UM SISTEMA DE SÍNTESE DE VOZ PARA A LÍNGUA PORTUGUESA

UNIVERSIDADE FEDERAL DO RIO DE JANEIRO

ESCOLA DE ENGENHARIA

DEPARTAMENTO DE ELETRÔNICA

Autor: _____
Cristiano Gaspar Machado

Orientador: _____
Márcio Nogueira de Souza

Examinador: _____
Antônio Claudio Gómez de Sousa

Examinador: _____
Luiz Wagner Pereira Biscainho

Rio de Janeiro - DEL
Fevereiro de 1997.

Índice

I. Introdução

Desde o início da ciência, o homem sempre procurou meios de facilitar sua vida. Estes meios procuravam substituir de alguma forma mais fácil aquilo que ele mesmo poderia fazer. A partir de um certo momento, este passou a perceber que as máquinas que fazia pretendiam apenas substituir o trabalho de uma máquina muito mais complexa e perfeita: o próprio homem. Foi esta consciência que originou uma curiosidade enorme em descobrir como o homem funcionava e descobrir um meio de copiar esta máquina o mais perfeitamente possível.

Esta curiosidade logo ganhou um alvo enorme, o qual diferenciava o homem de todos os outros animais: a fala, a capacidade de se comunicar e passar a experiência adiante. Para comprovar como esta curiosidade provém de muito tempo, as primeiras máquinas inventadas que tentam imitar a voz humana são datadas do século XIV. Esta precocidade no início das pesquisas nesta área é suficiente para demonstrar o quão complicada esta estrutura é, tendo em vista que ainda hoje não se tem um método definitivo para a síntese de voz e vários ramos da pesquisa estão em atividade.

Nos dias de hoje não se utilizam mais grandes foles controlados por pedais ou fitas magnéticas com gravações analógicas para se fazer a síntese da voz. Todas as pesquisas são feitas na área digital, seja com linguagens e computadores de uso genérico ou circuitos integrados completamente específicos. O desenvolvimento nesta área também agora não é mais uma questão de mera curiosidade, mas um avanço quase vital à humanidade, tendo em vista que cada dia mais estamos envoltos por mais e mais informações que serão muito melhor absorvidas e entendidas se nos forem passadas da maneira mais cômoda possível, ou seja, como se faladas por outra pessoa.

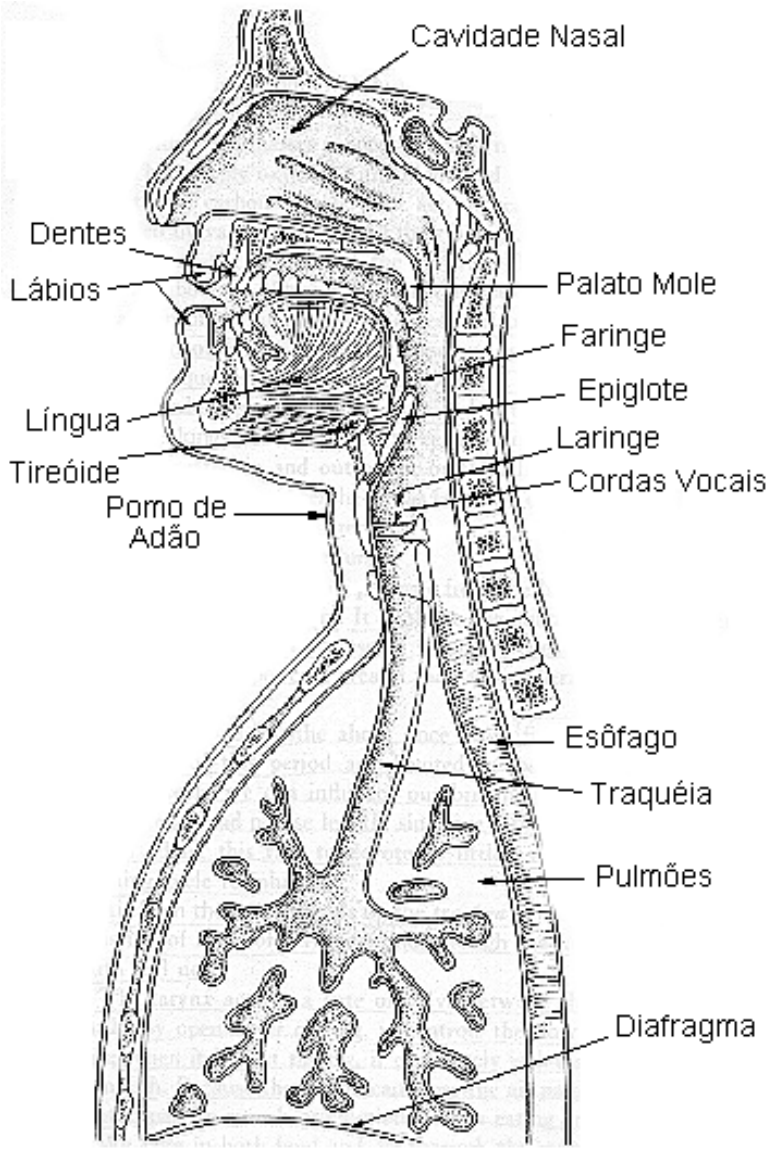
Além disto, pesquisas nesta área permitirão que pessoas com deficiências visuais, atualmente grandemente prejudicadas pela falta de opções, possam ter acesso fácil a materiais de leitura, equipamentos e situações hoje em dia quase inacessíveis.

Foi visando a estes objetivos que este projeto foi feito, esperando que seja ao menos uma contribuição no sentido de sanar esta “curiosidade” que existe em todos.

II. Uma Abordagem Sobre Produção de Voz

II.1 Os Órgãos Vocais

A voz é produzida por um conjunto de órgãos por isso denominados **órgãos vocais**. Os principais órgãos vocais são os pulmões, a traquéia, a laringe, a faringe, o nariz, e a boca (incluindo o palato, os dentes, a língua e os lábios). Juntos estes órgão formam um intrincado tubo que vai desde os pulmões até os lábios. A parte do tubo que fica acima da laringe é chamado de trato vocal, e consiste na faringe, boca e nariz. O formato do trato vocal pode variar muito, movendo-se o palato, a língua, os lábios e a mandíbula, que são por isto denominados **articuladores**.



A fonte de energia para a produção da voz é a constante corrente de ar que se produz ao se exalar. Esta corrente de ar que normalmente é inaudível pode tornar-se audível fazendo-se com que esta vibre rapidamente, o que normalmente é feito com as cordas vocais.

As cordas vocais são parte da laringe, e se constituem em uma barreira ajustável para o ar que vem dos pulmões. Quando falamos, fechamos e abrimos as cordas vocais rapidamente, transformando a corrente de ar que vem dos pulmões em uma série de pulsos de ar. Esta seqüência de pulsos de ar que se forma pode ser ouvida como um zumbido, que se torna mais agudo a medida que aumentamos a freqüência com que abrimos e fechamos as cordas vocais. As características deste zumbido são alteradas pelo trato vocal, alteração a qual depende do formato deste. Durante a fala, este formato é continuamente alterado, movendo-se a língua, os lábios e os articuladores, processo o qual nos habilita a produzir todos os diferentes sons da voz.

Além da vibração das cordas vocais, existem dois outros métodos de tornar a corrente de ar que vem dos pulmões em algum som audível. O primeiro método consiste em fazer uma constrição do tubo vocal em algum ponto, como por exemplo, colocando a língua próximo do palato e dos lábios. Deste modo o ar que passa sofre uma grande turbulência soando como um "chiado". Este é o som das consoantes fricativas (como s,ch,f).

Outro método consiste em bloquear completamente o fluxo de ar dos pulmões utilizando a língua ou os lábios, mas apenas por alguns instantes, logo depois, liberando de uma só vez este fluxo. Este som é o característico das consoantes explosivas (como p e g).

Alguns outros métodos podem ser empregados para a produção de voz, mas são raramente utilizados.

Deve-se notar que a função principal dos órgãos vocais não é a produção da fala, sendo necessários primariamente por outras funções vitais. Os pulmões, por exemplo, são os responsáveis pelo suprimento de oxigênio do organismo e pela liberação de gás carbônico resultante de reações internas. Este processo ocorre durante a respiração, a qual é produzida por vários músculos da caixa toráxica, por músculos do abdômen e pelo diafragma. Durante a fala, o diafragma se relaxa e os músculos do abdômen controlam o quanto o ar dos pulmões é pressionado contra o diafragma, jogando o ar para fora dos pulmões. Os músculos do peito também se contraem, reduzindo o tamanho da caixa toráxica. Quando se exala, a pressão do ar é aproximadamente 0.25% acima da pressão atmosférica, atingindo 1% acima em uma conversação.

