MAURO MENDES BATAN

PAULO TADASHI UKEI

SÍNTESE DE VOZ PARA O PORTUGUÊS BRASILEIRO

Projeto de formatura apresentado à Escola
Politécnica da Universidade de São Paulo

São Paulo
2001

MAURO MENDES BATAN
PAULO TADASHI UKEI

SÍNTESE DE VOZ PARA O PORTUGUÊS BRASILEIRO

Projeto de formatura apresentado à Escola Politécnica da Universidade de São Paulo

Orientador:
Prof. Doutor
Jorge Almeida Rady

São Paulo
2001

RESUMO

Este trabalho realizou a implementação de um sistema de síntese de voz a partir da introdução de um texto qualquer. A idéia básica consiste em varrer esse texto, originalmente em português brasileiro, e transforma-lo em som. Cada palavra é transformada separadamente em uma seqüência de fonemas com sua entonação e tempo.

Por sua vez, essa seqüência de fonemas é reconhecida na biblioteca de dífones. Que são segmentos que serão colados juntos para formar uma cadeia que possa ser pronunciada pelo computador através de uma placa de som.

ABSTRACT

This project was implemented a speech synthesis system from the introduction of any text. The basic idea consists of sweeping this text, originally into Brazilian Portuguese, and transforms it into sound. Each word, is transformed separately in a sequence of phonemes with its entonation and time.

In turn, this sequence of phonemes is recognized in the library of dífones. That they are segments that will be glue together to form a string that can be pronunciated for the computer through a sound board.

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

1. INTRODUÇÃO

1.1 Descrição De Projeto

A idéia básica é fazer um programa de computador que converte um texto em fala e pronuncia. Para isso será necessário atender varias especificações que serão definidas num próximo documento. As vantagens de se ter um pronunciador de texto são inúmeras. Qualquer um pode entender a mensagem sem treinamento ou concentração intensa. A mensagem pode ser recebida mesmo quando o usuário está envolvido em outras atividades, como andar, carregar objetos ou observar algo. Ou ainda o telefone convencional que pode ser utilizado para acesso remoto a informações.

Algumas das especificações que o programa deverá atender para que a síntese de voz tenha qualidade, ou melhor, seja inteligível são:

- Normalização de texto
- Pronúncia de palavra
- Prosódia
- Conexão de segmentos

A "**normalização de texto**" converte uma string como "João foi para casa." para uma série de palavras, "joão", "foi", "para", "casa", junto com um marcador que indica que um período aconteceu. Porém, isto se torna mais complicado quando strings como "João foi para casa a 150 km/h" onde "150 km/h" é convertido a "cento e cinqüenta quilômetros por hora".

E necessária a utilização de um arquivo de dicionário. Este arquivo conteria uma lista de palavras e suas respectivas pronúncias em fonemas.

A **pronúncia de palavra** é responsável em transformar a palavra em uma seqüência de fonemas. Assim como vemos nos dicionários.

Fazendo uma análise sintática na frase, podemos então definir a **prosódia** da frase. Com a devida entonação, duração (para dar o ritmo), entre outras características.

E por fim temos a **conexão dos segmentos**. O que não passa de "colarmos" juntos fonemas pré-gravados no formato WAV. E então com os segmentos já conectados, passamos por uma série de transformadas que deverão aplicar a prosódia para poder então ser pronunciado.

1.2 Introdução A Síntese De Fala

Foi preciso fazer uma pesquisa muito detalhada sobre os métodos existentes e mais utilizados. Pois não era nossa intenção desenvolver uma técnica totalmente nova mas sim adequar alguma existente para a língua portuguesa brasileira. Há o constante estudo das interfaces homem máquina para tornar o uso da máquina menos sofrido e mais natural possível. Uma destas interfaces é a fala. Tanto o reconhecimento quanto a síntese da fala. Mas fazer a máquina falar como um humano é complicado. Existem casos onde palavras ou até mesmo frases completas são pré-gravadas para serem reproduzidas pela máquina. Mas o ideal em que trabalhamos é o caso de se falar qualquer palavra que possa ser escrita. Não havendo assim a necessidade de armazenar todas as palavras do dicionário pré-gravadas. Mas para que isso seja possível, é preciso a utilização de algoritmos para conversão de texto para fala.

