de Voz

No sentido inverso ao do reconhecimento da fala, onde a entrada é a voz humana e a saída, uma ação do computador ou o texto digitado na tela, encontra-se a síntese de voz. A mesma recebe como entrada, um texto digitado ou em memória, e devolve, por meio de alto-falantes a leitura em voz alta do mesmo texto.

Este capítulo apresentará um breve histórico e evolução da sintetização de voz, apresentará também a sintetização de voz TTS (text-to-speech), ou seja, a reprodução de um texto falado (lido) pelo computador, suas fases e como se processa essa tecnologia.

3.1 Histórico

O primeiro estudo publicado sobre síntese de voz que se tem notícia foi em 1779. Este trabalho foi realizado por Christian Gottlieb Krazenstein da Academia Imperial de St. Petersburg. Esse estudioso inventou um instrumento que usava uma palheta vibrante e um constante fluxo de ar, como o mecanismo de um órgão musical. Mais tarde em 1791, Wolfgang von Kempelen de Viena criou uma máquina falante, a qual consistia de fole (pulmões), e uma palheta (cordas vocais). A forma da câmara de ressonância poderia ser alterada manualmente para gerar diferentes sons (timbres) de voz, da mesma maneira como a posição da língua, dos lábios, e maxilar alteram a forma do trato vocal. Na verdade, em outras palavras, essas máquinas foram baseadas no entendimento de algumas das características chaves do aparelho de produção da voz humana. A máquina de Kempelen, produziu mais que sons de vogais, chegando a produzir sentenças completas. Dava uma grande mão-de-obra fazer funcionar essa engenhoca, tanto é que dizem que trabalhadores especiais foram treinados por meses para usar a máquina, com o intuito de fazer ela gerar fala de forma inteligível.

Os estudos nessa área ficaram quase que paralisados durante décadas. Até que em 1939, Dudley apresentou uma outra máquina falante, chamada VODER (*Voice Operation Demonstrator*). Esse trabalho foi desenvolvido por Dudley nos laboratórios

Bell, pertencente a AT&T (American Telegraph & Telephony, USA) e foi apresentado na Feira Mundial na cidade de Nova Iorque. Esta máquina era constituída de dois geradores de sons independentes (ou excitação), um para sons periódicos (cordas vocais durante sons vozeados) e outro para ruído (turbulência causada pelas constrições no trato vocal). Um filtro operado manualmente imitava os efeitos do trato vocal. Para fins de demonstrações, pessoas especializadas foram treinadas para manusear o VODER. Devido a complexidade do equipamento, esse treinamento foi bastante longo. Este tinha um pedal de controle da freqüência operado pelo pé e 10 (dez) teclas para operar o sistema de controle das ressonâncias.

Os estudos com essas máquinas contribuíram para dar um melhor entendimento, mesmo que ainda elementar, ao nosso aparelho de produção da voz. Mesmo assim, esses dispositivos foram os primeiros passos importantes para construir sistemas que fossem capazes de produzir voz sintética convincente . Porque nos casos citados, eles produziram sons de voz inteligíveis e também porque eles foram baseados no conceito crítico de controle independente de uma fonte periódica (ou emissão de ruído) e a contribuição de uma variável de trato vocal. Essa idéia é a base da atual síntese de voz.

Em meados de 1965, o estudo de sistemas de síntese de voz a partir de texto teve um grande impulso, caracterizado por intensas pesquisas na área. Nessa época, havia ainda muitas dúvidas sobre a viabilidade dos sistemas e na qual iniciou-se o desenvolvimento das primeiras regras de conversão de texto para fonemas [ENDRES, 1983].

O início dos anos 70 marcou um novo período, com o aperfeiçoamento dos sintetizadores e dos algoritmos de conversão de texto para fonemas. Dois centros de pesquisas se destacaram nessa área, o "Bell Laboratories" pertencente à "American Telegraph & Telephony - AT&T", e o "Massachussets Institute of Technology -MIT" [UMEDA, 1976].

No início dos anos anos 80, o MIT apresentou o “MITalk”, um dos primeiros protótipos de sistemas de síntese de voz a partir de texto com vocabulário ilimitado [KAPLAN e LERNER, 1985].

Ainda nessa época surgiram os primeiros sistemas comerciais com vocabulário ilimitado, os quais vêm sendo continuamente aprimorados desde então.

Ainda em 1983, a "Infovox" colocou no mercado o sistema "SA 201/PC", capaz de sintetizar voz a partir de textos em Inglês, Francês, Espanhol, Alemão, Italiano, Suéco e Norueguês, desenvolvido a partir das pesquisas de Rolf Carlson no "Royal Institute of Technology of Stockholm" KLATT [1987].

Na segunda metade da década de 80, a "Berkeley Speech Technologies" apresentou o sistema "Text-to-Speech - T-T-S", originário das pesquisas de O'MALLEY (1990) na Universidade da Califórnia. Ainda nessa época, a AT&T lançou o sistema "Conversant", capaz de sintetizar voz a partir de textos em Inglês, Francês e Espanhol [AT & T do Brasil, 1994].

Em 1989, o "MITalk" produzia saída de voz a partir de textos em japonês e chinês [JAVKIN et al., 1989].

As pesquisas e estudos no Brasil, também já avançaram bastante. Tiveram seu início na Escola Politécnica da Universidade de São Paulo - EPUSP, através das pesquisas de CAMPOS (1980), sobre um sintetizador de voz para o idioma português, capaz de aceitar entradas na forma fonética da Língua Portuguesa [EGASHIRA, 1992].

No Instituto de Estudos da Linguagem da UNICAMP (Universidade de Campinas – SP), foi desenvolvido um mini léxico com cerca de 25.000 palavras, com informações sobre transcrição fonética, separação de sílabas e acentuação das palavras, para que possa ser utilizado em investigações e testes de regras de conversão de letras para fonemas [VIOLARO, 1993].

E em 1994, a Universidade Federal do Rio de Janeiro, desenvolveu o "Dosvox", para auxílio à deficientes visuais. Esse produto é formado por um conjunto de programas, tais como editor de texto e calculadora e é capaz de sintetizar voz a partir de texto utilizando um conjunto de regras de conversão de texto para fonemas [NCE – Núcleo de Computação Eletrônica, Rio de Janeiro, 1998].

3.2 Fases da Síntese de Voz

O processo de síntese de voz a partir de texto inicia-se com o processamento lingüístico para a determinação da estrutura fonológica das sentenças dos textos de entrada. Essa estrutura, constituída essencialmente pelos fonemas que formam as palavras, é analisada para a determinação dos parâmetros acústicos que serão utilizados posteriormente no controle do sintetizador de voz, o qual produzirá a fala. Este processo constitui uma tarefa ampla e complexa, e foi dividido em várias etapas por CROCHIERE e FLANAGAN (1990), conforme indica a Figura 1.

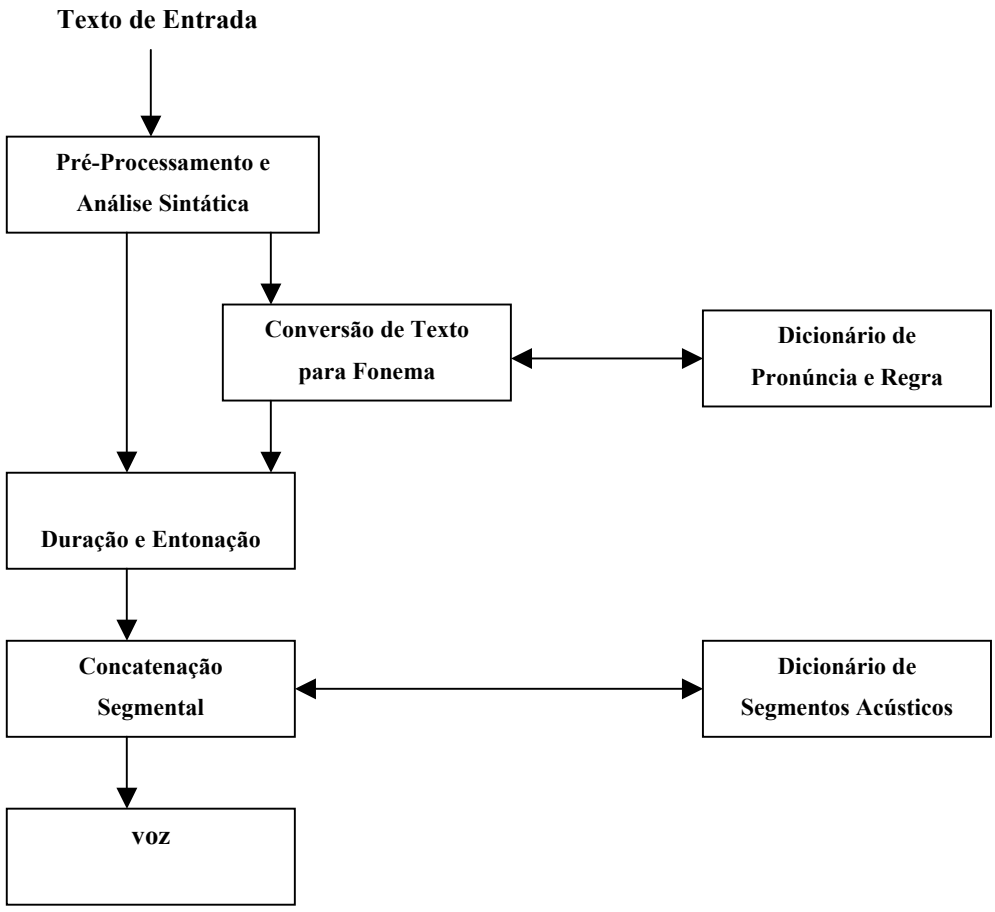

Figura 1 - Elementos de um sistema de síntese de voz a partir de texto.

[CROCHIERE e FLANAGAN, 1986]

Ao iniciar o processamento, o texto na sua forma original contendo abreviações, acrônimos, datas, caracteres não alfabéticos, acentuação e sinais de

pontuação, deve sofrer um pré-processamento para que seja utilizado nas fases seguintes.

Esse pré-processamento torna-se um tanto complexo, na medida em que podem ocorrer muitas ambigüidades. Por exemplo, um "**.**" (ponto) pode ser usado em uma abreviação ou no final de uma sentença. Um outro complicador são os números, estes constituem também uma séria dificuldade. Por exemplo, o número 26/10 pode ser uma data ou uma fração. Além disso, muitas abreviaturas comuns podem ter múltiplos sentidos, como nas seguintes sentenças: “O MM. Juiz, recebeu como homenagem uma placa com a espessura de 10MM. Que deveria ser lido: “O Meritíssimo juiz, recebeu como homenagem uma placa com a espessura de 10 milímetros”.

A análise sintática, mesmo que seja simples é necessária, para checar a pronúncia de determinadas palavras, como também sua correta entonação. Na maioria dos sistemas, as palavras são divididas em: *palavras de funçã*o, como as preposições e os artigos; e *palavras de conteúd*o, como substantivos e verbos. Normalmente, as palavras de conteúdo são pronunciadas com maior destaque, como na frase "Gato gosta ***de*** comer rato", onde a palavra destacada é pronunciada com menor ênfase em relação às demais. Conforme citam HIRSCHBERG et al. (1990).

A etapa seguinte à análise sintática, que consiste em determinar os fonemas correspondentes ao texto de entrada, tarefa executada normalmente através de um dicionário e de um conjunto de regras de conversão de letras para fonemas [ATAL e RABINER, 1986].

A conversão de uma palavra para a sua correspondente forma fonética inicia-se com a sua busca no dicionário, o qual contém informações fonéticas sobre algumas palavras. Caso não seja encontrada, a palavra deve ser submetida à aplicação de um conjunto de regras para a obtenção dos fonemas a partir de suas letras.

As informações fonéticas e sintáticas são utilizadas posteriormente para determinação da prosódia da mensagem, caracterizada por LAPORTE (1989) através das características de ritmo, acentuação, entonação e expressas como resultado combinado de parâmetros de duração, intensidade e "pitch". Cada sílaba em uma sentença tem uma duração específica, geralmente diferente das sílabas vizinhas, que determinam o ritmo da fala. A intensidade caracteriza o volume do som da fala, que em

uma sentença varia entre valores baixos e altos. O mesmo acontece com os valores de "pitch", os quais definem a entonação de uma sentença.