Normalmente respira-se uma vez a cada cinco segundos, dedicando igual tempo para inalar e exalar. Durante a fala normalmente altera-se esta seqüência de acordo com a frase e o tamanho da palavra que se está pronunciando. Visto que se fala apenas ao exalar, pode-se controlar este ciclo de modo a dedicar apenas 15% do tempo ao processo de inalar.

O ar que sai dos pulmões atravessa a traquéia, que consiste em um tubo de anéis de cartilagem, passando pela laringe, boca e nariz.

A laringe age como uma válvula entre os pulmões e a boca, isolando-os ou não. Devido a esta função, a laringe tem um papel fundamental enquanto se respira, come ou fala. Como se ingere ar e comida pela boca, o controle do caminho a ser tomado por cada um destes é feito pela laringe, permitindo que o ar passe pela traquéia até os pulmões e que a comida passe pelo esôfago. Além destas funções, aprendeu-se a utilizar a laringe para converter o contínuo fluxo de ar dos pulmões em uma série de pulsos de ar.

No topo da laringe está localizada a epiglote, que durante a ingestão de comida, é de grande auxílio para o direcionamento desta para o esôfago.

A função de válvula da laringe depende grandemente das cordas vocais, que são ligamentos internos da laringe que se estendem de um lado ao outro. O espaço entre as cordas vocais é chamado de glote. Quando as cordas vocais são pressionadas umas contra as outras, a passagem de ar está completamente bloqueada e a válvula está fechada.

Acima das cordas vocais estão localizadas as falsas cordas vocais, que também se estendem de um lado ao outro da laringe. As opiniões quanto à função das falsas cordas vocais no processo da fala se divergem.

Com isto percebe-se que a laringe forma uma tripla barreira ao tubo vocal, através da ação da epiglote, das falsas cordas vocais e das cordas vocais. Durante a ingestão de comida, todas estão fechadas, durante a respiração estão todas abertas.

Durante a fala, a epiglote e as falsas cordas vocais permanecem abertas, enquanto que as cordas vocais ficam fechadas. A pressão do ar que vem dos pulmões vai aumentando até que as cordas vocais são afastadas e o ar é liberado como em uma explosão. As cordas vocais então voltam a sua posição original e o processo se repete. É a repetição deste processo que forma o conjunto de pulsos de ar audíveis antes mencionado. A freqüência da vibração das cordas vocais é determinada pela velocidade com que ocorre a explosão e pela velocidade com a qual as cordas voltam ao estado inicial, o que se deve a uma combinação de fatores.

Alguns destes fatores são a densidade das cordas vocais, o seu tamanho e tensão. Existe também o efeito de vácuo criado na glote pela explosão de ar passando por este pequeno espaço em direção a um espaço maior logo acima. Este efeito aumenta a velocidade com que as cordas vocais retornam à sua posição inicial. O aumento da pressão do ar que vem dos pulmões também aumenta a freqüência de vibração das cordas vocais.

Durante a fala, altera-se continuamente o tamanho e a tensão das cordas vocais, assim como a pressão do ar que vem dos pulmões, para se obter a freqüência desejada. A faixa de freqüências das cordas vocais normalmente utilizada durante uma conversação é de 60 a 350Hz, ou seja, mais de duas oitavas. Durante a fala de uma pessoa normalmente utilizam-se as cordas vocais em uma faixa de freqüências de somente uma oitava e meia.

Algumas medidas possibilitaram a determinação de como os pulsos de ar variam durante um ciclo. O espectro destes pulsos de ar tem muitas componentes, mas todas são múltiplos da freqüência de vibração das cordas vocais. A amplitude destas componentes normalmente decai com o aumento da freqüência. Quando se fala alto ou se grita, as cordas vocais se abrem e se fecham mais rapidamente, permanecendo abertas por um tempo menor, o que aumenta a amplitude das freqüências mais altas e dá ao som um aspecto mais áspero. É por causa disto que se distingue uma fala em voz alta ou um grito mesmo quando o volume do que foi pronunciado é reduzido.

Este conjunto de pulsos que é produzido pelas cordas vocais não é ainda a voz que se ouve. Estes pulsos têm sua estrutura alterada pelo trato vocal, que se estende desde a glote até os lábios (ou seja, laringe e boca), tendo um ramo composto pela cavidade nasal.

A faringe é a parte do trato vocal que une a laringe e o esôfago com a boca e o nariz. O seu formato é alterado quando se ingere algo, movendo-se a língua para trás ou a laringe para cima, ou ainda contraindo-se as paredes da faringe, alterações as quais também ocorrem durante a fala. O tamanho da cavidade da faringe também pode ser afetado pela posição da língua, alterando as freqüências de ressonância do trato vocal.

A cavidade nasal se estende desde a faringe até as narinas, e é dividida em duas seções separadas pelo septo nasal. Alguns sulcos e dobras na cavidade nasal transformam alguns segmentos do nariz em intrincadas passagens de ar. Na parte inferior do nariz estão as amígdalas, que ocasionalmente aumentam de tamanho de modo a influenciar a corrente de ar dos pulmões e conferir ao som um aspecto normalmente chamado de anasalado.

A última e mais importante parte do trato vocal é a boca, cujo tamanho e formato pode variar mais que qualquer outra parte do trato vocal, ajustando-se as posições relativas do palato, da língua, dos lábios e dos dentes.

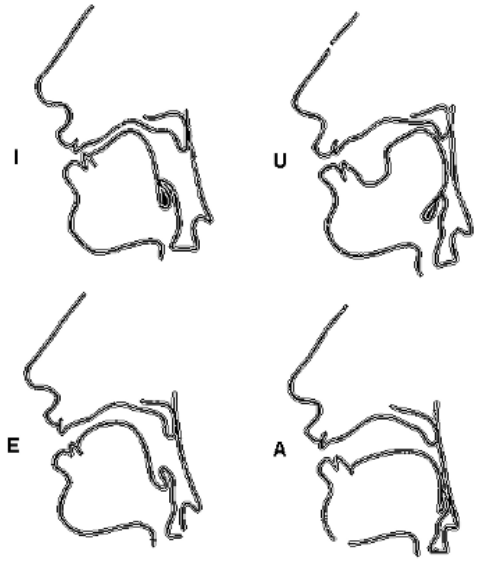
O mais flexível destes órgãos é a língua. Suas partes inferior, superior e central podem se mover independentemente, e todo conjunto pode se mover para frente, para trás, para cima e para baixo.

Os lábios afetam tanto o formato quanto o tamanho do trato vocal, podendo ser alongados ou arredondados em vários níveis e podendo também parar completamente o fluxo de ar.

Os lábios (junto com as bochechas) influenciam a fala de mais de um modo, não somente alterando o formato do trato vocal (e conseqüentemente o som produzido), mas em conjunto com os dentes, alterando o modo como se percebe o que se está falando, visto que são os únicos órgãos vocais normalmente visíveis. Os ouvintes podem adquirir grande parte da informação do que está sendo falado através da face, seja "lendo os lábios" de quem está falando ou seja entendendo as expressões de alegria ou tristeza. Normalmente se atribui menos importância a estas características do que elas têm.

Os dentes também podem afetar o trato vocal, podendo ser usados para restringir a passagem de ar colocando-se a língua ou os lábios próximo a estes (como ao se pronunciar a consoante v).

O último dos órgãos que moldam a cavidade da boca é o palato, que é o responsável pela passagem ou não da corrente de ar pela cavidade nasal.



Contorno do trato vocal durante a articulação de várias vogais

II.2 A Articulação da Língua Portuguesa

Ao se analisar uma língua pelo seu aspecto sonoro, considera-se sempre o que se denomina de **fonemas**. São fonemas as unidades sonoras básicas de uma língua, capazes de em união, descrever sonoramente todos os vocábulos desta.

A língua portuguesa apresenta o conjunto de fonemas descritos no quadro abaixo:

Símbolo	Exemplo	Transcrição Fonológica
/p/	**p**aca	/paka/
/b/	**b**ula	/bula/
/t/	**t**ara	/tara/
/d/	**d**ata	/data/
/k/	**c**ara, **qu**ero	/kara/, /kéro/
/g/	**g**ola, **gu**erra	/góla/, /géRa/
/f/	**f**aca	/faka/
/v/	**v**ala	/vala/
/s/	**s**ola, a**ss**a, mo**ç**a	/sóla/, /asa/, /mosa/
/z/	a**s**a, **z**ero	/aza/, /zéro/
/x/	me**ch**a, **x**á	/méxa/, /xa/
/j/	**j**aca, **g**ela	/jaka/, /jéla/
/m/	**m**ola	/móla/
/n/	**n**ata	/nata/
/ɳ/	ni**nh**o	/niɳo/
/l/	**l**ata	/lata/
/ʎ/	ca**lh**a	/kaʎa/
/r/	pa**r**a	/para/
/R/	**r**ota, ca**rr**oça	/Róta/, /kaRosa/
/a/	c**á**	/ka/
/é/	m**e**l	/mél/
/e/	s**e**da	/seda/
/i/	r**i**ca	/Rica/
/ó/	m**o**la	/móla/
/o/	t**o**la	/tola/
/u/	g**u**la	/gula/

Os sons produzidos durante a fala são divididos em dois grandes grupos: o de sons vozeados e o de sons não vozeados. Chama-se de sons vozeados aqueles em que as cordas vocais vibram durante sua produção e de sons não vozeados aqueles em que isto não ocorre.