A síntese de fala é objeto de estudo em universidades dentro e fora do país há anos. A figura a seguir mostra os avanços com o passar dos anos.

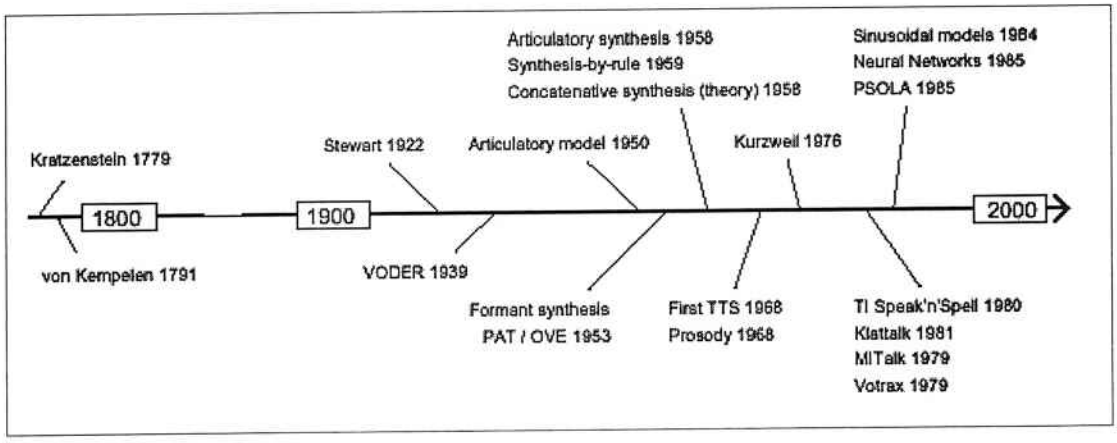


Figura 1 - Histórico da sintese de fala.

Não abordarei aqui aspectos da fisiologia humana uma vez que não é escopo do trabalho. Os algoritmos de conversão podem ser desmembrados em módulos para facilitar a abordagem na implementação.

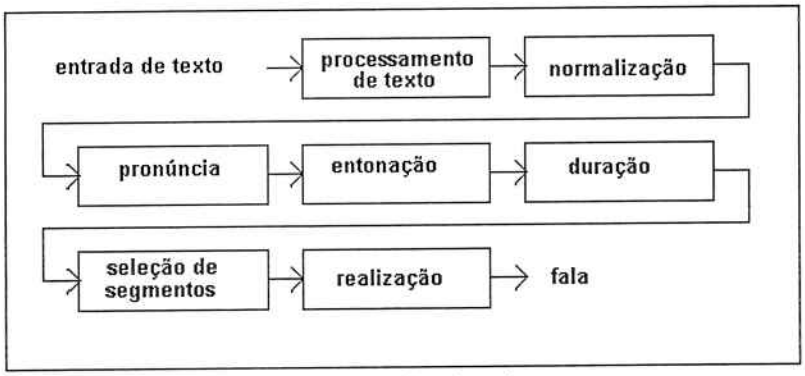


figura 2 - módulos de um sintetizador de voz.

1.3 Objetivos

O objetivo básico deste trabalho é apresentar a implementação de uma estrutura de "software" capaz de reproduzir um texto por meio de técnicas conhecidas para a síntese de voz.

Para isso há a necessidade da utilização de um conjunto de regras de transcrição de letras para fonemas, que permite efetuar a transcrição fonológica de palavras com base nas letras que a constituem. E também de técnicas de processamento de sinal para a síntese do sinal de voz.

1.4 Pesquisas Na Área

As primeiras tentativas de construir maquinas de produção de fala, embora sintetizando apenas 5 vogais, datam de 1779, por C. G. Kratzenstein. Poucos anos mais tarde, em 1791, W. R. von Kempelen demonstrou uma máquina muito mais sofisticada e capaz de produzir fala contínua, provando que o sistema humano de produção de fala podia ser modelado artificialmente. No mesmo ano publicava um livro descrevendo os seus estudos sobre produção de fala e as experiências de duas décadas até chegar a esta máquina.