A partir dessas características, são determinados os parâmetros acústicos da sentença a ser pronunciada, eventualmente com base em um dicionário de segmentos acústicos e finalmente a voz é produzida por um sintetizador.

3.3 Conversão de texto para Fonemas ao Nível de Palavras

3.3.1 Conversão de Letra para Fonema

Várias tentativas iniciais foram feitas com a finalidade de prever a pronúncia de palavras a partir das letras que as compunham, com base na hipótese de que uma letra ou par de letras poderia ser convertido para o fonema apropriado caso fosse examinado o contexto, ou vizinhança, na qual a palavra estava inserida. Para cada letra, as regras deveriam ser ordenadas de modo que as primeiras tratariam dos casos mais complexos, e o último caso corresponderia à tradução fonética "default" [KLATT, 1987].

O melhor algoritmo de conversão de letra para fonema desenvolvido na década de 70 foi o algoritmo de Hunnicut, utilizado nos sistemas "MITalk" e no "DECtalk", na opinião de KLATT (1987). Esse algoritmo era bem mais complexo e executava em uma primeira fase a eliminação dos afixos (prefixo e sufixo) da palavra. Em seguida, fazia a conversão das consoantes e finalmente as vogais eram transcritas. Aproximadamente 15 prefixos e 50 sufixos eram detectados e posteriormente, aproximadamente 500 regras eram aplicadas. Em testes realizados com palavras extraídas aleatoriamente de um dicionário, esse algoritmo atingia em média um índice de 65% de palavras transcritas corretamente KLATT (1987).

Ficou constatado que somente a conversão de letras para fonemas utilizada nos algoritmos existentes, não era possível fazer uma transcrição fonética correta de todas as palavras de uma Língua. A melhor saída foi a introdução nesses algoritmos de um *dicionário de exceçõe*s [HIRSCHBERG, 1990].

3.3.2 Dicionário de Exceções

Levando-se em consideração que apenas a conversão de letras para fonemas não é suficiente para uma transcrição fonética absolutamente correta de todas as palavras de uma Língua, principalmente na nossa Língua Portuguesa onde temos que levar em conta os acentos que utilizamos. A alternativa é a utilização de um *dicionário de exceçõe*s, que contenha palavras que falhem a essas regras. Como no exemplo citado por ALLEN (1976), onde ele diz que a letra "f" sempre é pronunciada como /f/, exceto em *of* (/ v/), onde soa como se fosse "v".

A vantagem da elaboração de um *dicionário de exceções* advém do fato de que um pequeno número de palavras repete-se inúmeras vezes em um texto aleatório [KLATT, 1987]. HIRSCHBERG et al. (1990) citam que um dicionário com apenas 150 palavras chega a cobrir 50% das palavras de um texto.

No entanto, a utilização de um dicionário contendo todas as palavras de uma Língua é inviável. Devido ao grande número de palavras que deveriam ser armazenadas, a velocidade de acesso a esse banco tornar-se-ia muito lenta [KLATT, 1987].

Essas considerações sugerem que um sistema híbrido, contendo um conjunto de regras de conversão de letras para fonemas e um dicionário de exceções, é uma solução adequada à conversão de textos em fonemas, pois um dicionário de exceções de tamanho moderado pode reduzir deficiências de um conjunto de regras de conversão de letras para fonemas [CHBANE, 1994].

3.4 Conversão de Texto para Fonemas no Idioma Português

A conversão de texto para fonemas no idioma Português, de maneira análoga a que ocorre na Língua Inglesa, deve ser executada em vários passos para que possa ser bem sucedida.

Partindo-se do texto de entrada, é necessária a execução de pré-processamento para eliminar abreviaturas, siglas, números e caracteres não alfabéticos, expandindo-os para as correspondentes palavras. Da mesma forma que acontece na língua inglesa, podem ocorrer ambigüidades, como no caso dos caracteres "1" e "2", que podem ser escritos respectivamente como *um* ou *um*a, e *dois* ou *duas* [EGASHIRA, 1992].

A análise sintática é necessária não apenas para determinar a correta entonação do "pitch", mas também, para determinar a correta pronúncia de palavras como *g[ó]st*o, verbo, e *g[o]st*o, substantivo.

Na Língua Portuguesa, as palavras homógrafas heterófonas (palavras com a mesma grafia porém com pronúncias diferentes), constituíram cerca de 3% do corpus de teste, nos estudos realizados por VIANA et al. (1991).

É indispensável o uso de um dicionário de exceções que contenha a transcrição de palavras que falhem a aplicação das regras. Como por exemplo, a letra "x", que pode ser associada a quatro fonemas diferentes, [x] na palavra *xal*e; [ks] na palavra *fix*o; [s] na palavra *texto* e [z] na palavra *exame*. A Tabela 2 apresenta algumas regras para a conversão fonética dos fonemas [x], [z] e [s], porém, deve-se notar mais uma vez que essas regras não serão suficiente para atender a todos os casos, devendo-se utilizar um dicionário de exceções.

Fonema	Ocorrência	Exemplo
[x]	- Início de palavra - Depois de "n" - Depois de "ai", "ei" e "ou"	xícara, xarope enxame, enxofre caixa, eixo, frouxo
[z]	- Palavras iniciadas com "ex" seguido de vogal	exame, exercício, exótico
[s]	- Seguido de consoante	texto, sexto

TABELA 2. Algumas regras para conversão fonética da letra "x" [ALIANDRO, 1974]

A nasalização das vogais ocorre em três situações distintas, quando a vogal é acentuada seguida de consoante nasal ([m] ou [n]); sempre que a vogal for seguida por consoante nasal e outra consoante, e quando a vogal estiver antes de consoante nasal em final de vocábulo. Essas regras como também alguns aspectos referentes a nasalidade das vogais foram destacados por CALLOU e LEITE (1990), ressaltando dificuldades na transcrição de vogais nasais quando estas não estão marcadas com o "~" (acento til).

Entretanto, existem exceções a essa regra, que não é capaz de distinguir *c[ã]minh*a, substantivo, de *c[a]minh*a, verbo. Nesses casos, as duas alternativas devem estar presentes no dicionário de exceções e a seleção da transcrição adequada dependente de análise sintática do texto de entrada.

3.5 Conclusão

A tecnologia de síntese de voz TTS (text-to-speech), tem evoluído bastante nos últimos anos tornando-se viável a sua aplicação à qualquer idioma, mesmo levando em consideração alguns aspectos (problemas) de grafia e fonética utilizados nos mesmos, demonstrados nesse capítulo.

No próximo capítulo, descreveremos como os sintetizadores trabalham as estruturas aqui tratadas, para produzirem voz com qualidade.

Capítulo 4 Sintetizadores de Voz

As técnicas básicas utilizadas para sintetizar a voz humana são duas: uma baseada em regras que define parâmetros da voz e fala de uma pessoa, e outra baseada na concatenação (encadeamento) de segmentos de voz, previamente gravados por um locutor. Ambas geram fonemas sintetizados que, combinados com outros parâmetros, são capazes de formar palavras e frases.

Este capítulo apresenta os sintetizadores de voz mais utilizados nos sistemas aplicativos comerciais disponíveis no mercado hoje, assim como também os métodos, técnicas e algoritmos utilizados nos mesmos, fazendo uma descrição clara e objetiva de como eles sintetizam a voz a partir de texto.

4.1 Métodos, Técnicas e Algoritmos para Síntese de Voz

Fala sintetizada pode ser produzida através de vários métodos diferentes. Todos eles apresentam alguns benefícios e deficiências. Os métodos normalmente são classificados conforme abaixo:

- Síntese Articulatória, que tenta modelar o sistema de produção de fala humana diretamente.
- Síntese de Formantes, que modela as freqüências do sinal de som da fala ou transfere a função de área vocal baseado em fonte-filtro-modelo.
- Síntese de Concatenação, que usa amostras de pré-gravação, de comprimentos diferentes de fala natural.
- Síntese LPC (*Linear Predictive Coding*), que sintetiza o sinal de voz usando o resíduo para criar um sinal de fonte e usa os formantes para criar um filtro.

O método de Formantes e o método de Concatenação são geralmente os mais usados em sistemas de síntese de voz atualmente. A síntese de Formantes foi dominante durante algum tempo atrás, mas hoje o método de Concatenação está ficando cada vez mais popular. O método Articulatório ainda é muito complicado para implementações de alta qualidade, mas pode surgir como um método potencial no futuro [LEMMETTY, 1999].

4.2 Síntese Articulatória

A síntese Articulatória tenta modelar os órgãos vocais humanos tão perfeitamente quanto possíveis, assim é potencialmente o método promissor para produzir fala sintética de alta qualidade. Ela envolve tipicamente modelos do órgão articulador humano e cordas vocais.

Os articuladores normalmente são modelados como uma área situada entre a glote e a boca. O primeiro modelo de articulador foi baseado em uma área vocal onde funciona a região da laringe e lábios, para cada segmento fonético [KLATT, 1987]. Para síntese articulatória baseada em regras, os parâmetros de controle podem ser por exemplo: abertura dos lábios, protuberância dos lábios, altura da ponta da língua, posição da ponta da língua, altura da língua, posição de língua e abertura do vélum. Fonetização ou parâmetros de excitação podem ser abertura da glote, tensão das cordas vocais e pressão pulmonar [KRÖGER, 1992].

O dados para modelos de articulação normalmente é derivado de análise de Radiografia de fala natural. Porém, estes dados são normalmente radiografados somente em 2-D, quando a real área vocal é naturalmente em 3-D. Assim, a síntese articulatória baseada em regra é muito difícil de se implementar, devido à indisponibilidade de dados suficientes dos movimentos dos órgãos articuladores durante fala. Outra deficiência com síntese articulatória é que os dados de uma Radiografia não caracterizam as massas ou graus de liberdade dos órgãos articuladores [KLATT, 1987]. Além disso, os movimentos da língua são tão complicados que é quase impossível fazer uma modelagem precisa.

Este método é consideravelmente complexo em relação aos outros métodos comuns [KRÖGER, 1992], [RAHIM, 1993]. A síntese articulatória é usada raramente

em sistemas de síntese de voz atualmente, mas desde que os métodos de análise em desenvolvimento e os recursos de computação estão evoluindo rapidamente, poderá ser um método de síntese potencial no futuro.

4.3 Síntese de Formantes

Provavelmente o método de síntese mais utilizado durante décadas passadas, foi o de síntese de formantes, que está baseado em fonte-filtro-modelo. Ele possui duas estruturas básicas em geral que são: paralelo e cascata. Mas para um melhor desempenho a combinação destes é normalmente usado. Síntese de formantes também provê um número infinito de sons, tornando-se assim mais flexível que o método de concatenação.

Geralmente são exigidos pelo menos três formantes para produzir uma fala inteligível e até cinco formantes para produzir fala de alta qualidade. Cada formante normalmente é modelado como um ressonador bipolar que habilita ambas freqüências formantes (freqüência de pólo-par) e sua largura de banda a ser especificada [DONOVAN, 1996].

Síntese de formantes está baseada em um conjunto de regras determinadas por parâmetros necessários para sintetizar uma expressão vocal desejada [ALLEN, 1987]. Os parâmetros de entrada podem ser, por exemplo, o quociente do tempo de abertura da glote em relação à duração do período total [HOLMES, 1990].

A estrutura do sintetizador de formantes em cascata (figura 4.1), consiste em ressonadores de passsagem de banda conectados em série e a saída de cada ressonador, é ligada à entrada do seguinte. A estrutura em cascata precisa somente das freqüências dos formantes como informação de controle. A vantagem principal da estrutura em cascata é que a amplitude relativa dos formantes para as vogais não precisam de controles individuais [ALLEN, 1987].