Esta divisão quase que se confunde com a divisão entre vogais e consoantes, pois todas as vogais são sons vozeados e praticamente todas as consoantes são não vozeadas.

Durante a produção das vogais, o que as distingue, visto que todas produzem a vibração das cordas vogais é a posição da língua e dos lábios.

Esta classificação se dá em três níveis:

- quanto à zona de articulação, ou seja, de acordo com a região da boca em que se dá a maior elevação da língua, assim podendo ser **anteriores**, **centrais** e **posteriores**;
- pela elevação da região mais alta da língua; por este critério podendo ser altas, **médias** e **baixas**;
- quanto ao timbre, podendo ser **abertas** ou **fechadas**.

Sendo assim, classifica-se a vogal /a/ por exemplo como vogal oral, central, baixa.

A classificação de todas as vogais se encontra no quadro abaixo:

		Anteriores	**Centrais**	**Posteriores**
Altas		/i/		/u/
Médias	**Fechadas**	/e/		/o/
	Abertas	/é/		/ó/
Baixas			/a/	

As consoantes, diferentemente das vogais, se caracterizam não por produzirem a vibração das cordas vocais, mas por constituírem do modo como se cria obstáculos à passagem de ar dos pulmões antes da produção de uma vogal , constituindo-se basicamente de um "ruído".

Classificam-se então as consoantes do seguinte modo:

- Quanto ao **modo de articulação**: indica-se o tipo de obstáculo encontrado pela corrente de ar ao passar pela boca. São **oclusivas** aquelas produzidas com obstáculo total; são **constritivas** as produzidas com obstáculo parcial. As constritivas se subdividem em **fricativas** (o ar sofre fricção), **laterais** (o ar passa pelos lados da cavidade bucal) e **vibrantes** (a língua ou o véu palatino vibram);
- Quanto ao **ponto de articulação:** indica-se o ponto da cavidade bucal onde se encontra o obstáculo à corrente de ar. As consoantes podem ser **bilabiais** (os lábios entram em contato), **labiodentais** (o lábio inferior toca os dentes incisivos superiores), **alveolares** (a língua toca os alvéolos dos incisivos superiores),

palatais (a língua toca o palato duro ou céu da boca) e **velares** (a língua toca o palato mole, ou véu palatino);

- As consoantes podem ainda ser **surdas** ou **sonoras**, de acordo com a vibração das cordas vocais (as cordas vocais vibram já durante a produção da consoante, embora esta vibração seja decorrente da produção da vogal que será pronunciada em seguida), e ainda **orais** ou **nasais**, de acordo com a participação das cavidades bucal e nasal no seu processo de emissão.

Sendo assim, por exemplo classifica-se a consoante /v/ como consoante oral, constritiva, fricativa, labiodental, sonora.

A classificação de todas as consoantes se encontra no quadro abaixo:

Cavidades Bucal e Nasal		Orais						Nasais
Modo de Articulação		Oclusivas		Constritivas				
				Fricativas		Laterais	Vibrantes	
Cordas Vocais		Surdas	Sonoras	Surdas	Sonoras	Sonoras	Sonoras	Sonoras
Ponto de Articu-lação	Bilabiais	/p/	/b/					/m/
	Labiodentais			/f/	/v/			
	Linguodentais	/t/	/d/					/n/
	Alveolares			/s/	/z/	/l/	/r/	
	Palatais			/x/	/j/	/λ/		/ɳ/
	Velares	/k/	/g/				/R/	

II.3 Um Modelo Físico da Produção de Voz

Como visto, a produção de voz se inicia com a transformação, pelas cordas vocais, da corrente de ar vinda dos pulmões em pulsos de ar, que irão passar então por todo o trato vocal. Este trato vocal funciona então como todo tubo repleto de ar, ou seja, como uma cavidade ressonante. Sendo assim, esta cavidade ressonante possui certas freqüências naturais, às quais responde melhor que em outras freqüências.

Partindo-se destas considerações assume-se então que as cordas vocais produzam uma série de pulsos com freqüência de 100Hz como mostrados na figura II.3.1. O espectro de freqüências destes pulsos é composto de vários pulsos em freqüências múltiplas de 100Hz, os quais são aplicados à cavidade ressonante, que "molda" estes de acordo com sua própria curva de resposta em freqüência.

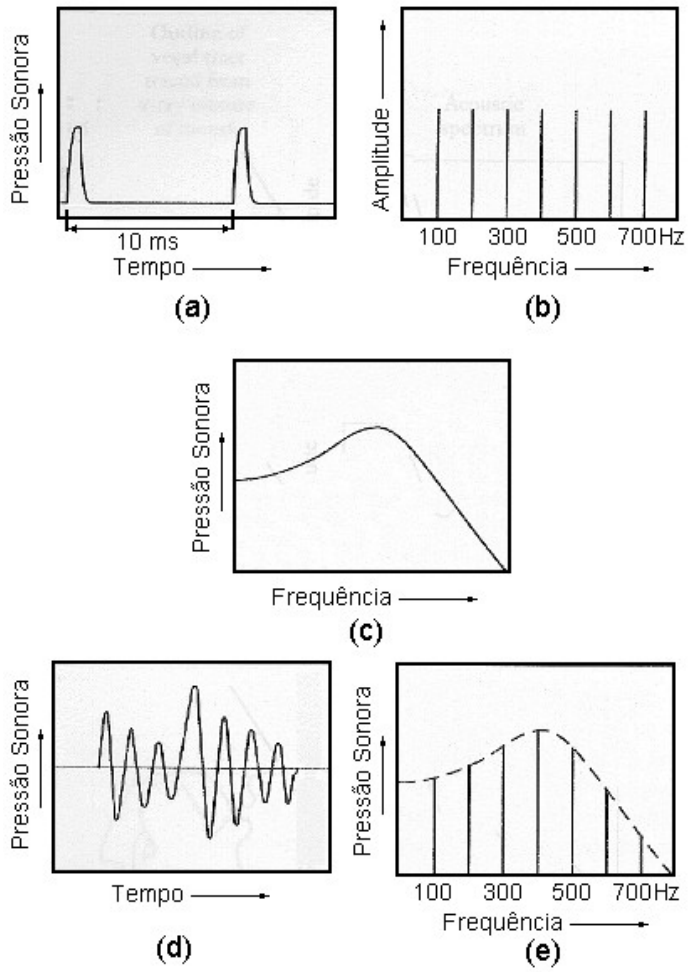
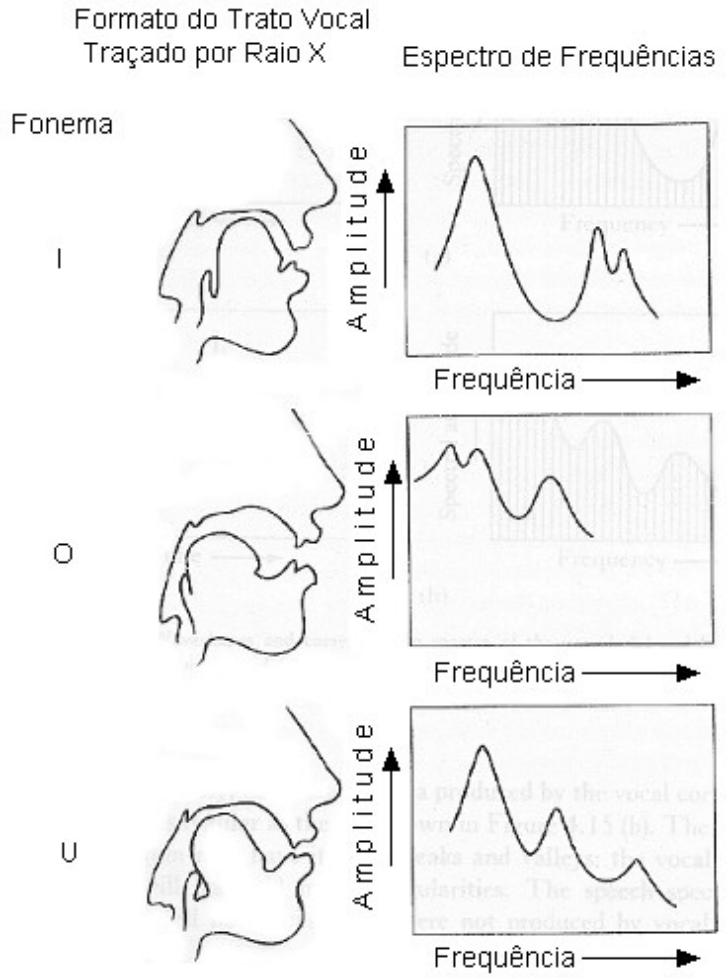


Figura II.3.1: (a) os pulsos de ar provenientes dos pulmões; (b) o espectro de freqüências dos pulsos de ar vindos dos pulmões; (c) resposta em freqüência de uma cavidade ressonante; (d) e (e) são respectivamente a forma de onda e a resposta em freqüência, respectivamente, do sinal produzido quando uma série de pulsos como em (a) é aplicado a uma cavidade ressonante como em (b).