Em 1835, Wheatstone demonstrou, na *Dublin Association for the Advances of Science*, uma máquina construída com base nos princípios descritos no livro de von Kempelen. Esta máquina usava um fole para fornecer ar a um ressonador feito de couro, sendo a sua secção alterada pela mão de um operador. A outra mão manipulava quatro comandos que geravam constrições de modo a produzir consoantes.

Mas já a partir da década de 60, o estudo de sistemas de síntese de voz a partir de texto teve um grande impulso, caracterizado por intensas pesquisas na área. Porém havia dúvidas sobre viabilidade de sistemas de síntese de voz.Iniciou-se então o desenvolvimento das primeiras regras de conversão de texto para fonemas.

Os anos 70 marcaram um novo período, com o aperfeiçoamento dos sintetizadores e dos algoritmos de síntese de voz. Dois centros de pesquisas se destacaram nessa área, o "Bell Laboratories" pertencente à "AT&T", e o "Massachussets Institute of Technology -MIT". No final da década de 70 e início dos anos 80, começaram a aparecer os primeiros protótipos de sistemas de síntese de voz a partir de texto com vocabulário ilimitado, com destaque para o "MITalk", desenvolvido no MIT durante a década de 70, sob a supervisão de Jonathan Allen.

Ainda nessa época surgiram os primeiros sistemas comerciais com vocabulário ilimitado, os quais vêm sendo continuamente aprimorados desde então e que ainda não tem um representante significativo para o português brasileiro. Dennis Klatt fez uma ampla descrição cronológica sobre a evolução de equipamentos vendidos comercialmente. O primeiro sistema comercial com vocabulário ilimitado foi uma

máquina de leitura para deficientes visuais, lançado em 1976, pela "Kurzweil", baseado no "chip Votrax SC-01", que era capaz de produzir voz a partir de material impresso. Dois anos depois, em 1978, surgiu o conversor de texto para voz "Type-n-Talk", da Votrax, baseado também no mesmo "chip". Entretanto, havia falta de inteligibilidade da voz gerada por esses dois sistemas.

Em 1982, a "Speech Plus Inc." lançou o "Prose-2000", sistema de conversão de texto para voz baseado no "MITalk". No mesmo ano, a "Street Eletronics" lançou o "Echo", sistema de baixo custo baseado no "chip TMS-5220", da "Texas Instruments".

Em 1983, a "Digital Equipment Corporation - DEC" lançou o "DECtalk", que teve origem no conversor de texto para voz "Klattalk", também desenvolvido no MIT por Dennis Klatt.

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".

Na segunda metade da década de 80, a "Berkeley Speech Technologies" apresentou o sistema "Text-to-Speech - TTS", 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. Desde então, esses sistemas vêm sofrendo constantes atualizações com a finalidade de aumentar a inteligibilidade e a naturalidade da voz produzida, e também, suportar novos idiomas. Por exemplo, um ano após o lançamento do "Prose 2000", já havia uma implementação para aceitar textos em espanhol (OLABE et al., 1983), e em 1989, o "MITalk" produzia saída de voz a partir de textos em japonês e chinês (JAVKIN et al., 1989).

Diversas outras pesquisas vêm sendo realizadas em âmbito mundial, como o sistema de síntese de voz a partir de textos em Chinês desenvolvido por LEE, TSENG e OUH-YOUNG (1989) e o sistema para textos em Árabe produzido por EL-IMAN (1989).

Além desses estudos, outros sistemas de síntese de voz de baixo custo têm surgido no mercado, como as placas para microcomputador padrão PC-AT "Sound

Blaster PRO" da "Creative Labs" e a "Mwave LS2000" da "IBM", acompanhadas do programa "Monolog" da "Creative Labs", capaz de fazer a conversão de textos em Inglês para voz, com qualidade bastante aceitável.