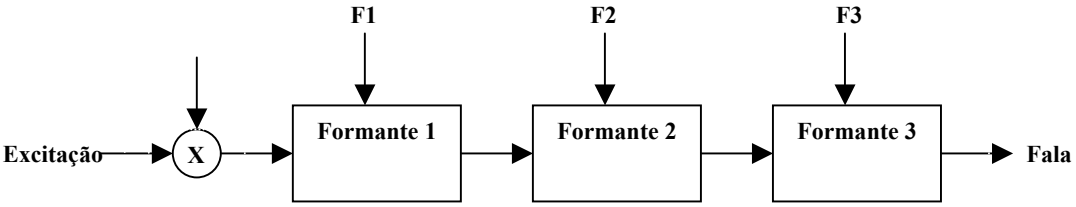

Figura 2. Estrutura básica do sintetizador de formantes em cascata [Allen, 1987]

A estrutura em cascata foi considerada melhor para sons não-nasais e também porque precisa de menos informação de controle do que a estrutura paralela, é então um instrumento mais simples. Porém, com o modelo em cascata a geração de sons fricativos e de sons plosivos (ver seção 2.2), torna-se um problema [ALLEN, 1987].

O sintetizador de formantes em paralelo (figura 3) consiste em ressonadores conectados em paralelo. Às vezes ressonadores extras para sons nasais são usados. O sinal de excitação é aplicado a todos os formantes simultaneamente e suas saídas são somadas. As saídas adjacentes dos ressonadores devem ser somadas em fase oposta, para evitar zeros não desejados ou anti-ressonância na freqüência de resposta [O'SAUGHNESSY, 1987]. A estrutura paralela habilita o controle de largura de banda para cada formante individualmente, por isso, necessita também de mais informação de controle.

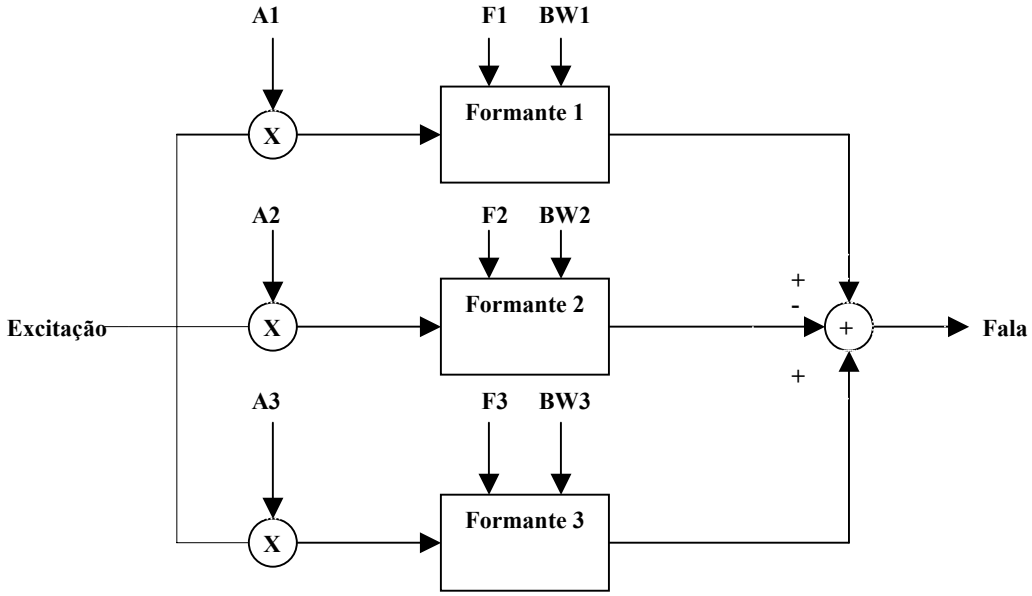

Figura 3. Estrutura básica do sintetizador de formantes em paralelo [O'Saughnessy, 1987]

A estrutura do sintetizador de formantes em paralelo foi considerada melhor para sons nasais, sons fricativos e sons de consoantes, exceto para algumas vogais. Só que estas podem ser normalmente modeladas com a estrutura em cascata.

Houve controvérsia sobre essas duas estruturas em relação a qualidade e características dessas. É fácil de ver que utilizando apenas um método é difícil alcançar boa qualidade, assim alguns esforços foram feitos para melhorar e combinar estes modelos básicos [KLATT, 1980].

4.4 Síntese de Concatenação

A pré-gravação conectando expressões vocais naturais, provavelmente é o modo mais fácil para produzir fala sintética soando de forma inteligível e natural. Porém, normalmente os sintetizadores de concatenação são limitados a um orador e uma voz, e normalmente requerem mais capacidade de memória que outros métodos.

Um dos aspectos mais importantes em síntese de concatenação é determinar o correto comprimento da unidade de fala. A seleção do comprimento da unidade, é normalmente um intercâmbio entre unidades mais longas e unidades mais curtas. Com unidades mais longas, obtém-se alta naturalidade, menos concatenação são efetuadas e

bom controle de coarticulação é alcançado; mas, quanto mais unidades são exigidas, mais memória também é consumida. Com unidades mais curtas, é consumida menos memória, mas a concatenação destas, exige procedimentos mais difíceis e complexos. Nos sistemas atuais essas unidades são normalmente palavras, sílabas, fonemas, difones e às vezes trifones.

A palavra é talvez a unidade mais natural para texto escrito, como também para alguns sistemas de mensagens com vocabulário muito limitado. Concatenação de palavras é relativamente fácil de se executar e a sua pronúncia e entonação podem ser armazenados com essas unidades também. Porém, há uma grande diferença com palavras faladas isoladamente, pois o som destas pode variar quando são faladas dentro de uma frase [ALLEN, 1987]. Porque há centenas de milhares de palavras diferentes e nomes próprios em cada idioma, palavra não é uma unidade satisfatória para qualquer bom sistema de TTS irrestrito.

O número de sílabas diferentes em cada idioma é consideravelmente menor que o número de palavras, mas o tamanho do banco de dados dessas unidades, normalmente, ainda é muito grande para sistemas de TTS. Por exemplo, há aproximadamente 10.000 (dez mil) sílabas em inglês. Ao contrário das palavras, o efeito de pronúncia e entonação não é armazenado em unidades; então usar sílabas como uma unidade básica não é muito razoável; pois não há como controlar efeitos de prosódia em cima da oração. Até o momento, não existe muito fundamento no uso de palavra ou sílaba nos sistemas de TTS. Os atuais sistemas de síntese estão baseados principalmente no uso de fonemas, difones, trifones ou uma boa combinação destes.

Os fonemas provavelmente são as unidades mais usadas em síntese de fala, porque eles são a representação lingüística normal da fala. O montante de unidades básicas normalmente está entre 40 (quarenta) e 50 (cinqüenta), que são claramente bem menor se comparados a outras unidades [ALLEN, 1987]. O uso de fonemas dá uma maior flexibilidade nos sistemas baseados em regras. Porém, alguns sons que não têm uma posição firmemente designada, como os sons plosivos por exemplo; são mais difíceis de sintetizar. A articulação destes também deve ser formulada como regra. Fonemas às vezes são usados como entrada em sintetizadores baseados em difones por exemplo.

Difones (ou par de sons) é definido como sendo o ponto central da parte fixa de um som de um fonema ao ponto central do seguinte, assim eles contêm as transições entre os sons adjacentes. Isso significa que o ponto de concatenação estará na região de estado mais fixa do sinal, o que reduz a distorção de pontos de concatenação. Outra vantagem com difones é que os efeitos de entonação e pronúncia não necessitam mais serem formulados como regras. O número de unidades normalmente é de 1500 (hum mil e quinhentos) a 2000 (dois mil), que certamente ainda exige um bom consumo de memória, tornando a coleta dos dados mais difícil comparado a fonemas. Porém, a quantidade de dados ainda é tolerável e com outras vantagens, difone é uma unidade muito satisfatória para síntese de voz baseada em TTS. O número de difones pode ainda ser reduzido invertendo transições simétricas, como por exemplo: /us/ por /sa/.

São usadas raramente unidades segmentárias mais longas, como trifones ou tetrafones. Trifones são parecidos com difones, mas contém um fonema entre os pontos de estado fixo (meio fonema - fonema - meio fonema). Em outras palavra, um trifone é um fonema com uma esquerda e direita específica dentro de um contexto. Para o inglês, são requeridos mais de 10.000 (dez mil) unidades [HUANG, 1997].

Concatenar as amostras de fala natural normalmente é muito demorado. Porém, algum trabalho pode ser feito automaticamente fazendo um pré-processamento adequado no texto de entrada. A implementação de regras para selecionar amostras corretas para concatenação também deve ser feita muito cuidadosamente. [HON, 1998].

Há vários problemas em síntese de concatenação comparada a outros métodos.

- Distorção de descontinuidades em pontos de concatenação que podem ser reduzidos usando difones ou alguns métodos especiais para suavisar o sinal.
- Exigências de memória normalmente muito alta, especialmente quando unidades de concatenação longas são usadas, como sílabas ou palavras.
- A recuperação de um registro no banco de dados de amostras de fala é normalmente demorado.

Alguns dos problemas podem ser resolvidos com os métodos descritos a seguir, além do que, o uso de métodos de concatenação está aumentando devido a alta capacidade de processamento dos computadores atuais [DONOVAN, 1996].

4.4.1 PSOLA

O método de síntese PSOLA (*Pitch Synchronous Overlap and Add*), foi desenvolvido originalmente na França Telecom (CNET - *Centre National d'Etudes Télécommunications*). Permite a concatenação de amostras de fala pré-gravadas e provê bem o controle de duração e *pitch*, assim é usado em alguns sistemas comerciais de síntese de voz, como o ProVerbe e HADIFIX [DONOVAN, 1996].

Há várias versões do método PSOLA e todos eles trabalham em essência do mesmo modo. A versão TD-PSOLA (*Time Domain - Pitch Synchronous Overlap and Add*), é geralmente a mais usada devido a sua eficiência de processamento [KORTEKAAS, 1997].

O algoritmo básico consiste em três passos: 1) O passo de análise, onde o sinal original da fala primeiramente é dividido e separado em períodos curtos; 2) O passo de modificação de cada sinal de análise para sinal de síntese; 3) O passo de síntese, onde estes segmentos são recombinados através de sobreposição e concatenação ([CHARPENTIER, 1989], [VALBRET, 1991]).

4.4.2 MBROLA

O projeto do sintetizador MBROLA (*Multi Band Resynthesis OverLap Add*) foi criado pelo Laboratório de TCTS (*Théorie des Circuits et de Traitement du Signal*) na Faculdade Politécnica de Mons, Bélgica. A meta principal do projeto é desenvolver um sintetizador de voz multilingual, para propósitos não comerciais e não militares, tendo como um dos objetivos o fomento à pesquisa acadêmica, especialmente na geração de prosódias (pronúncias adequadas, ver seção 3.2). MBROLA é um método similar ao PSOLA (*Pitch Synchronous Overlap and Add*), mas ficou sendo chamado de MBROLA, devido ao fato de que PSOLA foi patenteado como uma marca registrada da CNET (*Centre National d'Etudes Télécommunications*). (DUTOITt et al.1999)

O sintetizador MBROLA versão 3.01d está baseado em concatenação de difones. Ele possui uma lista de fonemas com informações de prosódia como contribuição e produz amostras de fala de 16 bits (linear), e a freqüência de amostragem do banco de dados de difones.

Os bancos de dados de difones estão atualmente disponíveis para uma série de linguagens, entre elas: inglês americano/britânico, português brasileiro, holandês, francês, alemão, romano e espanhol, com timbre de voz masculino e feminino. Para vários outros idiomas, tal como o Estoniano; também estão sendo desenvolvidos banco de dados de difones.

O dados de entrada requeridos pelo MBROLA contém um nome de fonema, uma duração em milissegundos, um *pitch* e a freqüência. Por exemplo, a entrada "_ 51 25 114" indica que o sintetizador deverá produzir um silêncio (_) de 51 ms, e pôr um *pitch* de 114 Hz a 25% de 51ms.