A cavidade ressonante exemplificada na figura II.3.1 possui somente uma freqüência de ressonância, mas o trato vocal possui muitas. Estas freqüências de ressonância, as quais são determinadas pelo próprio formato do trato vocal, determinam a resposta em freqüência deste, e é esta resposta que "molda" o som final.

As freqüências de ressonância do trato vocal são denominadas ***formantes,*** e qualquer trato vocal tem seu próprio conjunto de formantes. Quando o palato mole esta erguido, fechando a cavidade nasal o trato vocal tem aproximadamente 17 centímetros, contando desde a epiglote até os lábios. Para um tubo deste tamanho, com seção uniforme, as principais freqüências de ressonância seriam 500Hz, 1500Hz, 2500Hz, 3500Hz e 4500Hz. Como normalmente a seção do trato vocal varia consideravelmente, os formantes não ficam uniformemente espaçados como em um tubo regular. O formante de menor freqüência é então

chamado primeiro formante, sendo os outros denominados segundo, terceiro e assim por diante em ordem crescente de freqüência.



Como se pode notar, as freqüências de ressonância do trato vocal não são necessariamente iguais a nenhum dos harmônicos dos pulsos de ar produzidos pelos pulmões e cordas vocais, o que se deve à completa independência dos órgãos que determinam estas freqüências.

Percebe-se, então, que na formação dos fonemas, existem dois conjuntos de parâmetros que são determinantes no som produzido: as freqüências de ressonância do trato vocal, ou seja, os formantes e a freqüência dos pulsos de ar produzidos pelo conjunto pulmões+cordas vocais.

Estes parâmetros são responsáveis tanto pela diferenciação entre um fonema e outro quanto pela diferenciação entre fonemas iguais produzidos por pessoas diferentes, os quais "soam" de forma diferente mas são identificados como o mesmo.

Os parâmetros que essencialmente diferenciam um fonema de outro são as freqüências de ressonância do trato vocal (formantes). Como o trato vocal não é igual de uma pessoa para outra, existindo peculiaridades a cada um (uma cavidade ou protuberância por exemplo), estas peculiaridades geram formantes que caracterizam e diferenciam uma voz da outra. Sendo assim cada fonema possui um conjunto definido de formantes, sendo a eles acrescidos formantes característicos de cada trato vocal.

A freqüência dos pulsos de ar que passam pelo trato vocal determina basicamente o quão "grave" ou "aguda" é uma voz. A voz dos homens normalmente é mais grave porque as cordas vocais masculinas são mais rígidas, produzindo pulsos de ar com uma freqüência menor, enquanto que com as mulheres acontece o contrário.

III. Métodos de Síntese de Voz

III.1 Introdução

Ao se observar e classificar os métodos de produção de voz, percebe-se que estes variam imensamente com relação à forma com que se aborda o problema. Isto se deve inicialmente a que alguns destes métodos foram criados no final da década de 70, início de 80, quando o armazenamento digital de informação e a velocidade de processamento eram ainda problemas.

Pode-se classificar as técnicas de produção de voz em dois grandes grupos:

- Síntese de voz a partir de análise de parâmetros
- Síntese de voz por edição de voz gravada

As técnicas de síntese de voz a partir de análise de parâmetros eram de grande utilidade, pois, apesar de exigirem um "hardware" específico e bastante complicado para a produção de voz, exigiam um pequeno processamento e pequena quantidade de dados armazenados. Atualmente estas técnicas continuam sendo empregadas, mas de outra maneira. O "hardware" específico foi substituído por técnicas digitais de filtragem e modelagem, permanecendo apenas a idéia inicial.

III.2 Síntese a partir de análise de parâmetros

As técnicas de produção de voz baseadas na análise de parâmetros procuram decompor a voz em parâmetros, utilizando a modelagem física da produção de voz humana abordada no capítulo anterior.

A voz tem como origem a excitação do trato vocal produzida pela corrente de ar vinda dos pulmões, que pode ser modelada por pulsos periódicos ou por um ruído branco. Em ambos os casos o trato vocal responderá a esta excitação podendo ser modelado com um filtro digital com três ou quatro freqüências de ressonância. Como a resposta do trato vocal às excitações não é constante no tempo, o filtro digital também deve ser variante no tempo.

Para este tipo de síntese de voz, então, deve-se estimar a partir da voz natural os seguintes parâmetros: os intervalos de sons sonoros e de sons surdos, o "pitch" (freqüência dos pulsos de ar produzidos pelas cordas vocais) dos sons sonoros, a intensidade do sinal de

voz e os parâmetros do filtro digital (mais especificamente as freqüências de ressonância deste filtro).

Durante a produção de voz estes parâmetros mudam, e análises da voz demonstram que é necessária a atualização destes valores cerca de 100 a 200 vezes por segundo.

Pode-se perceber que como são utilizados cerca de 10 parâmetros com um número de bits pequeno, a quantidade de informação a ser armazenada para a reprodução de cada som é pequena.

A partir deste conceito básico, foram criados vários métodos de obtenção destes parâmetros a partir da voz natural. Um destes métodos, por exemplo, considera que a voz é a convolução de dois componentes, um representando a função de excitação do trato vocal e o outro correspondendo à resposta deste trato. Sabendo-se que a componente relativa à resposta do trato vocal varia lentamente com a freqüência e que a componente relativa à excitação do trato vocal varia rápida e periodicamente com a freqüência, estas duas componentes podem ser separadas através de filtros apropriados (passa-baixas para a obtenção da resposta ao impulso do trato vocal e passa-altas para a a obtenção da excitação).

Este tipo de síntese atualmente tem como grande vantagem a possibilidade de se alterar os aspectos básicos da voz, como sua entonação e cadência.

III.3 Síntese por edição de voz gravada

A síntese por edição de voz gravada consiste no método mais simples de produção artificial de voz. Inicialmente este método consistia simplesmente em um banco de palavras gravadas, que eram reproduzidas de modo a formar frases. Este método, embora extremamente rudimentar ainda é utilizado, como por exemplo em secretárias eletrônicas e saldos automáticos de agências bancárias. A limitação deste método, é claro, está no banco de palavras armazenadas, que para a Língua Portuguesa, por exemplo, seria composto de algo em torno de 100.000 palavras. Além disto, a fala produzida desta forma “soa” extremamente artificial.

Atualmente as técnicas que se baseiam neste método procuram dividir as palavras em unidades menores, que juntas formem todo o universo de palavras que o sintetizador se propõe a falar.

Estas técnicas se dividem em dois grupos:

- Síntese por fonemas
- Síntese por difones

A síntese de voz por fonemas propõe a divisão das palavras nas suas unidades sonoras básicas, os fonemas. Sendo assim, o banco de sons seria composto pelo som de cada um dos possíveis fonemas da língua que juntos formariam a palavra desejada. Por exemplo, a palavra *chave* seria formada pela junção dos fonemas /x/+/a/+/v/+/e/.

Para a codificação das palavras em fonemas é utilizado um banco de regras específico à lingua que o sintetizador se propõe a falar. Estas regras inclusive permitem a flexibilidade de se falar com estilos diferentes, como por exemplo falar como um carioca ou como um paulista. De acordo com o falar de um carioca, um banco de regras deveria associar o fonema /x/ à letra s em final de palavra, enquanto que de acordo como falar paulista o fonema deveria ser /s/.

Como se pode perceber, também, a quantidade de informações armazenadas para a síntese de toda uma língua é pequena.

O problema deste tipo de síntese é que, analisando-se o espectro da voz, percebe-se que a quase totalidade da energia de uma palavra se encontra nas vogais, consistindo as consoantes apenas no modo como se passa de uma vogal para outra. Este fato dificulta muito a inteligibilidade das consoantes quando armazenadas em separado, principalmente por formarem grupos com características semelhantes (consoantes constritivas, fricativas, constritivas, vibrantes, etc).

É exatamente este problema que a síntese de voz por difones tenta resolver. Nesta técnica, as palavras são divididas em unidades maiores, que se confundem até com as sílabas. Deste modo não se armazena uma consoante em separado, mas sempre acompanhada de uma vogal. Por exemplo a palavra *chave* seria formada pelos difones /xa/+/ve/.

Esta alteração no método implica primariamente em se utilizar um banco de sons maior, que tem agora de ser composto de todas as possíveis combinações de fonemas existentes na língua que o sintetizador se propõe a falar.