No Brasil, estudos de síntese voz tiveram 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 do Português. ESQUIVEL, em 1984, apresentou um sistema de síntese de voz em tempo real a partir de texto, no qual sinais adicionais eram acrescentados ao texto para a correta pronúncia de determinados sons.

Na Universidade de Campinas - UNICAMP, estudos sobre síntese de voz a partir de texto irrestrito foram realizados posteriormente (EGASHIRA, 1992). Foram feitos trabalhos sobre pré-processamento de texto com a finalidade de permitir a correta elocução de números, abreviaturas e caracteres não alfabéticos.

Mais recentemente, vem sendo comercializado um produto de baixo custo para auxílio à deficientes visuais, o "Dosvox", desenvolvido na Universidade Federal do Rio de Janeiro. 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.

2. CARACTERÍSTICAS DA VOZ

2.1 - Principais Características

A fala humana distingüe-se de outros sistemas simbólicos, como os gestos por exemplo, por ser segmentada em unidades menores que se apresentam em número finito para cada idioma e possibilitam recombinação de modo a expressar idéias diferentes, os fonemas. Combinado-se esses fonemas e variando a quantidade bem como sua ordem, somos capazes de ocasionar alterações no significado do som, formando assim uma palavra. Os fonemas são produtos do nosso aparelho fonador.

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.

O aparelho fonador humano é constituído pelas seguintes partes, indicadas na Figura 3

- 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 produção de som,

- a laringe, na qual se localizam as cordas vocais, que produzem a energia sonora utilizada na fala,

- e as cavidades supralaríngeas (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 supralaríngeas, são também conhecidas como trato vocálico.

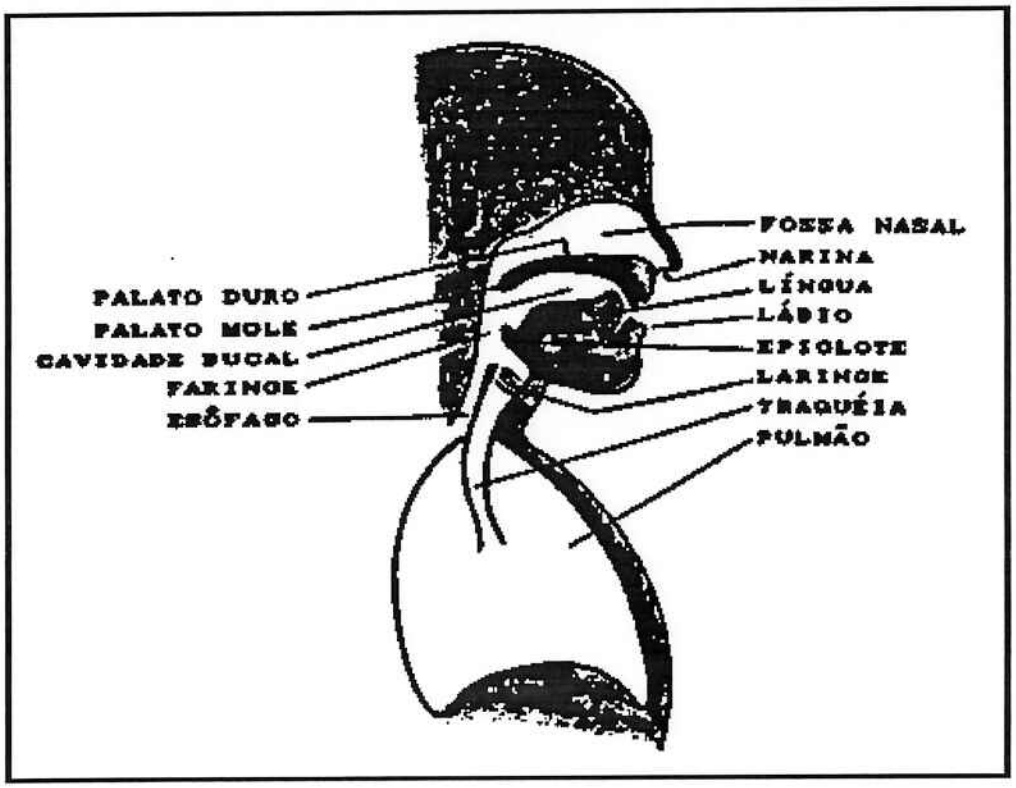


FIGURA 3 - Diagrama esquemático do aparelho fonador humano.