4.5 Síntese LPC

O método LPC (*Linear Predictive Coding*) é uma das técnicas de síntese de voz muito poderosa. Ele é um dos métodos mais úteis para codificar a voz com boa qualidade a uma baixa taxa de bit. O LPC provê estimativas extremamente precisas de parâmetros de voz e é relativamente eficiente no processamento de síntese de voz.

O LPC analisa o sinal de fala formantes, removendo os efeitos do sinal e avaliando a intensidade e freqüência do sinal restante. O processo de remover os formantes é chamado de filtro inverso e o sinal restante é chamado de resíduo.

O LPC sintetiza o sinal de voz invertendo o processo: usa o resíduo para criar um sinal de fonte e usa os formantes para criar um filtro (como se fosse uma espécie de tubo), a fonte (sinal) percorre pelo filtro, gerando como resultado a voz (fala).

Como sinais de voz variam com o tempo, este processo é terminado em pedaços curtos de sinal de voz, que são chamados de "*frames*" (quadros). Normalmente com uma taxa de 30 (trinta) a 50 (cinqüenta) *frames* por segundo é possível produzir um discurso inteligível com boa compreensão.

O problema básico do sistema de LPC é determinar os formantes do sinal de voz. A solução básica é uma equação diferencial que expressa cada amostra do sinal como uma combinação linear de amostras prévias. Essa equação é chamada de "predição linear" e é por isso que é chamado de Codificação de Predição Linear.

Os coeficientes da equação diferencial (os coeficientes de predição) caracterizam os formantes, assim o sistema de LPC precisa calcular estes coeficientes. A estimativa é terminada minimizando o erro entre o sinal previsto e o sinal atual. Este é um problema direto em princípio. Na prática, envolve (1) a computação de uma matriz de valores de coeficientes, e (2) a solução de um conjunto de equações lineares. Vários métodos (auto-correlação, covariação, formulação recursiva) podem ser usados para assegurar convergência a uma solução eficiente em computação.

Pode parecer surpreendente que o sinal possa ser caracterizado por uma predição linear simples. Quando isso ocorre, para que funcione, o tubo não deve ter nenhuma matriz lateral. (Em condições matemáticas, matrizes laterais introduzem zeros que requerem equações muito mais complexas.)

Para vogais ordinárias, a área vocal é representada bem por um único tubo. Porém, para sons nasais, a cavidade do nariz forma uma matriz lateral. Então, sons teoricamente nasais requerem um algoritmo diferente e mais complicado. Na prática, esta diferença é ignorada e será codificada como resíduo.

Se os coeficientes de predição são precisos e tudo funciona corretamente, o sinal de voz pode ser filtrado de forma inversa e o resultado será uma fonte pura. Para esse sinal, é bastante fácil de extrair a freqüência e a amplitude e os codificar.

Porém, algumas consoantes são produzidas com corrente de ar turbulenta, gerando como resultado um som fricativo. Felizmente, a equação de predição não se preocupa se a fonte é periódica ou caótica (não-periódica).

Isto significa que para cada *frame* (quadro), o codificador LPC tem que decidir se a fonte é boa ou ruim; se for boa, calcula a freqüência; em ambos os casos, calcula a intensidade e codifica a informação de forma que o decodificador pode desfazer todos estes passos.

Infelizmente, as coisas não são tão simples assim. Uma das razões é que quando falamos, emitimos uma combinação de zumbidos e assobios (por exemplo, as consoantes iniciais em "this zoo" e a consoante mediana em "azure"). A voz não parecerá ser reproduzida corretamente por um codificador de LPC simples.

Outro problema é que, inevitavelmente, qualquer inexatidão na estimação dos meios dos formantes, mais informação de voz é partida no resíduo. Há aspectos de sons nasais que não são modelados pelo LPC (conforme discutido acima), por exemplo, terminará em resíduo. Há outros aspectos do som de voz que o LPC não modela; Matrizes laterais introduzidas pelas posições da língua de algumas consoantes e ressonâncias na traquéia são alguns exemplos.

Então, o resíduo contém informações importantes a respeito de como a fala deveria soar. Então síntese de LPC sem estas informações resultará em fala de má qualidade. Para melhorar a qualidade, deveríamos também enviar o sinal de resíduo, e o resultado da síntese soaria melhor. Infelizmente, se utilizarmos desta técnica, a compressão do sinal original da fala juntamente com os vários bits do resíduo, poderá se tornar inviável.

Foram feitas várias tentativas de codificar o sinal de resíduo de um modo eficiente, visando sons de voz de melhor qualidade sem aumentar muito a taxa de bits. Os métodos mais prósperos usam um codebook [CAMPOS, 1996], onde os resíduos são colocados em uma tabela fixa de acordo com o seu respectivo desenho. Na operação, o analisador compara o resíduo com todas as entradas no codebook, escolhe a entrada que mais se aproxima, e envia o código correspondente para aquela entrada. O sintetizador recebe este código busca o seu correspondente no codebook e utiliza-o para excitar o filtro de formantes. Esquemas deste tipo são chamados de CELP (*Code Excited Linear Prediction*).

Para o CELP ter bons resultados, o codebook deve ser grande o bastante para incluir todos os vários tipos de resíduos. Mas se o codebook for muito grande, o tempo de varredura na tabela para procurar o código do resíduo correspondente, será muito grande também. O maior problema é aquele em que é requerido um código diferente para cada freqüência da fonte (pitch da voz), o que implicaria em construir um codebook extremamente grande.

Este problema pode ser resolvido usando dois codebooks pequeno em vez de um muito grande. Um codebook será fixo e conterá a maioria códigos para representar um pitch de resíduo. O outro codebook será adaptável, inicializando-se com vazio e será preenchido durante as operações, com cópias dos resíduos remanescentes. Assim, o

codebook adaptável terá uma lista variável de registros para pesquisa, com a quantidade de pitch necessários.

Com o algoritmo CELP, é possível produzir fala natural de boa qualidade, a uma taxa de 4800 (quatro mil e oitocentas) frames (quadros) por segundo (CAMPOS et al., 1996).

4.6 Outros Métodos e Técnicas

Vários outros métodos e experiências, foram feitos para melhorar a qualidade de fala sintética. Foram estudadas e descritas variações e combinações de vários métodos, mas ainda não há um consenso sobre qual é o melhor método. Fala sintetizada pode ser manipulada naturalmente como fala normal através de processamento de algoritmos. Por exemplo, somando-se um pouco mais de eco ao sinal de voz, podemos produzir fala mais agradável. Porém, este adicionamento pode aumentar facilmente a carga de processamento computacional do sistema.

Foram feitas algumas experiências para mostrar o uso de uma combinação dos métodos de síntese básicos, porque diferentes métodos apresentaram sucesso diferentes na produção de fonemas individuais. Síntese de Tempo-Domínio pode produzir segmentos de fala natural de alta qualidade, mas a combinação de vários segmentos de fala sintetizada pode ser descontínua nos limites dos segmentos e se uma variação de aumento de faixa da freqüência fundamental (*F0*) é requerida, a complexidade aumentará mais ainda. Por outro lado, síntese de formantes produz fala mais homogênea e permite um bom controle da freqüência fundamental (*F0*), mais o timbre da voz soa mais sintético. Esta aproximação conduz a um sistema híbrido que combina a síntese de Tempo-Domínio e síntese de Freqüência-Domínio. A idéia básica de um sistema híbrido é mostrada na figura 4 [FRIES, 1993].

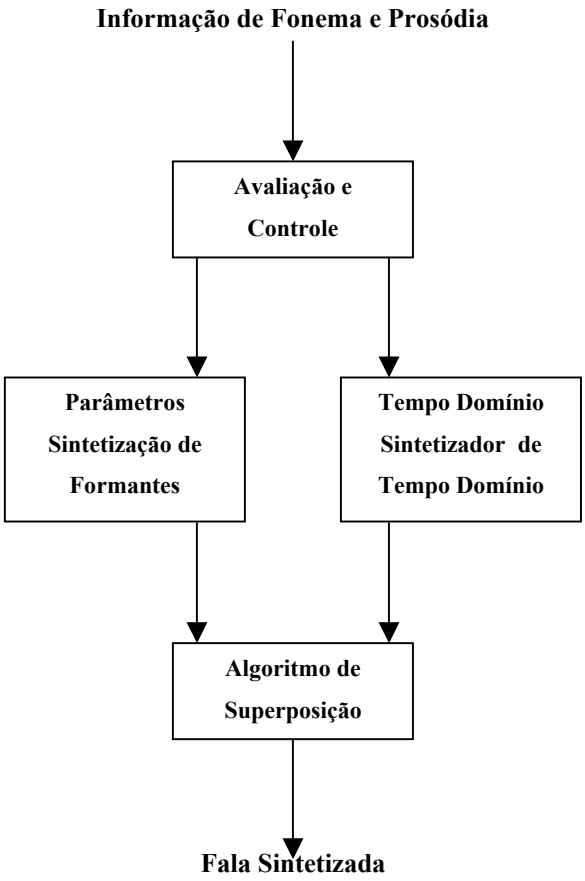

Figura 4. Idéia Básica do sistema de síntese híbrido [FRIES, 1993].

Vários métodos e técnicas para determinar os parâmetros de controle de um sintetizador podem ser utilizados. A inteligência artificial (IA) criou um método chamado Redes Neurais Artificiais (ANN), que foi usado para controlar parâmetros de síntese, tais como duração, ganho e freqüência fundamental (*F0*) [SCORDILIS, 1989] e [KARJALAINEN, 1998]. Redes Neurais também é utilizado na síntese de fala, onde é utilizado um conjunto de elementos ou nodos, análogos aos neurônios do cérebro. Estes elementos ou nodos são processados interconectados em uma rede, que pode identificar os dados expostos de acordo com seus padrões.

4.7 Conclusão

Existem diferentes métodos para se produzir a fala sintetizada, dentre os quais pode-se destacar: Síntese Articulatória, Síntese de Formantes, Síntese de Concatenação, Síntese LPC, Síntese Híbrida e Síntese por Redes Neurais Artificiais (ANN). Todos eles apresentam alguns benefícios e deficiências de forma que ainda não há um consenso de qual é o melhor dos métodos.

O método de Formantes e o método de Concatenação são geralmente os mais usados em sistemas de síntese de voz atualmente.O método Articulatório ainda é muito complicado para implementações de alta qualidade, mas pode surgir como um método potencial no futuro.

O método LPC é uma técnica de produção de voz muito poderosa, podendo gerar fala de alta qualidade com uma baixa transmissão de taxa de bits.

O método Híbrido também pode gerar voz com alta qualidade, mas dependendo das combinações dos seguimentos de fala, isto pode não ser verdadeiro.

O método por Redes Neurais Artificiais (ANN) é também bastante promissor, pois ele usa um conjunto de elementos ou nodos, análogos aos neurônios do cérebro humano.

No próximo capítulo apresento a minha análise crítica dos assuntos até aqui tratado em especial dos métodos de sintetização de voz.

Capítulo 5 Análise Comparativa dos Métodos de Síntese de Voz

A fala é a principal maneira de comunicação entre as pessoas. Ainda hoje, na nossa sociedade, as pessoas com deficiência vocal (pessoas mudas) enfrentam certas dificuldades no seu dia-a-dia. Essas mesmas dificuldades também fazem parte da vida dos deficientes visuais. A síntese de voz ***text-to-speech*** (TTS) tornou-se uma grande aliada na melhoria da qualidade de vida dessas pessoas. Nos diferentes métodos para se produzir Fala Sintetizada os quais são descritos nesse estudo, todos eles apresentam alguns benefícios e deficiências e como já frisamos, ainda não há um consenso de qual é o melhor entre eles. Em todo caso, método de Concatenação (utilizando difones) e o método de Formantes são geralmente os mais usados em sistemas de síntese de voz comercializados atualmente.