Além deste aumento no banco de sons, agora deve-se, além de transcrever as palavras em fonemas, transcrever estes fonemas em difones, de modo a utilizar o banco de sons.

As implementações de sintetizadores de voz utilizando estes métodos de síntese por edição de vos gravada têm produzido resultados satisfatórios, principalmente se considerarmos a inteligibilidade da voz produzida. Normalmente esta possui uma boa

inteligibilidade, tendo o método como ponto fraco apenas apouca flexibilidade no que diz respeito a como "soa" a voz produzida, que normalmente "soa" de modo artificial.

IV. O Sintetizador

IV.1 Introdução

Dentre os métodos apresentados escolheu-se fazer a síntese de voz por edição de voz gravada, pois a idéia do projeto é fazer um sintetizador genérico capaz de ser incorporado aos mais diferentes projetos. Desta forma utilizou-se como "hardware" padrão um computador padrão PC 486 com 8 MB de memória e Windows, que é atualmente a plataforma mais difundida entre os usuários comuns, além de uma placa de som.

Como características básicas do sintetizador foram selecionadas: capacidade de let um texto de tamanho indefinido, fazer a transcrição texto/som em "real time", ou seja, sem a necessidade de um processamento antes da fala do texto, e capacidade de ser utilizável em diferentes tipos de "hardware" (diferentes marcas de placas-mãe, diferentes tipos de placas de som etc).

Devido à ampla gama de possibilidades de tipos de implementação do sintetizador e também ao pouco conhecimento efetivo de como se comporta a inteligibilidade de uma palavra construída de diversas maneiras, foram feitos vários testes iniciais que ajudaram a determinar como proceder à implementação.

IV.2 Testes Iniciais

A primeira dúvida a surgir foi com que qualidade se deveria gravar a voz para que ainda fosse possível a sua utilização no banco de sons do sintetizador. O método utilizado para determinar isto foi fazer a gravação de uma mesma palavra várias vezes com uma determinada qualidade. Posteriormente separavam-se vários trechos de som destas várias pronúncias e se os uniam de modo a formar a palavra original, testando-se então o quanto se poderia manipular os trechos de som e ainda assim utilizá-los no banco de sons.

Após estes testes foi determinada como qualidade de gravação de som uma amostragem a 11kHz com 8 bits de quantização.

Após isto determinado, precisava-se determinar como seria este banco de sons: se por fonemas ou difones. Primeiramente testou-se um banco de sons por fonemas, por ser este mais simples. Para este teste montou-se um banco de fonemas reduzido, de modo a se poder formar algumas palavras simulando o que faz o sintetizador (estas palavras foram formadas

unindo-se manualmente em um editor de arquivos .wav todos os arquivos correspondentes aos sons). Todos os bancos de sons utilizados, tanto na fase de teste quanto no sintetizador final foram construídos manualmente, ou seja, para se conseguir o som /pa/ grava-se por exemplo a palavra *apagado* e retira-se manualmente o trecho de som correspondente ao som "pa".

Com este pequeno banco de fonemas montado verificou-se que a inteligibilidade das palavras produzidas era muito baixa pois perdia-se completamente a informação contida na articulação de um som para outro, o que na prática influencia em muito o som final.

Com esta conclusão testou-se então a síntese por difones, montando-se um banco de sons que desta vez é composto por arquivos contendo trechos de som correspondentes aos finais de som + transições + inícios de som. Por exemplo, a palavra *pato* seria composta neste banco pelos seguintes trechos de som:

1. silêncio + início da consoante p
2. final da consoante p + início da vogal a
3. final da vogal a + início da consoante t
4. final da consoante t + início da vogal o
5. final da vogal o + silêncio.

Com esta estrutura não se perde a informação contida na transição de um som para outro, pois a união dos arquivos de som se dá no meio de uma consoante ou no meio de uma vogal, que se caracterizam por ter comportamento consoante nestes trechos.

Os testes com este banco demonstraram um ganho muito grande na inteligibilidade das palavras produzidas, tendo como pontos negativos os seguintes fatos: grande aumento no número de arquivos do banco de sons e maior dificuldade na confecção do banco de sons. Esta maior dificuldade é proveniente da constatação de que para cada grupo de sons, deveria-se ter arquivos de som com tamanhos diferentes, ou seja, não se pode fazer o "corte"do som de maneira igual para todos os sons. Por exemplo para se formar o som /ata/ deve-se unir os arquivos correspondentes aos sons:

1. silêncio + início da vogal a
2. final da vogal a + início da consoante t
3. final da consoante t + início da vogal a
4. final da vogal a + silêncio

Neste exemplo a transição entre os arquivos 2 e 3 deveria ocorrer exatamente no meio da consoante t, o que na prática não é verdade, pois esta consoante tem toda sua informação contida em um intervalo de tempo muito pequeno. Sendo assim deve-se nestes casos testar o melhor ponto de "corte" do som, que no caso deixa a maior parte da informação do som da consoante t no arquivo 3 em detrimento do arquivo 2.

Ao se terminar os testes para definir o tipo de banco de sons a ser utilizado, procurou-se também perceber a importância que a tonicidade com que a palavra é pronunciada influencia na inteligibilidade desta, já que até este ponto todos os testes haviam sido feitos com toda a palavra sendo pronunciada com uma mesma intensidade, sem acentuar sua sílaba tônica. Analisando-se isto percebeu-se inicialmente que, ao contrário do que instintivamente nos é proposto, a sílaba tônica não é assim caracterizada pela amplitude do som desta (que instintivamente deveria ser maior), mas sim pelo prolongamento da vogal desta sílaba por aproximadamente 1/3 do tempo normal das outras sílabas. Sendo assim, para caracterizar uma sílaba tônica basta inserir um trecho de som equivalente a este tempo.

Os testes com o acréscimo da tonicidade demonstraram também uma grande melhora na inteligibilidade das palavras sintetizadas.

Com o tipo de banco de sons definido surge então mais um problema de definição de implementação: como fazer a reprodução do som final. Analisando-se o banco de sons percebe-se que este é composto de arquivos que variam de 0,12 segundos a 0,22 segundos, o que significa que não pode haver praticamente nenhum intervalo entre a reprodução dos arquivod de som (um nível aceitável de intervalo seria de aproximadamente 1 décimo do tamanhjo médio dos arquivos que corresponde a aproximadamente 0,017 segundo ou 17 ms).

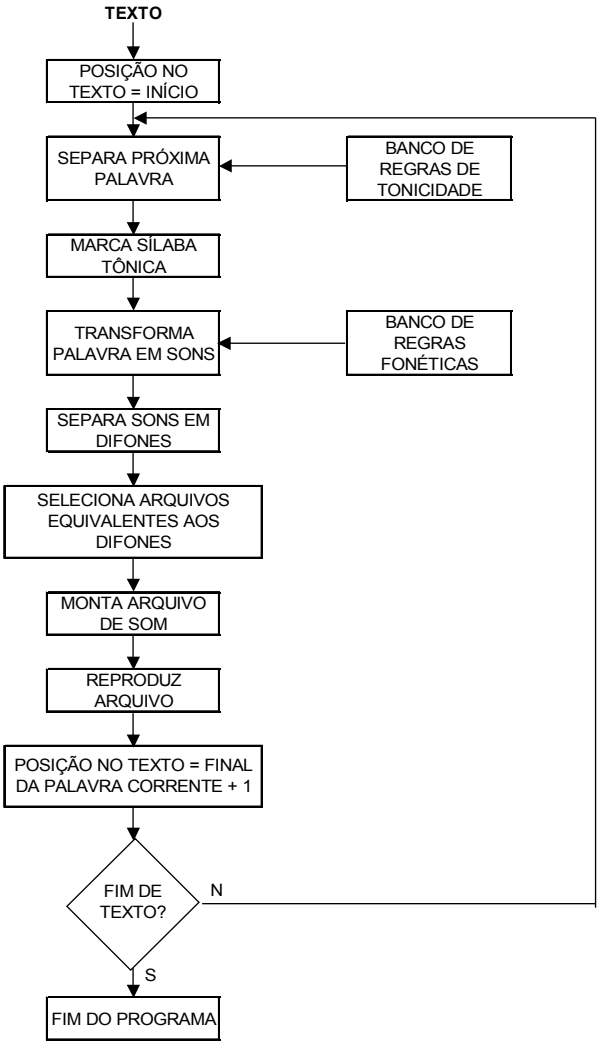
A idéia inicial era de carregar todos os dados necessários ao sintetizador em memória (inclusive o banco de sons) e a partir destes fazer toda a síntese da voz. Sendo assim, iria-se simplesmente reproduzir os arquivos de som necessários à reprodução da palavra um após o outro. Este método não se mostrou satisfatório, mesmo quando se fez o acionamento dos "drivers" da placa de som diretamente, minimizando ao máximo o tempo entre a reprodução de um arquivo de som e outro. Tendo isto em vista, optou-se por manter os arquivos do banco de sons no disco rígido, carregando em memória apenas os arquivos necessários para a reprodução da palavra corrente. Com estes arquivos, monta-se, então, um outro arquivo único contendo toda a palavra a ser pronunciada e, então, reproduz-se este arquivo. Este método se mostrou satisfatório, já que o tempo entre a reprodução de um arquivo e outro não mais existe

e o tempo necessário para a confecção do arquivo final contendo toda a palavra é pequeno o suficiente para que não atrapalhe a continuidade do texto a ser falado. Além disto tinha-se a preocupação de que com este método houvesse um grande número de acessos ao disco, acessos estes quase que ininterruptos durante a reprodução falada do texto, o que não ocorreu, devido ao fato dos sistemas operacionais fazerem um bom tratamento de "cache" de disco, pré-armazenando estes dados de grande utilização em memória.

IV.3 Implementação

Para o desenvolvimento do sintetizador foi escolhida a linguagem Delphi versão 2.0 da Borland, que se baseia na estrutura da linguagem Pascal, só que orientada a objetos e em ambiente Windows.

A estrutura do programa pode ser resumida como:

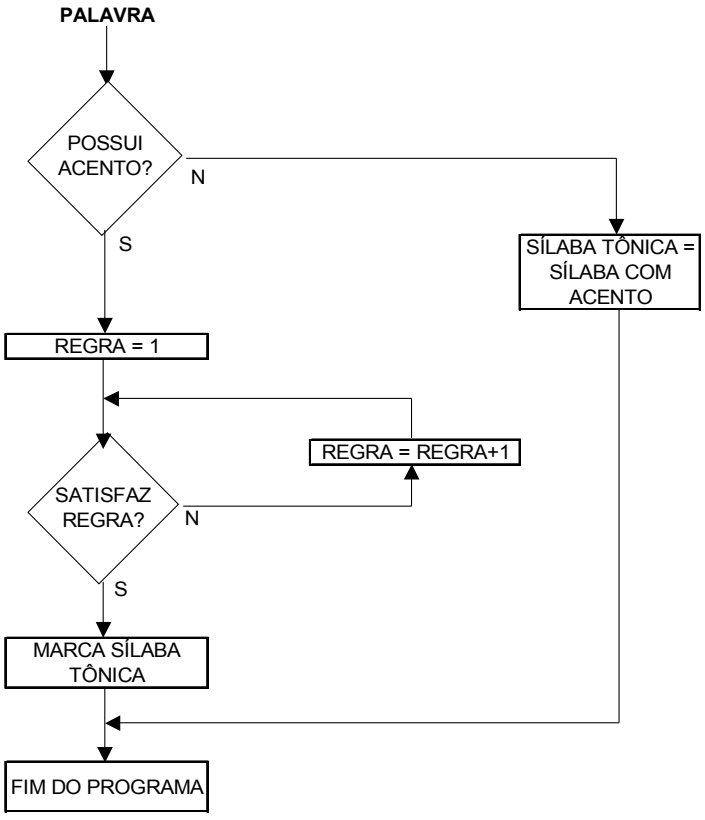


Entre as estruturas apresentadas acima destacam-se o banco de regras de tonicidade e o banco de regras fonéticas, que nada mais são que os conjuntos de regras que determinam qual será a sílaba tônica da palavra e qual o som de cada letra, respectivamente.

Estas estruturas são importantes porque são elas as únicas responsáveis pelo sintetizador falar certo ou errado uma palavra. Várias das regras são estabelecidas apenas pelo senso comum, não havendo nenhuma regra formal na Língua Portuguesa para esta. Este senso demonstra-se bastante razoável quando implementado. Para se testar a eficácia deste, basta escrever em um papel uma palavra que não exista e pedir para várias pessoas lerem esta palavra, constatando-se que a maioria das pessoas vai ler a palavra de uma mesma maneira, seguinto um certo "conjunto de regras" implícitas.

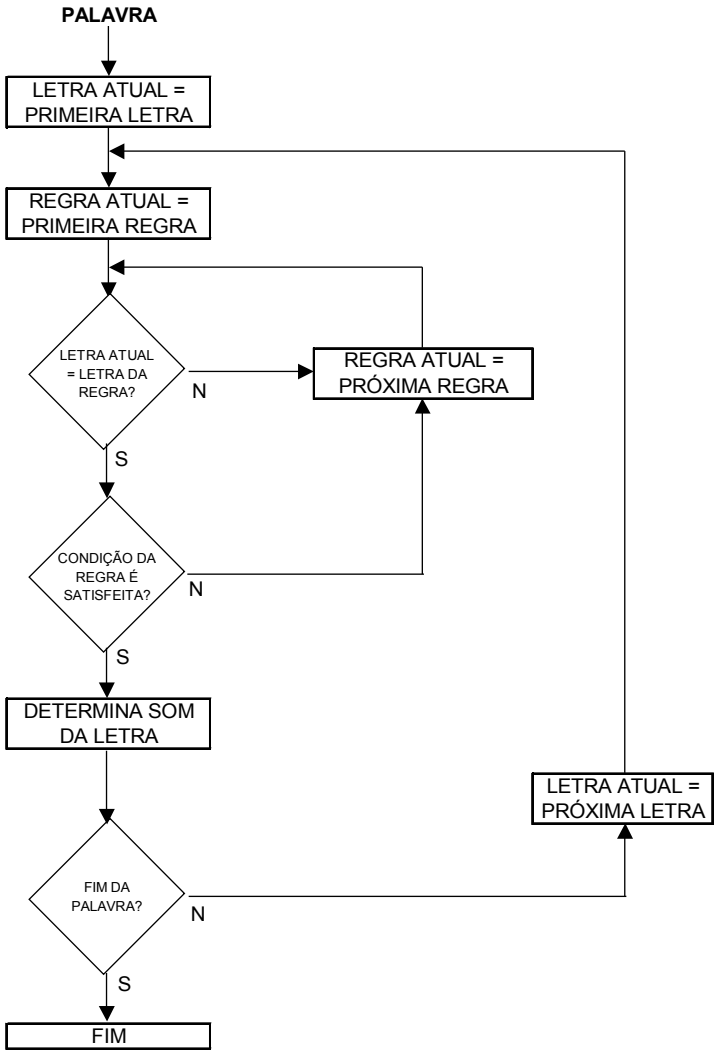
A utilização de somente um banco de regras deste tipo não é suficiente para apresentar 100% da eficácia, já que existem várias exceções. Uma maneira de contornar estas exceções seria a criação de um banco de exceções que seria consultado antes de utilizarem as regras padrões.

O banco de regras de tonicidade funciona da seguinte forma:



No caso das regras de tonicidade, estas foram criadas a partir das regras de acentuação da Língua Portuguesa ao contrário. Na Língua Portuguesa existem palavras proparoxítonas (a sílaba tônica é a antepenúltima), paroxítonas (a sílaba tônica é a penúltima) e oxítonas (a sílaba tônica é a última). Uma as regras de acentuação diz que todas as palavras proparoxítonas devem ser acentuadas, portanto todas as regras do banco de regras já estão restringidas às palavras paroxítonas e oxítonas. É utilizando-se as regras de acentuação desta forma inversa que se montou o banco de regras.

Quanto ao banco de regras fonéticas, este foi criado empiricamente, fazendo-se testes com várias frases e percebendo a necessidade de adaptação de algumas regras, eliminação de outras ou criação de novas regras. A estrutura de análise deste banco está descrita no diagrama abaixo.



O banco de regras está disposto de tal modo que todas as regras para uma determinada letra estão consecutivas sendo a última regra a que satisfaz a qualquer situação.