Este capítulo apresenta uma análise comparativa entre os métodos de Concatenação e Formantes, utilizados na síntese TTS. Isto é feito através de análises teóricas e experimentação. Na análise teórica foram analisados os parâmetros de: complexidade, tamanho do dicionário de fonemas, versatilidade para implantação de idiomas e processamento. Na experimentação foi analisado o parâmetro de tempo de respostas, que avaliou o tempo entre a solicitação da síntese de voz e da apresentação da voz propriamente dita. E ainda na experimentação, foi realizada uma pesquisa de campo e a partir desta efetivou-se uma avaliação da qualidade de voz, utilizando softwares que implementam os métodos analisados. A análise de voz foi feita utilizando-se a pré-gravação das sentenças (frases) e fonemas baseados nas técnicas de fonoaudiologia para avaliação qualitativa da voz.

5.1 Métodos de Concatenação e de Formantes

Estes métodos, devido a sua grande utilização e aceitação na implementação de softwares TTS, foram intensamente estudados e analisados durante a elaboração desta dissertação. A seguir é feito um resumo dos métodos de Formantes e de Concatenação.

5.1.1 Concatenação

Uma forma de produzir o sinal de fala sintetizado consiste em definir as trajetórias dos parâmetros pela concatenação de seqüências de valores extraídos de pedaços de fala natural. Neste caso é necessário definir qual a dimensão das unidades a concatenar. Unidades demasiadamente longas obrigam a um extenso inventário que englobe todas as seqüências possíveis. Por outro lado, uma vez que não existem regras para incorporar os efeitos de coarticulação, as unidades têm de ser suficientemente longas para capturar esses efeitos [OLIVEIRA, 1996].

Sendo o centro do segmento fonético a sua zona mais estável, este então deve ser um bom candidato a início de unidade. Esta propriedade levou à sugestão do *difone* como unidade mínima de um sistema de síntese [PETERSON et al. 1958], definido como o segmento acústico deste o centro de um segmento fonético até ao centro do segmento seguinte. O número de difones necessários para cobrir todas as combinações de segmentos seria assim igual ao quadrado do número de segmentos fonéticos da língua. No entanto, nem todos os pares de segmentos ocorrem e o número de difones pode ser substancialmente reduzido. Os fenômenos de coarticulação estendem-se muitas vezes para além do segmento seguinte e por esse motivo é comum a utilização de algumas unidades mais longas, como o trifone, meias-sílabas, sílabas, ou mesmo palavras inteiras. Outro fator que faz aumentar a dimensão do inventário fonético de um sistema de síntese é a inclusão de variações alofônicas dos segmentos fonéticos.

Uma das grandes vantagens dos sistemas de concatenação é de que as unidades podem ser extraídas diretamente da fala natural sem ser necessário conhecer e modelar muitos dos detalhes com relevância perceptual. Esta característica possibilita um mais rápido desenvolvimento e alteração do sistema, sendo assim mais fácil criar o sintetizador. Por outro lado, temos como desvantagens a distorção de descontinuidades em pontos de concatenação, que podem ser reduzidos usando difones ou alguns

métodos especiais para suavizar o sinal. Tem ainda o problema do uso de unidades de concatenação longas, como palavras por exemplo, as quais exigem muita memória, dificultando assim o processamento.

5.1.2 Formantes

Os múltiplos tipos de sons emitidos por um tubo acústico como o trato vocal por exemplo, logo cedo sugeriram a sua modelação por circuitos ressonadores de segunda ordem. A forma de associação destes filtros dividiu inicialmente os sintetizadores em modelos em cascata (FANT, 1960), onde a amplitude das freqüências formantes é imposta pela relação entre as diversas freqüências e larguras, e modelos em paralelo (HOLMES, 1973), com controle individual da amplitude de cada formante. O modelo cascata/paralelo proposto por Klatt (KLATT, 1980) veio permitir o uso simultâneo dos dois modelos. Este modelo caracteriza-se por conter duas representações do trato vocal: uma usando uma associação em cascata de cinco filtros de segunda ordem representando as ressonâncias e com controle de energia à entrada, e outra usando a associação em paralelo de seis filtros idênticos aos primeiros, mas com controle individual das amplitudes. O modelo em cascata é normalmente utilizado para sintetizar sons vozeados, pois modela corretamente um tubo acústico excitado num dos extremos e onde a amplitude de cada ressonância é imposta pelas freqüências e larguras de banda de todas as formantes. Nos sons não vozeados, em que a excitação do trato vocal pode ter diferentes localizações, é mais conveniente o controle individual da amplitude de cada ressonância. Neste caso, acrescentou-se uma ressonância adicional para sintetizar o ruído de alta freqüência presente em certas consoantes alveolares [s ,z].

O efeito da cavidade nasal é modelado no modelo em cascata com um filtro com um pólo e um zero. Quando o som não é nasalizado, o zero é colocado de forma a cancelar o pólo, ou seja, na síntese de sons nasais, aumenta-se a freqüência do zero de forma que simultaneamente expõe-se o pólo e reduz a amplitude da primeira formante. No modelo paralelo é suficiente o controle independente das amplitudes das ressonâncias associadas ao pólo nasal e à primeira formante [OLIVEIRA, 1996].

Nos sons vozeados, o trato vocal é excitado com um modelo paramétrico do fluxo de ar na glote, baseado no modelo polinomial de Rosenberg (ROSENBERG, 1971), e que é repetido ao ritmo da freqüência fundamental pretendida. A turbulência

produzida na glote e a aspiração são modeladas por ruído de baixa freqüência a $-6dB$. A excitação para os sons fricativos é também feita com ruído de baixa freqüência, neste caso não é modelado. Esta última excitação pode ser colocada diretamente na saída do trato vocal, para modelar a turbulência produzida ao nível labial [KLATT, 1980].

Têm sido propostas algumas alterações ao modelo de síntese de formantes. Uma delas, propostas por Lalwani e Childers em 1991, propõe a possibilidade de variar o número de ressonâncias não apenas em função da freqüência de amostragem e do comprimento do trato vocal, mas dependendo também do som a sintetizar. O inconveniente desta alteração é que a introdução ou remoção de ressonadores durante a síntese conduz facilmente a sons transitórios indesejados. A solução encontrada consistiu em utilizar vários bancos de filtros de síntese em paralelo de forma que a variação entre conjuntos de parâmetros consecutivos possa ser feita, não por interpolação, mas pela soma pesada das saídas dos diversos bancos, cada um com o seu conjunto de parâmetros (VERHELST e NILENS, 1986). Desta forma resolve-se também o problema dos artefatos produzidos nos modelos tradicionais quando há variações demasiadamente rápidas dos valores das freqüências formantes.

Uma das vantagens no método de Formantes é que são utilizadas regras para converter fonemas em som e se isto for bem explorado, podem ser facilmente incluídas regras para muitos tipos de sons. E ainda tem o fato de que cada fonema é criado independentemente como uma função de seu contexto inteiro. Por outro lado, temos como desvantagens a grande complexidade na criação do sintetizador, e ainda, se não soubermos identificar o tipo de som produzido não é possível a criação de novas regras. Além do que, fala macia e natural é possível produzir, requerendo regras muito complicadas.

5.2 Métricas adotadas

Para realizar uma avaliação comparativa dos métodos de síntese de voz, é necessário primeiramente definir as métricas que serão utilizadas nas avaliações. Para este estudo, definimos as métricas descritas abaixo.

5.2.1 Métricas para a avaliação teórica

Estas métricas estão associadas a forma de implementação quando empregadas nos softwares de sintetização TTS. São as seguintes:

- Complexidade da implementação. É o parâmetro que determina os diversos graus de dificuldades na construção do software.
- Tamanho do dicionário de fonemas: É o parâmetro que determina o tamanho do dicionário de informações fonéticas e regras de conversão de letras para fonemas.
- Versatilidade para implantação de idiomas: É o parâmetro que verifica a possibilidade de se utilizar vários idiomas no sintetizador. De acordo com a escolha do operador o software pode sintetizar voz no idioma que se desejar. Por exemplo, no idioma português, idioma inglês, idioma holandês, idioma francês, idioma alemão, idioma espanhol, idioma estoniano, idioma russo e etc...
- Processamento: É o parâmetro que determina a quantidade de memória utilizada na execução dos softwares implementados em cada método. Qual o método que consome menos memória para sintetizar voz.

5.2.2 Métricas para a avaliação prática

Estas métricas estão associadas ao desempenho do ponto de vista do usuário final. Elas são:

- Tempo de respostas: este é o tempo entre a solicitação da síntese de voz e da apresentação da voz propriamente dita. Na execução do experimento, os softwares analisados tiveram um tempo de resposta muito próximos, cujas diferenças são imperceptíveis pelos usuários.
- Qualidade da sintetização: é a qualidade de voz percebida pelos usuários quando da utilização dos softwares que foram implementados com os métodos de Concatenção e Formantes.

5.3 Análise Teórica

Esta seção apresenta uma avaliação teórica das técnicas de síntese de voz Concatenação e Formantes. A tabela 3 resume os resultados da avaliação utilizando as

métricas apresentadas na seção anterior. Nas subseções que seguem, cada uma das métricas avaliadas estão melhor detalhadas.

Parâmetro	Concatenação	Formantes
Complexidade	É complexo dimensionar as unidades a concatenar. Utilizando-se unidades curtas como difones por exemplo, é menos complexo que controlar freqüências formantes.	É complexo controlar a variação do número de ressonâncias em função da freqüência. Isto torna-se mais complexo que concatenar unidades curtas de fala.
Tamanho do Dicionário	O número de difones necessários para cobrir todas as combinações de segmentos é igual ao quadrado do número de segmentos fonéticos da língua. Como nem todo par de segmento ocorre isto reduz bastante o número de difones, tornando assim o dicionário bastante pequeno.	Usa regras para converter fonemas em som. Se isto for bem explorado, podem ser facilmente incluídas regras para muitos tipos de sons. Muitas regras requer um dicionário bastante extenso.
Versatilidade	Unidades para concatenação podem ser extraídas diretamente da fala natural sem ser necessário conhecer e modelar muitos dos detalhes com relevância perceptual. Criar um banco de dados de difones para um idioma é mais rápido e fácil que modelar freqüências formantes para o mesmo.	Podem ser incluídas regras para muitos sons facilmente caso se saiba o que eles são. Se não soubermos identificar os vários tipos de sons que podem ser produzidos pelos fonemas de um idioma, torna-se impossível modelar as suas freqüências formantes.
Processamento	Muito menos computação (se utilizado unidades curtas). Pouca memória é utilizada para processamento do sintetizador.	Muita computação. Mais memória é utilizada para processamento do sintetizador.

TABELA 3. Análise comparativa (Teórica): Concatenação X Formantes

5.3.1 Complexidade

No método de Concatenação é complexo dimensionar as unidades a concatenar. Unidades demasiadamente longas obrigam a um extenso inventário que englobe todas as seqüências possíveis. A distorção de descontinuidades em pontos de concatenação podem ser reduzida usando difones ou alguns métodos especiais para suavizar o sinal. O uso de unidades de concatenação longas, como palavras por exemplo, exigem muita memória, dificultando assim o processamento.

No método de Formantes é complexo controlar a variação do número de ressonâncias em função da freqüência de amostragem e do comprimento do trato vocal, devido a sua dependência do som a sintetizar. A introdução ou remoção de ressonadores durante a síntese conduz facilmente a sons transitórios indesejados. A solução encontrada é a utilização de vários bancos de filtros de síntese em paralelo de forma que a variação entre conjuntos de parâmetros consecutivos possa ser feita, não por interpolação, mas pela soma pesada das saídas dos diversos bancos, cada um com o seu conjunto de parâmetros. É grande a complexidade na criação do sintetizador.

A síntese de Formantes é mais complexa, pois não é simples modelar todas as freqüências formantes dos vários tipos de sons produzidos pelos fonemas de uma língua.

5.3.2 Tamanho do Dicionário

No método de Concatenação, o número de difones necessários para cobrir todas as combinações de segmentos é igual ao quadrado do número de segmentos fonéticos da língua. No entanto, nem todos os pares de segmentos ocorrem e o número de difones pode ser substancialmente reduzido. Os fenômenos de coarticulação estendem-se muitas vezes para além do segmento seguinte e por esse motivo é comum a utilização de algumas unidades mais longas, como o trifone, meias-sílabas, sílabas, ou mesmo palavras inteiras. Outro fator que faz aumentar a dimensão do inventário fonético é a inclusão de variações alofônicas dos segmentos fonéticos.

No método de Formantes podem ser facilmente incluídas regras para muitos tipos de sons. Cada fonema é criado independentemente como uma função de seu

contexto inteiro. Se não soubermos identificar o tipo de som produzido não é possível a criação de novas regras. Portanto o número de regras é proporcional ao número de fonemas utilizados na língua (idioma).

Modelar todas as freqüências formantes dos sons produzidos pelos fonemas de um idioma exige um dicionário de grande capacidade.

5.3.3 Versatilidade

No método de Concatenação, as unidades podem ser extraídas diretamente da fala natural sem ser necessário conhecer e modelar muitos dos detalhes com relevância perceptual. Esta característica possibilita um mais rápido desenvolvimento e alteração do sistema, sendo assim mais fácil criar o sintetizador.

No método de Formantes, são utilizadas regras para converter fonemas em som e se isto for bem explorado, podem ser facilmente incluídas regras para muitos tipos de sons. Porém o controle rígido dessas regras torna-se trabalhoso.

Portanto concatenar unidades curtas de fala para um determinado idioma é mais versátil do que criar regras para converter fonemas em sons desse idioma.

5.3.4 Processamento

No método de Concatenação, utilizando-se unidades curtas de fala para concatenar, tais como: fonemas, difones e trifones; é exigida pouca memória para processamento dos mesmos. E é possível produzir voz com alta qualidade.

No método de Formantes, a modelagem dos circuitos ressonadores como também o controle adequado do conjunto de regras, exige normalmente muita memória para processamento dos mesmos. Fala macia e natural é possível reproduzir mas requer regras complicada.

O método de Concatenação utiliza pouco processamento e memória para sintetizar voz em relação ao método de Formantes.

5.4 Análise Experimental: Avaliação da Qualidade de Voz

Para a avaliação da qualidade de voz, foi necessário a realização de um experimento para se chegar a um resultado. Realizada a revisão bibliográfica constatou-se que não havia nenhum trabalho, em língua portuguesa, que realizasse a avaliação da qualidade de voz sintetizada utilizando a análise dos dados de forma qualitativa. A medição da qualidade da síntese de voz, tradicionalmente, tem sido subjetiva, isto é, ouvindo e comparando a qualidade do sinal de voz. Isto nos estimulou a sugerir a metodologia aqui empregada.

Nesta seção, inicialmente será apresentado alguns aspectos relacionados a percepção da fala. Em seguida são apresentadas formas de medição da qualidade de voz. Na seqüência é apresentada a metodologia proposta e os resultados obtidos.

5.4.1 Percepção da Fala

De acordo com RUSSO e BEHLAU (1993), a percepção dos sons da fala envolve um sistema de interação complexa que ultrapassa a realidade da simples detecção de sinais acústicos. As características acústicas dos sons da fala são consideravelmente mais complexas do que as dos sons utilizados na avaliação audiológica, tais como: tons puros, “clicks” e ruídos. Alem disso, o estímulo de fala precisa ser identificado, categorizado e reconhecido em sua forma. Assim, o processo de percepção da fala possui uma estreita relação com a atividade motora cognitiva envolvida em sua produção.

A percepção da fala apresenta uma série de etapas, iniciando-se com a *audibilidade*, isto é, com a detecção do som. A partir da audibilidade temos a recepção da informação sonora, a *discriminação* entre sons de diferentes espectros, o *reconhecimento* ou a comparação do que foi ouvido com experiências anteriores, a *memória* ou retenção e evocação de elementos da fala e, finalmente, a *compreensão* da mensagem falada [RUSSO e BEHLAU, 1993].

Além das etapas acima referidas, Keith(1982) destaca três fatores que fazem parte do processo lingüístico e cognitivo do indivíduo ao receber um sinal de fala, a saber: *análise-síntese*, *seqüenciação* e *fechamento auditivo*. A *análise-síntese* é a decomposição e a integração das informações de fala recebidas simultânea ou

alternadamente; a *sequenciação* auditiva é a capacidade de ordenar os estímulos sonoros e, por fim, o *fechamento auditivo* é a reconstrução da mensagem sonora, quando parte desta foi omitida, ou quando o ouvinte realiza suplência mental, mesmo antes do término da fala.

Desta maneira, a percepção dos sons da fala inclui a recepção e interpretação dos padrões de fala; a discriminação entre sons de diferentes espectros, durações, características temporais, formas seqüenciais e ritmo; o reconhecimento, a memorização e a compreensão de unidades de fala dentro de determinado sistema lingüístico.

Para a efetividade da transmissão da mensagem existe uma redundância de pistas acústicas de que o ouvinte vai se valer, de acordo com a situação e o contexto da comunicação. A redundância de pistas acústicas é uma garantia natural da transmissão da mensagem. Basicamente, a energia do sinal de fala deve ser suficientemente audível e os elementos acústicos desse sinal devem ser passíveis de discriminação, o que envolve a segmentação destes em unidades menores, as quais serão armazenadas na memória para comparação, reconhecimento e compreensão.

Ainda segundo, RUSSO e BEHLAU (1993), o sucesso de um ouvinte para compreender a fala depende de processos supra-liminares, diretamente relacionados aos seguintes fatores: atenção a mensagem; intensidade da mensagem; intensidade do ruído; tipo de material de fala; coarticulação e fatores suprasegmentais; sensação de freqüência (“pitch”); sensação de intensidade (“loudness”); fatores temporais, ritmo e velocidade; qualidade vocal do falante; articulação e pronúncia.

5.4.2 Medindo a qualidade de voz

Não existe ainda um padrão específico para medição da qualidade de voz gerada pelos sintetizadores. No caso da telefonia, o padrão adotado para medição da qualidade de voz é o MOS (*Mean Option Score*) [P.800, 1996]. O MOS é uma medida subjetiva da qualidade da voz oriunda da telefonia, que é usado para avaliar a qualidade de voz em chamadas telefônicas.

Ao usar MOS com ouvintes humanos, um grupo de pessoas ouve o áudio gerado pelo sintetizador e dá suas opiniões sobre a qualidade da voz de acordo com as

características que se queira avaliar, pontuando-se com os conceitos especificados na tabela 4.

Qualidade	MOS
Excelente	5
Bom	4
Regular	3
Insatisfatório	2
Ruim	1

Tabela 4 . Escala MOS [P.800, 1996].

O MOS é um processo que funciona bem na telefonia e tem sido muito utilizado pelos pesquisadores na área, mas pode ser difícil e dispendioso de executar. Também há o inconveniente da necessidade de se formar um grupo de pessoas aptas para a experimentação.

5.4.3 Metodologia

Para medir a qualidade de voz sintetizada pelos métodos de Concatenação e de Formantes, foi definida uma metodologia própria, envolvendo a consulta a um grupo de pessoas para avaliar a qualidade de voz percebida por este grupo. Este grupo ouvirá uma série de sentenças e fonemas sintetizados por softwares de síntese de voz. Cada sentença é sintetizada diversas vezes, utilizando softwares de TTS. Após ouvir cada "leitura", o consultado deve avaliar a qualidade de voz que será registrada via o preenchimento de um questionário (figura 5). Cada pessoa do grupo, individualmente, ouvirá o som da voz que lerá o texto e fonemas grafados no formulário e atribuirá seus respectivos conceitos segundo a métrica do MOS. Optamos por selecionar pessoas adultas e com audição perfeitamente sadia para o sucesso do experimento.

As etapas desta metodologia são apresentadas a seguir:

a) Delimitação do Universo (Descrição da população)

A pesquisa teve como universo pesquisado, um grupo de 20 (vinte) pessoas adultas, com audição perfeitamente sadias, reflexos e percepções normais. Funcionários

públicos, lotados na área de informática, cursando ou já detentores de curso superior e potenciais usuários da tecnologia TTS.

b) Definição dos Formulários

Baseado na bibliografia de Fonoaudiologia [GAMA, 1994] e [RUSSO & BEHLAU, 1993]; e na consulta realizada a fonoaudióloga Drª Cecília Medeiros Oliveira, foi definido um formulário (figura 5) para avaliar a qualidade de voz. Este formulário visou avaliar os quesitos de naturalidade, entendimento e pronúncia.

O formulário definido foi dividido em duas partes:

- Parte **I Avaliação de Compreensão de Sentenças:** Esta parte do formulário foi elaborada segundo o teste de compreensão de sentenças inserido na avaliação GASP (Glendonald Auditory Screening Procedure), desenvolvida por Erber (1982). O estímulo utilizado são sentenças que foram adaptadas a um vocabulário menos infantil, como a proposta original. Todas as formas de pronomes interrogativos são utilizados nas 10 (dez) sentenças criadas. Desse modo, essa avaliação está em um nível mais alto de complexidade lingüística e de contexto de fala. Ela tenta reproduzir uma situação mais próxima da conversação diária.

- Parte II **Avaliação de Detecção de Fonemas: esta parte foi** elaborado a partir dos sons propostos por Ling (1978) e com o acréscimo das consoantes fricativas */v/, /z/* e da nasal */m/*. As fricativas foram escolhidas devido à faixa de freqüência que elas abrangem, além da intensidade média dentre os fonemas do seu grupo. O fonema */m/* foi escolhido devido à importância da característica da nasalidade dos fonemas em português, observada por Ling (1992).

FORMULÁRIO

Avaliação de Compreensão de Sentenças

Nome:

Nr. Ordem	Sentenças/Estímulo	Quesitos/Notas/Softwares											
		Naturalidade				Entendimento				Pronúncia			
		A	B	C	D	A	B	C	D	A	B	C	D
01	Quando é o seu aniversário?												
02	Que tipo de comida você gosta?												
03	Onde fica sua casa?												
04	Qual o animal que você mais gosta?												
05	Como você veio até aqui?												
06	Quem o está acompanhando?												
07	Quantas blusas você está vestindo?												
08	Em que você trabalha?												
09	O que você está achando deste teste?												
10	Quantos anos você tem?												
T o t a i s													

Notas (MOS) 5-Excelente 4-Bom 3-Regular 2-Insatisfatório 1-Ruim

Avaliação de Detecção de Fonemas

Nome:

Fonemas	/a/				/i/				/u/				/m/				/s/				/v/				/x/				/z/			
Softwares	A	B	C	D	A	B	C	D	A	B	C	D	A	B	C	D	A	B	C	D	A	B	C	D	A	B	C	D	A	B	C	D
Respostas																																

● Não detectou o fonema + Detectou o fonema

Software A: MBROLA Synthesizer version 3.01d

Software B: TextAloud version 1.408

Software C: Audio Book Creator (ABC) version 2004.2.1

Software D: ShadiSoft Speak version 1.8.84

Figura 5 – Formulário de Pesquisa

c) Softwares Utilizados

Para avaliar e comparar a qualidade de voz produzida por softwares de TTS utilizando os métodos de concatenação e formantes, a presente pesquisa procurou determinar um conjunto de softwares que utilizam os métodos avaliados. Os softwares identificados e testados foram aqueles apresentados na tabela 5.

Software	Versão	Fabricante	Método utilizado
MBROLA Synthesizer	3.01d	Faculté Polytechnique de Mons TCTS	Concatenação
TextAloud	1.40 8	NextUp Technologies	Concatenação
Áudio Book Creator (ABC)	2004.2.1	Audio Book Creator	Formantes
ShadiSoft Speak	1.8.84	ShadiSoft.Com	Formantes

TABELA 5. Softwares utilizados na experimentação

Para não haver nenhum tipo de direcionamento, não foi esclarecido aos ouvintes as características dos softwares utilizados na sintetização e nem dos métodos que o mesmo utilizavam. Os softwares foram apenas identificados como A, B, C e D.

d) Geração dos arquivos de voz

As sentenças e fonemas usados no formulário (figura 5) foram inseridos (digitados) nos editores dos softwares da Tabela 5 a fim de produzir uma série de arquivos de áudio para cada uma das sentenças e fonemas.

e) Aplicação da Pesquisa

Do grupo de ouvintes selecionados, foi solicitada à presença individual de cada um deles a uma sala equipada com microcomputador, onde estavam instalados os softwares da tabela 5 e colocado a sua disposição o formulário. Após uma rápida explicação de como preencher os formulários, os mesmos ouviam cada um dos áudios reproduzidos pelos softwares sintetizadores e anotavam o seu conceito.

f) Tratamento dos Dados

Para o tratamento dos dados coletados, foi utilizado o processo de elaboração de tabelas e gráficos que retrataram os resultados obtidos através das questões objetivas dos questionários. Além da totalização de escores brutos e média aritmética, a principal técnica estatística empregada para a comparação de grupos de unidades amostrais foi a *análise de variância* e *testes de aleatorização*, que são aplicáveis a dados univariados e multivariados. Testes de aleatorização geram com base nos próprios dados as probabilidades usadas para julgar a significância das diferenças entre grupos. Os dados coletados foram introduzidos e processados nos softwares STATISTICA versão 5.1 e Microsoft EXCEL 2000, onde foram gerados os resultados dos testes.