O banco de regras utilizado é composto das 88 regras listadas abaixo:

1. A letra A sucedida de M em final de palavra tem som de Ã
2. A letra Â sucedida de M em final de palavra tem som de Ã
3. A letra A sucedida de N em final de palavra tem som de Ã
4. A letra Â sucedida de M em final de palavra tem som de Ã
5. A letra A sucedida de NH tem som de Ã
6. A letra A sucedida de M tem som de Ã
7. A letra A sucedida de N tem som de Ã
8. A letra Á tem som de A
9. A letra Â tem som de Ã
10. A letra Ã tem som de Ã
11. A letra À tem som de A
12. A letra A tem som de A
13. A letra B tem som de B
14. A letra C sucedida de H tem som de X
15. A letra C sucedida de E tem som de S
16. A letra C sucedida de I tem som de S
17. A letra C tem som de C
18. A letra Ç tem som de S
19. A letra D tem som de D
20. A letra E sucedida de M tem som de Ê
21. A letra E sucedida de N tem som de Ê
22. A letra É tem som de E
23. A letra Ê tem som de Ê
24. A letra E tem som de Ê
25. A letra F tem som de F
26. As letras GU sucedidas de E têm som de G
27. As letras GU sucedidas de I têm som de G
28. A letra G sucedida de E tem som de J
29. A letra G sucedida de I tem som de J

30. A letra G tem som de G
31. A letra H não tem som
32. A letra Í tem som de I
33. A letra I tem som de I
34. A letra J tem som de J
35. As letras LH tem som de ! (fonema LH)
36. A letra L precedida de N e sucedida de vogal tem som de L
37. A letra L precedida de R e sucedida de vogal tem som de L
38. A letra L precedida de S e sucedida de vogal tem som de L
39. A letra L precedida de vogal e sucedida de vogal tem som de L
40. A letra L sucedida de vogal tem som de L
41. A letra L tem som de U
42. A letra K tem som de C
43. A letra M precedida de A em final de palavra tem som de U
44. A letra M tem som de M
45. As letras NH têm som de @ (fonema NH)
46. A letra N tem som de N
47. A letra O sucedida de M tem som de Ô
48. A letra O sucedida de N tem som de Ô
49. A letra Ó tem som de O
50. A letra Ô tem som de Ô
51. A letra O tem som de Ô
52. A letra Õ tem som de Ô
53. A letra P tem som de P
54. As letras QU sucedidas de E têm som de C
55. As letras QU sucedidas de I têm som de C
56. As letras QÜ têm som de CU
57. As letras QU têm som de CU
58. As letras RR têm som de R
59. A letra R em início de palavra tem som de R
60. A letra R precedida de S tem som de R
61. A letra R precedida de N tem som de R

62. A letra R precedida de vogal e sucedida de consoante tem som de R
63. A letra R tem som de & (fonema "*ére*")
64. As letras SS Têm som de S
65. A letra S em início de palavra tem som de S
66. A letra S precedida de B tem som de S
67. A letra S precedida de P tem som de S
68. A letra S precedida de N tem som de S
69. A letra S precedida de R tem som de S
70. A letra S em final de palavra tem som de S
71. A letra S tem som de Z
72. A letra T tem som de T
73. A letra Ú tem som de U
74. A letra Ü tem som de U
75. A letra U tem som de U
76. A letra V tem som de V
77. A letra W tem som de V
78. A letra X em final de palavra tem som de CS
79. A letra X precedida de Ê e sucedida de vogal tem som de Z
80. A letra X precedida de E e sucedida de vogal tem som de Z
81. A letra X precedida de A em início de palavra e sucedida de vogal tem som de CS
82. A letra X precedida de O em início de palavra e sucedida de vogal tem som de CS
83. A letra X precedida de A e sucedida de I tem som de S
84. A letra X precedida de O e sucedida de I tem som de S
85. A letra X tem som de X
86. A letra Y tem som de I
87. A letra Z em final de palavra tem som de X
88. A letra Z tem som de Z

Quando estas regras são aplicadas, obtém-se uma "pseudo-palavra" que contém apenas a informação de tonicidade e sons da palavra original. A partir desta então, é que se deve identificar quais difones deverão ser utilizados.

O banco de difones utilizado é composto de 374 arquivos de som, os quais possuem a seguinte estrutura de nomes: PRIMEIRO SOM + SEGUNDO SOM + TONICIDADE. Por exemplo, o difone ***sé*** é representado pelo arquivo SEAGU.WAV (S+E+AGUDO). Esta forma de representação simplifica a equivalência entre o difone e o arquivo, não sendo necessário consultar uma tabela de equivalências.

Após a identificação de todos os difones estes são carregados em memória, construindo-se um único arquivo de som composto da união de todos. Este arquivo então é reproduzido.

Ao chegar-se a esta fase, teoricamente o sintetizador estaria terminado, passando então a uma fase de testes finais. Logo no início destes testes percebeu-se que as palavras que possuíam os encontros consonantais envolvendo das letras R e L não estavam com uma boa inteligibilidade. Percebeu-se então que estes encontros consonantais possuem uma estrutura de formação bastante parecida com o encontro de uma vogal e uma consoante, ou seja, a transição entre um som e outro possui grande quantidade de informação. Esta informação estava sendo passada à palavra pronunciada pelo sintetizador de forma "quebrada". Por exemplo, na palavra ***macro*** o encontro consonantal cr tinha toda sua informação passada apenas pelo difone que o precedia e pelo que sucedia, ou seja, pelos difones ***ac*** e ***ro***.

Desta forma resolveu-se fazer mais uma pequena alteração no projeto, apenas em sua fase final. Ao invés de traduzir os conjuntos de sons ***acro*** da palavra macro nos difones ac+ro, traduzi-los em ***ac***+***cro***. Isto implica o aumento do banco de difones, que agora deverá possuir os "sons" BR, CR, DR, FR, GR, PR, TR, BL, CL, FL, GL, PL com todas as possíveis vogais, o que corresponde a 96 arquivos de som, aumentando o banco de difones para 470 arquivos.

IV. Resultados

V.1 Introdução

Fisicamente a fala de uma palavra é simplesmente uma onda de pressão sonora com certas características. Estas características são: intensidade, duração, freqüências fundamentais e outras, as quais são facilmente medidas. Quando, no entanto, se pensa no que é percebido pelo ouvido quando esta palavra é pronunciada, os parâmetros a serem considerados são difíceis de ser determinados.

Não existe uma relação direta entre os parâmetros físicos de uma palavra e o que importa neste para a inteligibilidade da palavra. O estabelecimento destas relações já se mostrou muito difícil, quando se do que se pode ser estabelecido por um senso comum. Sendo assim o que se faz é realizar testes com um conjunto de palavras com características semelhantes e conhecidas (como palavras faladas pelo telefone, palavras pronunciadas em locais cheios de ruído, palavras de um sintetizador e outras situações em que as palavras por um motivo ou outro não estão com todas as suas características originais), estabelecendo critérios de avaliação.

Normalmente estes testes são de três tipos:

- Testes de Articulação:

Este tipo de teste pretende quantificar o grau de inteligibilidade da fala em diferentes situações. Normalmente, este é realizado pronunciando-se algumas palavras ou frases a algumas pessoas, que têm de escrevê-las, repeti-las ou até responder questões sobre estas. A medição consiste na quantidade de palavras ou frases que são corretamente percebidas.

O resultado deste tipo de teste depende grandemente do conjunto de palavras utilizado, que normalmente é composto de 20 a 40 itens. Neste conjunto devem aparecer todos os possíveis sons da língua, ocorrendo com a mesma freqüência relativa que na fala cotidiana. Este tipo de conjunto de palavras é o que se chama de palavras ***foneticamente balanceadas***.

Deve-se perceber também que o resultado obtido com o teste com palavras é bem diferente do com frases, sendo este último o mais adequado a se classificar o grau de inteligibilidade. Isto ocorre porque, no nosso próprio dia a dia, estamos acostumados a

identificar certas palavras apenas pelo sentido geral de uma frase. Alguns testes realizados demonstram que se pode obter uma conversação normal com um índice de reconhecimento de palavras na ordem de 50%.

Esta compensação de alguns sons ou até mesmo palavras pode ser facilmente percebido com um experimento visual, observando-se a seguinte frase:

A CR**NÇ* F*L* M**T*

Nesta frase, emobra várias letras estejam faltando, a maioria das pessoas percebe que está escrito A CRIANÇA FALA MUITO. O que ocorre com a fala é semelhante.

- Testes de Qualidade:

Neste tipo de teste procura-se avaliar a qualidade do som que se produz, ou seja, quão natural a fala aparenta. Esta naturalidade pode ser traduzida em um som sem "ruídos" estranhos, sem aparência "robotizada" ou passando as emoções condizentes com o que está sendo falado.

Este teste é realizado apenas perguntando a opinião das pessoas que escutam a fala em diversas situações, e atribuindo a esta uma nota em um determinado intervalo.

- Testes de Compreensão:

Ao se ouvir um trecho pronunciado por um sintetizador de voz, pode-se entender todas as palavras pronunciadas e ainda assim não se entender o sentido do trecho. Isto ocorre porque se despende muito tempo entendendo quais palavras foram pronunciadas, sem sobrar tempo para entender o sentido da frase.

Os testes que pretendem medir este parâmetro são realizados colocando-se algumas pessoas para ouvir um trecho de cerca de um minuto, após o que são realizadas algumas perguntas sobre o que foi dito.