A compilação dos dados por meio deste método, ou seja a digitação em planilha eletrônica, permitiu estruturá-los em blocos de informações comuns, agilizar o tratamento dos dados e preservá-los para a realização de pesquisas futuras sobre sintetização de voz.

Cada software foi analisado e comparado individualmente e posteriormente avaliados em grupos de acordo com o método implementado. Grupo AB, implementados pelo método de Concatenação e o grupo CD, implementados pelo método de Formantes.

5.4.4 Resultados da Experimentação

Após a aplicação da pesquisa no universo de 20 (vinte) pessoas, foi realizada a tabulação dos dados do formulário.

a) Resultado da tabulação da parte I

A tabulação de cada conceito atribuído aos quesitos (naturalidade, entendimento e pronúncia) para cada software, foram calculadas as médias que estão demonstradas na Tabela 6.

SOFTWARES	QUESITOS/MÉDIAS		
	NATURALIDADE	ENTENDIMENTO	PRONÚNCIA
A	3,495	4,440	3,155
B	3,405	4,350	2,985
C	3,270	4,265	2,945
D	3,185	4,155	2,795

TABELA 6. Média de Escores

Tendo como base as médias calculadas na Tabela 6, estas enquadradas nos níveis dos conceitos da Tabela MOS, obtivemos os resultados demonstrados na Tabela 7.

SOFTWARES	QUESITOS/CONCEITO FINAL		
	NATURALIDADE	ENTENDIMENTO	PRONÚNCIA
A	REGULAR	BOM	REGULAR
B	REGULAR	BOM	INSATISFATÓRIO
C	REGULAR	BOM	INSATISFATÓRIO
D	REGULAR	BOM	INSATISFATÓRIO

TABELA 7. Conceitos finais

O gráfico 1 apresenta a análise de variância dos resultados apresentados na tabela 6.

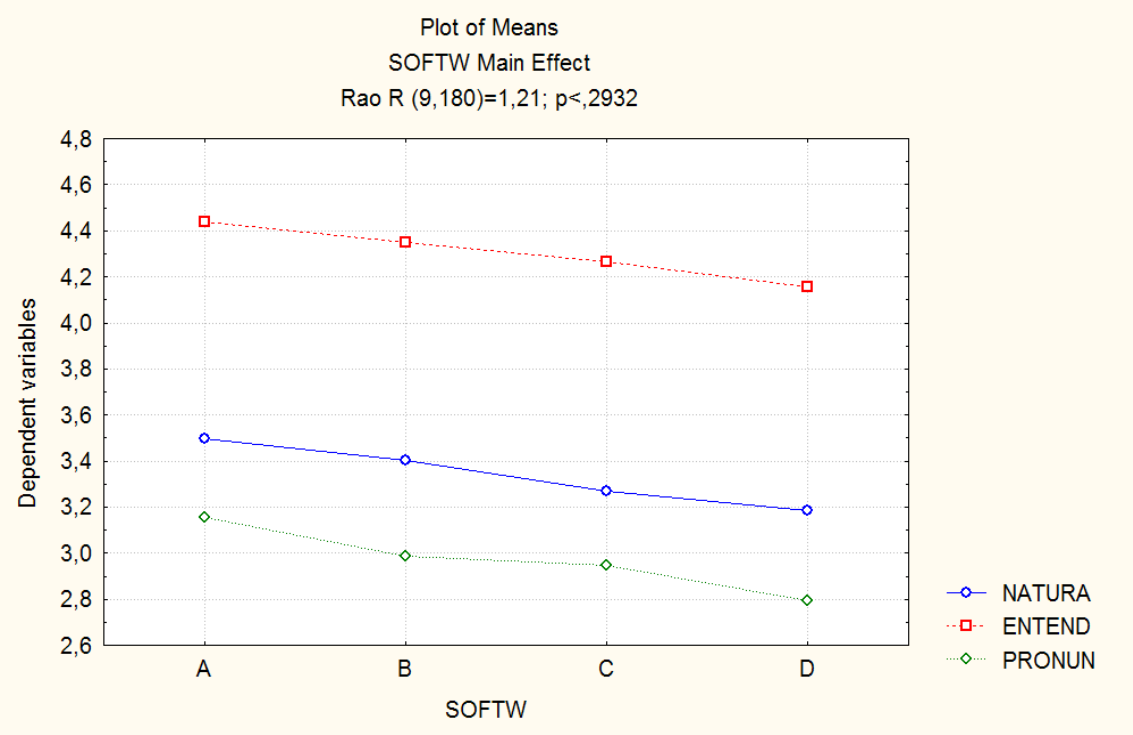

GRÁFICO 1. Análise de Variância (Software X Quesitos).

Ficou evidenciado então que não há uma diferença significativa, pois $p < 0,2932$, para a diferença ser significativa o resultado de p deveria ser: $p < 0,05$. Isto só ocorreria se as médias fossem bastantes divergentes e conforme o demonstrado na Tabela 6, fica evidenciado que elas estão muito próximas não acusando diferença significativa. Portanto, para os softwares avaliados a qualidade de voz obtida é a mesma para todos.

Como não há uma diferença significativa não podemos avaliar o possível efeito dos tratamentos. Resta-nos apenas realizar a avaliação comparativa entre os grupos de softwares. Grupo AB, grupo de softwares implementados pelo método de Concatenação e Grupo CD, que é o grupo de softwares implementados pelo método de Formantes e por fim, comparar ABxCD (Concatenção X Formantes).

Avaliação comparativa dos Método de Concatenação e Formantes

O resultado da avaliação da média de escore do grupo AB (Concatenação) e do grupo CD (Formantes) está demonstrado na Tabela 8.

SOFTWARES	QUESITOS/MÉDIAS		
	NATURALIDADE	ENTENDIMENTO	PRONÚNCIA
AB	3,45	4,395	3,07
CD	3,2275	4,21	2,87

TABELA 8. Médias de Escores dos Grupos

O resultado do confronto das médias dos grupos AB x CD, está demonstrado na Tabela 9.

Parâmetro	Concatenação	Formantes
Média Geral de Escores	3,6383	3,43583

TABELA 9. Média Geral de Escores: Concatenação X Formantes

De acordo com os resultados acima obtidos, apesar de não haver diferença significativa na análise de variância, o confronto direto nos possibilita afirmar que neste caso os softwares implementados pelo método de Concatenação são os melhores.

b) Resultado da tabulação da Parte 2

A tabulação dos percentuais de fonemas não detectados pelos softwares A, B, C e D, apontados no Formulário (parte 2) , está demonstrado no Gráfico 2.

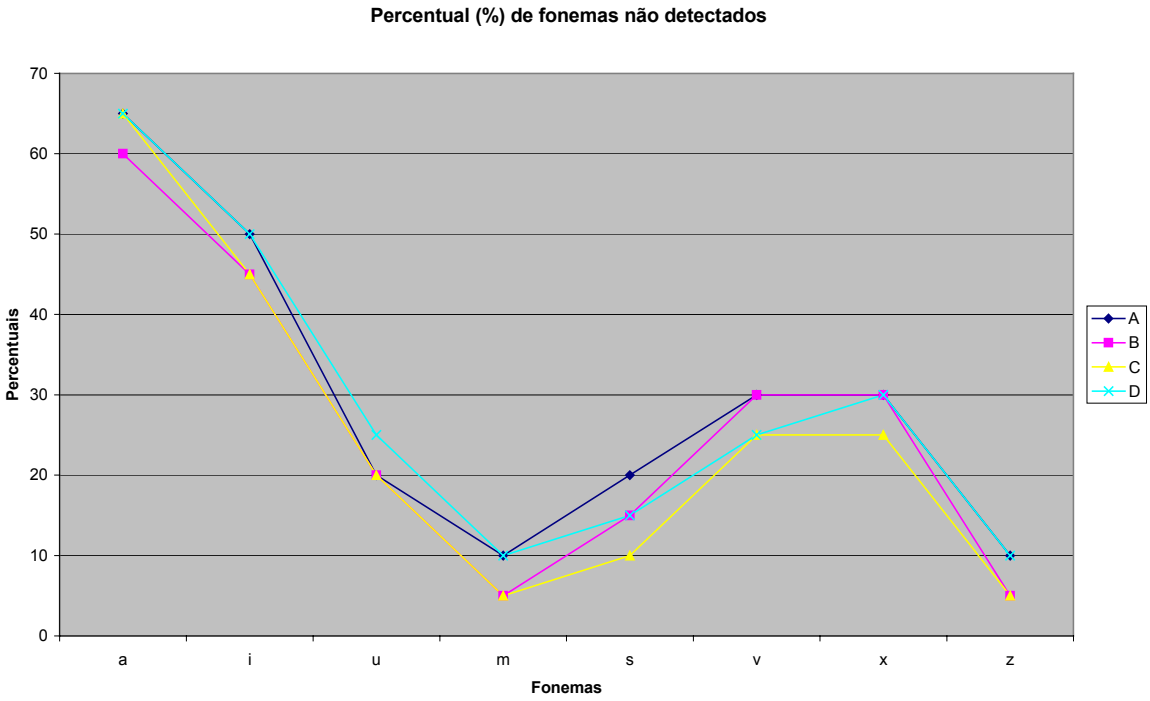

GRÁFICO 2. Percentual (%) de fonemas não detectados

O que fica evidenciado a grande proximidade de percentuais de fonemas não detectados pelos softwares utilizados na experimentação. O que implica que todos eles tem deficiência na detecção desses fonemas, devendo assim, serem ajustados em suas implementações para detecção correta dos mesmos.

5.5 Conclusão

Neste capítulo foi apresentado a análise comparativa entre os métodos de Concatenação e Formantes através de análises teóricas onde foram analisados os parâmetros de: complexidade, tamanho do dicionário de fonemas, versatilidade para implantação de idiomas e processamento. E também através da experimentação, onde foi analisado o parâmetro de tempo de respostas e realizada uma pesquisa de campo que a partir desta efetivou-se uma avaliação da qualidade de voz.

Na análise teórica o método de Concatenação foi melhor em todas as métricas aqui estipuladas, quais sejam: complexidade, tamanho do dicionário de fonemas, versatilidade para implantação de idiomas e processamento.

Na análise prática (experimentação) os resultados obtidos indicaram que para os softwares utilizados a qualidade de voz gerada é a mesma nos dois métodos (Concatenação e Formantes).

Apesar de não haver diferença significativa no experimento prático realizado, a análise teórica demonstrou que o método de Concatenação é a melhor alternativa para o desenvolvimento de um sintetizador, pois além da sua grande facilidade de desenvolvimento e baixo custo de implementação, o sinal de voz gerado é de altíssima qualidade.

Capítulo 6 Conclusões

A síntese de voz TTS é de suma importância para o nosso dia a dia, pois ela nos ajuda no manuseio de certos equipamentos eletrônicos (caixas de auto-atendimento bancário, por exemplo), como também resolve em parte os problemas sociais de pessoas portadoras de deficiência vocal e deficiência visual. A qualidade dos produtos desenvolvidos com essa tecnologia alcança um nível adequado para várias aplicações, tais como multimídia e telecomunicações.