V.2 Resultados Obtidos

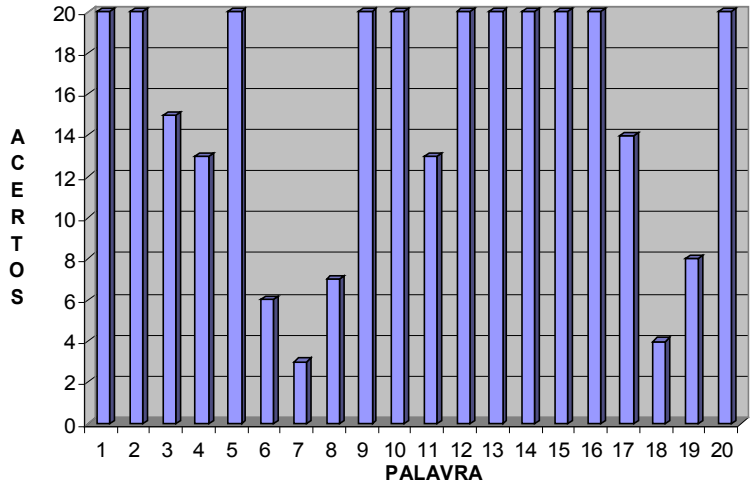
1. Teste de Articulação

Este teste foi realizado tanto com palavras como com frases. Para o teste com palavras foram utilizadas as seguintes palavras:

1. ABACATE
2. MORANGO
3. CARRO
4. ZUMBI
5. SALADA
6. DIDÁTICA
7. GRUTA
8. FILHO
9. PALHAÇO
10. GAROTO
11. CANHÃO
12. VIOLETA
13. JARRO
14. LAPISEIRA
15. NOTÍCIA
16. XÍCARA
17. LACRADO
18. PLÁSTICO
19. CADERNO
20. PROFISSÃO

Todos os testes foram realizados com um grupo de 20 pessoas. Para o teste de articulação com palavras, estas foram pronunciadas pelo sintetizador e cada pessoa repetia a palavra pronunciada. Foram considerados apenas dois tipos de resultados: acertos e erros. Foi

considerado como acerto toda vez que repetia-se corretamente logo após o sintetizador pronunciar a palavra uma só vez. O próximo gráfico representa os resultados.



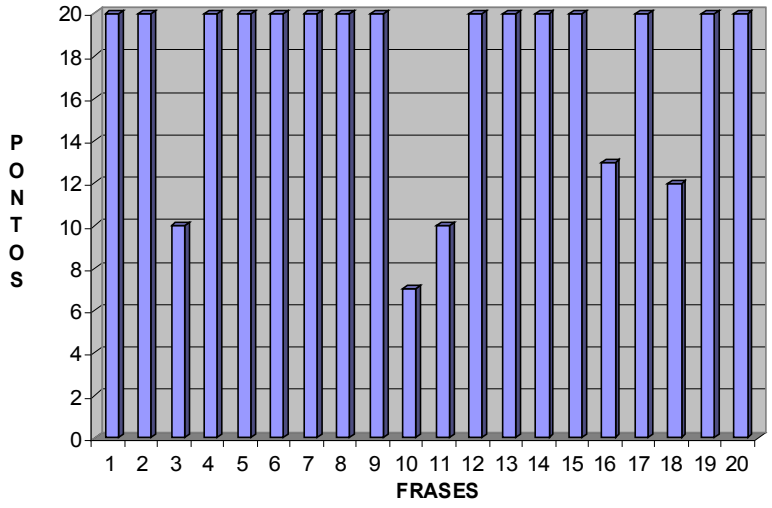
Estes resultados representam uma taxa de acertos de 75,75%.

Para o teste com frases foram utilizadas as seguintes frases:

1. A MENINA ESTÁ BONITA.
2. O PROFESSOR DEU UMA BELA AULA HOJE.
3. TODOS GOSTARAM DA APRESENTAÇÃO DOS MÚSICOS.
4. A ELEIÇÃO PARA PREFETO É AMANHÃ.
5. A MATÉRIA DO JORNAL FOI BASTANTE DISCUTIDA.
6. O AUDITÓRIO FICOU LOTADO PARA O COMUNICADO DO REITOR.
7. VILLA LOBOS FOI UM DOS MELHORES COMPOSITORES DE TODOS OS TEMPOS.
8. O SUPER-HOMEM É UM DOS PRIMEIROS HERÓIS DOS QUADRINHOS.
9. O MEU CHEFE FOI ALMOÇAR COM O PRESIDENTE DA EMPRESA.
10. A MINHA VOVÓ COZINHA DIVINAMENTE BEM.
11. A MINHA CASA ENTRARÁ EM OBRAS NA PRÓXIMA SEMANA.
12. A PROFESSORA FOI ELOGIADA POR TODOS.
13. TODOS FICARAM SURPRESOS COM SUA HUMILDADE.
14. A MÚSICA DO NORDESTE ESTÁ CADA VEZ GANHANDO MAIS FORÇA.
15. O BRASIL ESTÁ ENTRANDO EM UMA NOVA FASE.

16. BRASÍLIA FOI CONSTRUÍDA POR JUSCELINO.
17. O FUTEBOL REALMENTE É A PAIXÃO DOS BRASILEIROS.
18. ENTRE OS ESPORTES, O TÊNIS E O GOLFE SE DESTACAM COMO OS MAIS ELEGANTES.
19. A LUA GIRA EM TORNO DA TERRA.
20. PARECE QUE AGORA TUDO VAI BEM.

Para o teste de articulação com frases, estas foram pronunciadas pelo sintetizador e cada pessoa repetia toda a frase pronunciada. Foram considerados apenas três tipos e resultados: acertos completos (quando a frase foi repetida completamente correta), acertos parciais (quando a frase foi repetida com no mínimo 70% de acerto) e erros (quando a frase foi repetida com mais de 30% de erro). O próximo gráfico representa os resultados, onde os acertos completos foram considerados como 1 ponto e os parciais como 0,5 ponto.



Estes resultados representam uma taxa de acertos de 88,12%.

2. Teste de Qualidade

Ao grupo de 20 pessoas utilizado para o teste foi apresentado o seguinte questionário:

Atribua notas variando de 0 a 10 às seguintes perguntas:

1. Quanto o som do sintetizador parece poluído por ruídos?
2. Quanto a voz do sintetizador parece “robotizada”?
3. O quanto você acha que a voz do sintetizador é identificável, ou seja, possui características próprias bem definidas?
4. Qual é a sua nota para a voz do sintetizador em geral?

Para estas perguntas obteve-se as seguintes médias:

1. Média: 3,4
2. Média: 4,2
3. Média: 7,3
4. Média: 8,5

3. Teste de Compreensão

Este último teste compreendeu em fazer cada uma das 20 pessoas ouvirem o seguinte trecho:

“Atualmente o Brasil vem passando por períodos que nunca foram vistos em sua história. Depois da implantação do plano de estabilização econômica o governo agora tem que fazê-lo continuar. Muitos falam que estes planos podem ser afetados na época das eleições, quando historicamente todos os políticos só pensam em se reeleger.”

Após isto, pediu-se que fosse explicado sobre que o texto falava. Pôde-se perceber que todas as pessoas entenderam o que o texto pretendia passar, embora fosse relatado por todos que realmente algumas palavras do texto não foram entendidas por falta de tempo aliada à falta de entonação, já que o sintetizador não enfatiza as partes importantes das frases, o que normalmente ocorre em uma fala normal.

V.3 Conclusões

Os resultados dos testes acima comprovam que o resultado final foi bastante satisfatório. Estes resultados demonstram que o sintetizador pode ser utilizado sem qualquer restrição para aplicações que exijam frases não muito longas.

O nível de acertos do teste de articulação com palavras e com frases demonstra que o nível de inteligibilidade do sintetizador é mais que suficiente para estabelecer um diálogo normal. O acréscimo de dificuldade de entendimento para textos longos se deve principalmente à falta de "emoções" durante a fala. Percebe-se claramente que estas "emoções" são de grande importância na percepção de textos, já que naturalmente faz-se uma seleção do que realmente é importante e do que não é. Estas "emoções" consistem tanto na entonação natural ao se fazer uma pergunta ou uma exclamação, quanto no aumento de intensidade da fala em trechos de importância.

Quanto aos aspectos de aplicabilidade, o projeto se mostra grandemente versátil, já que não exige nenhuma plataforma especial, e também pode ser utilizado junto com outras aplicações, já que consome muito pouco em termos de processamento, memória e acesso a disco. Como foi feito na linguagem Delphi, esta permite que o programa seja compilado na forma de uma DLL (Dinamic Link Library), ou seja, transformando-se em uma função capaz de ser utilizada em qualquer outro programa.

Como melhoramentos ao projeto, pode-se analisar como se comportam estas "emoções" que são passadas ao texto, de modo a tornar o sintetizador mais "amigável" e possível de ser utilizado em aplicações aonde a naturalidade e fluência em textos sejam fundamentais.

VI. Bibliografia

- Sound Blaster – O Livro Oficial
  Peter M. Ridge, David M. Golden, Ivan Luk e Scott E, Sindorf
  Makron Books

- The Speech Chain – The Physics and Biology of Spoken Language
  Peter B. Denes, Elliot N. Pinson
  W. H. Freeman and Company
  New York, New York

- Um Sistema de Síntese de Voz por Difones do Idioma Português
  Alexis de Souza Esquivel
  Tese M. Sc. COPPE UFRJ 1984