Fala sintetizada pode ser produzida através de vários métodos diferentes. Todos eles apresentam alguns benefícios e deficiências. Os métodos normalmente são classificados conforme abaixo:

- Síntese Articulatória, que tenta modelar o sistema de produção de fala humana diretamente.
- Síntese de Formantes, que modela as freqüências do sinal de som da fala ou transfere a função de área vocal baseado em fonte-filtro-modelo.
- Síntese de Concatenação, que usa amostras de pré-gravação, de comprimentos diferentes de fala natural.
- Síntese LPC (*Linear Predictive Coding*), que sintetiza o sinal de voz usando o resíduo para criar um sinal de fonte e usa os formantes para criar um filtro.

O método de Concatenação e o método de Formantes são geralmente os mais usados em sistemas de síntese de voz atualmente.

O presente trabalho objetivou fazer uma análise comparativa entre os métodos de Concatenação e Formantes, utilizados na síntese TTS, através de análises teóricas e experimentação.

Para atingir este objetivo foram definidas métricas teóricas e práticas, para que fosse possível avaliar os resultados.

A análise comparativa entre os métodos de Concatenação e Formantes através de análises teóricas onde foram analisados os parâmetros de: complexidade, tamanho do dicionário de fonemas, versatilidade para implantação de idiomas e processamento. Como também a experimentação, onde foi analisado o parâmetro de tempo de respostas e realizada uma pesquisa de campo, nos permite chegar ao seguinte resultado:

Na análise teórica o método de Concatenação foi melhor em todas as métricas definidas acima.

Na análise prática (experimentação) os resultados obtidos indicaram que para os softwares utilizados a qualidade de voz gerada é a mesma nos dois métodos (Concatenação e Formantes).

Em suma, através da experimentação observamos que a subjetividade torna-se um empecilho considerável para a avaliação dos métodos de síntese de voz, mesmo assim, concluímos que os métodos aqui analisados, quais sejam método de Concatenação e método de Formantes, tornan-se eficazes dependendo da maneira pela qual serão implementados, ou seja, quais ferramentas (linguagem de programação, ambiente operacional, etc...) serão utilizadas na construção do sintetizador. Cada método possui suas vantagens e desvantagens, mas dependendo das habilidades e conhecimentos do implementador, um deles se tornará mais eficiente. Nesta experimentação concluímos que o método de Concatenação leva uma pequena vantagem em relação ao método de Formantes, fato comprovado na análise teórica. O resultado final da análise dos softwares utilizados na experimentação é que os mesmos são equivalentes, pois isto ficou comprovado quando da aplicação da Análise de Variância, onde ficou constatado que não houve diferença significativa entre as médias.

Baseado nos resultados da análise teórica afirmamos que o método de Concatenação é a melhor alternativa para o desenvolvimento de um sintetizador de voz humana.

Levando-se em consideração alguns aspectos apontados neste trabalho de pesquisa, bem como temas ainda não explorados na Literatura, conclui-se que alguns pontos merecem atenção especial, sobre os quais serão feitas algumas sugestões para futuros estudos:

- Realização de um estudo estatístico para a preparação de um conjunto de regras com maior abrangência, propiciando melhores resultados na conversão fonética de palavras através de regras.
- Realização de estudos para propor um modelo de Sintetizador Jurídico, que seria utilizado para ler as peças processuais em um tribunal do júri.

Capítulo 7 Referências

CALLOU,D.; LEITE,Y. **Iniciação à Fonética e à Fonologi**a. 1 ed. Rio de Janeiro, São Paulo, Jorge Zahar Editor Ltda. 1990.

CROCHIERE,R.E.; FLANAGAN,J.L. Speech Processing: an Evolving Technology. **AT & T Technical Journa**l, v. 65, n. 5, p. 2-11, Sept/Oct. 1990.

CUNHA,C.; CINTRA,L. **Nova Gramática do Português Comtemporâne**o. 2 ed. Rio de Janeiro, Editora Nova Fronteira. 1985.

EGASHIRA,F.; VIOLARO,F. **Síntese de Voz a Partir de Text**o. Campinas, Faculdade de Engenharia Elétrica da Universidade Estadual de Campinas, 1993.(Publicação FEE 01/93).

CEGALLA, D.P. **Novíssima Gramática da Língua Portuguesa**. 16 ed. São Paulo, Companhia Editora Nacional. 1977.

FLANAGAN,J.L. **Speech Analysis Synthesis and Perceptio**n. 2 ed. New Jersey, Springer-Verlag, 1972.

AL-SUWAIYEL,M.; HOROWITZ,E. Algorithms for Trie Compaction. **ACM Transactions on Database System**s, v. 9, n. 2, p. 243-63, June 1984.

LUCCHESI,C.L.; KOWALTOWSKI,T. **Applications of Finite Automata Representing Large Vocabularie**s. Campinas, Departamento de Ciência de Computação, 1991.

OLIVEIRA,L.C.; VIANA,M.C.; TRANCOSO,I.M. **A Rule-Based Text-to-Speech System for Portuguese**. Lisboa, INESC/IST/CLUL, 1994.

ALLIANDRO, H. **The Portuguese-English Dictionar**y. 10 ed. New York, Pocket Book. 1974.

FERREIRA, A.B.H. **Novo Dicionário da Língua Portugues**a. 2 ed. Rio de Janeiro, Editora Nova Fronteira, 1986.

PAIS,C.D. Elementos de Fonologia Estrutural. In: **Manual de Lingüístic**a. Petrópolis, Editora Vozes Ltda, 1978, p. 9-80.

VIOLARO,F. Panorama de Investigações em Processamento de Fala no Brasil. In: **Encontro de Processamento da Língua Portuguesa Escrita e Falada**, 1., Lisboa, 1993.

CAMPOS,G.L. **Síntese de Voz para o Idioma Português**. São Paulo, 1980. Tese (Doutorado em Engenharia) - Escola Politécnica, Universidade de São Paulo.

CASAES,E.J. **Descrição Acústico-Articulatória dos Sons da Voz para um Modelo dos Sons do Português do Brasi**l. São Paulo, 1990. Tese (Doutorado em Engenharia) – Faculdade de Filosofia, Letras e Ciências Humanas, Universidade de São Paulo.

CHBANE, DIMAS TREVISAN. **Desenvolvimento de Sistema para Conversão de Textos em Fonemas no Idioma Português**. São Paulo, 1994. Dissertação (Mestrado em Engenharia) - Escola Politécnica, Universidade de São Paulo.

ALLEN,J. Synthesis of Speech from Unrestricted Text. **Proceedings of the IEE**E, v. 64, n. 4, p. 433-42, Apr. 1987.

AINSWORTH,W.A. A System for Converting English Text into Speech. **IEEE Transactions on Audio and Electroacoustic**s, v. 21, n. 3, p. 288-90, June 1973.

APPEL,A.W.; JACOBSON,G.J. The World's Fastest Scrabble Program. **Communications of the AC**M, v. 31, n. 5, p. 572-8, May 1988.

ATAL,B.S.; RABINER,L.R. Speech Research Directions. **AT & T Technical Journa**l, v. 65, n. 5, p. 75-88, Sept./Oct. 1986.

ALLEN, J. Reading Machines for the Blind: The Technical Problems and the Methods Adopted for Their Solution. **IEEE Transactions on Audio andElectroacoustic**s, v. 21, n. 3, p. 259-64, June 1973.

COKER,C.H.; UMEDA,N.; BROWMAN,C.P. Automatic Synthesis from Ordinary English Text. **IEEE Transactions on Audio and Electroacoustic**s, v. 21, n. 3, p. 293-8, June 1973.

ESQUIVEL,A.S. Um Sistema de Síntese de Voz. In: Congresso Nacional de Informática, 18., São Paulo, 1985. **Anai**s. São Paulo, Sucesu, 1985. p. 776-82.

FLANAGAN,J.L.; COKER,C.H.; RABINER,L.R.; SCHAFER,R.W.; UMEDA, N. Synthetic Voices for Computers. **IEEE Spectru**m, v. 7, p. 22-45, Jan. 1970.

GROSS,M. The Use of Finite Automata in the Lexical Representation of Natural Language. In: M. Gross and D. Perrin, ed. **Eletronic Dictionaries and Automata in Computational Linguistic**s. Berlin, Springer-Verlag, Berlin, 1989. p. 34-50 (Lectures Notes in Computer Science 377)

HAGGARD,M.P.; MATTINGLY,I.G. A Simple Program for Synthesizing British English. **IEEE Transactions on Audio and Eletroacoustic**s, v.16, n. 1, p. 95-9, Mar. 1968.

HERTZ,S.R. From Text to Speech with SRS. **Journal of the Acoustical Society of Americ**a, v. 74, n. 4, p. 1155-70, Oct. 1982.

HERTZ,S.R.; KADIN,J; KARPLUS,K.J. The Delta Rule Development System for Speech Synthesis from Text. **Proceedings of the IEE**E, v. 73, n. 11, p. 1589- 601, Nov. 1985.

HIRSCHBERG,J.B.; RIEDERER, S.A.; ROWLEY,J.E.; SYRDAL,A.K. Voice Response Systems: Technologies and Applications. **AT & T Technical Journa**l, v. 65, n. 5, p. 42-51, Sept./Oct. 1990.

KLATT,D.H. Linguistics Uses of Segmental Duration in English: Acoustics and Perceptual Evidence. **Journal of the Acoustical Society of Americ**a, v. 59, n. 5, p. 1208-21, May 1976.

KLATT,D.H.; KLATT,L.C. Analysis, Synthesis, and Perception of Voice Quality Variations Among Female and Male Talkers. **Journal of Acoustical Society of Americ**a, v. 87, n. 2, p. 820-56, Feb. 1990.

LAPORTE,E. Applications of Phonetic Description. In: M. Gross and D. Perrin, ed. **Eletronic Dictionaries and Automata in Computational Linguistic**s. Berlin, Springer-Verlag, Berlin, 1989. p. 65-78 (Lectures Notes in Computer Science 377)

LEE,L.S.; TSENG,C.H,; OUH-YOUNG,M. The Synthesis Rules in a Chinese Text-to-Speech System. **IEEE Transactions on Acoustics, Speech, and Signal Processin**g, v. 37, n. 9, p. 1309-20, Sept. 1989.

OLABE,J.C.; SANTOS,A.; MARTÍNEZ,R.; MUÑOZ,E.; MARTÍNEZ, M.;QUILIS,A.; BERNSTEIN,J. **Real Time Text to Speech Conversion System for Spanish**. In: H.W. Schüssler, ed. EURASIP Signal Processing: Theories and Applications, 2., Erlangen, 1983.

UMEDA,N. Linguistic Rules for Text-to-Speech Synthesis. **Proceedings of the IEE**E, v. 64, n. 4, p. 443-51, Apr. 1976.

VIANA,M.C.; ANDRADE,E; OLIVEIRA,L.C.;TRANCOSO,I.M. Ler_PE: Um Utensílio para o Estudo da Ortografia do Português. In: Encontro da Associação Portuguesa Lingüística, 7., Lisboa, 1991. **Anai**s. Lisboa, APL, 1991. p. 474-89.

OLIVEIRA, L.M.V.C. **Síntese de Fala a Partir de Texto.** Tese de Doutorado. Universidade Técnica de Lisboa, 1996.

YOUNG,S.J.; FALLSIDE,F. Speech Synthesis from Concept: A Method for Speech Output from Information Systems. **Journal of Acoustical Society of Americ**a, v. 66, n. 3, p. 685-95, Sept. 1989.

ETSnet. **Toefl on line** : Test of english as a foreign language. Disponível em:

<http://www.toefl.org>.

P.800, ITU-T Recommendation, ***Methods for Subjective Determination of Transmission Quality***, 1996.

Erber NP (1982) Auditory Training, Washington DC: AG Bell Assoc for the Deaf.

GAMA, MÁRICA REGINA. **Percepção da Fala: Uma Proposta de Avaliação Qualitativa**, Pancast, 1994.

IEDA RUSSO & MARA BEHLAU. **Percepção da Fala: Análise Acústica do Português Brasileiro**, Lovise, 1993.