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, conforme ilustra o esquema da Figura 4 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.

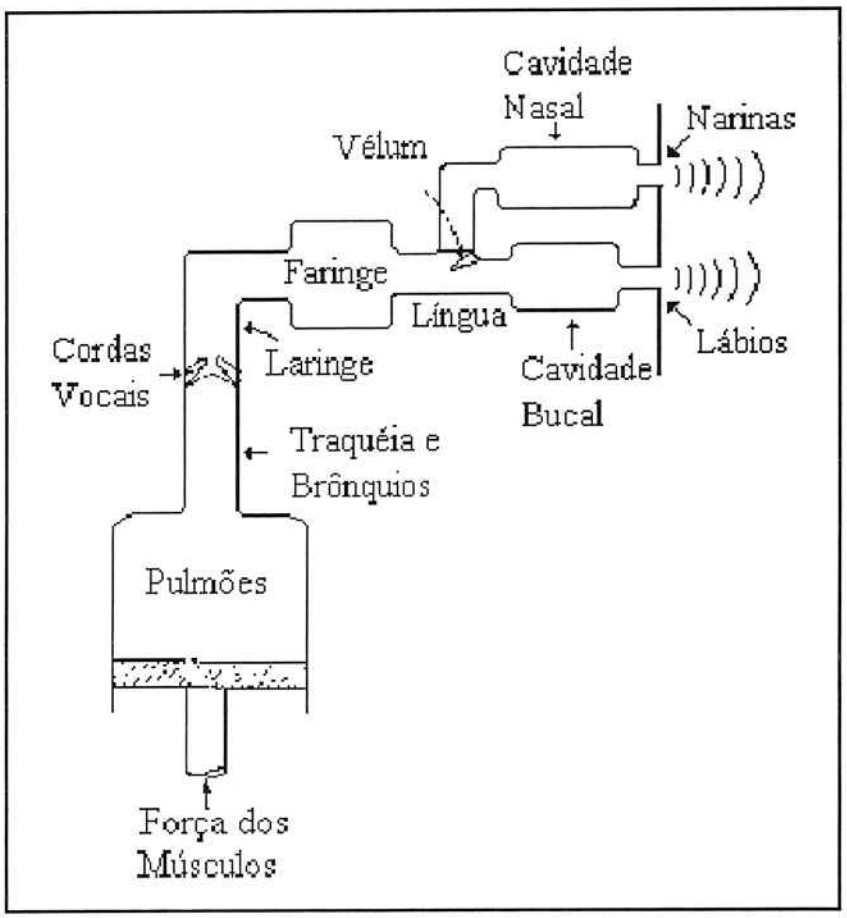


FIGURA 4 - Diagrama esquemático dos componentes funcionais do trato vocálico.

Um órgão essencial na produção de som é a faringe, 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 se pretende 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.

2.2 – A Voz

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 plosivos*. O modo como esses sons são produzidos foi descrito detalhadamente por vários autores.

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" (tom) 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 surdos*.

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.

À medida que os sons, gerados por qualquer uma das formas acima descritas, propagam-se pelo trato vocálico, apresentam alteração em seu espectro de freqüências e com ressonância em determinadas freqüências.

Estas freqüências são denominadas *freqüências formantes* do som, ou simplesmente *formantes*, 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 for 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 formantes estão aproximadamente 10% acima das formantes masculinas.

A estrutura das formantes é comumente representada através de espectrogramas sonoros, conforme exemplifica a Figura 5, na qual está indicado o espectrograma sonoro da frase "Noon is the sleepy time of the day", obtido por FLANAGAN et al. (1970), com suas três freqüências formantes representadas por linhas tracejadas. O eixo das abscissas corresponde ao tempo de elocução da frase, enquanto que o eixo das ordenadas corresponde às freqüências, sendo que os padrões escuros ocorrem nas freqüências com intensidade sonora.

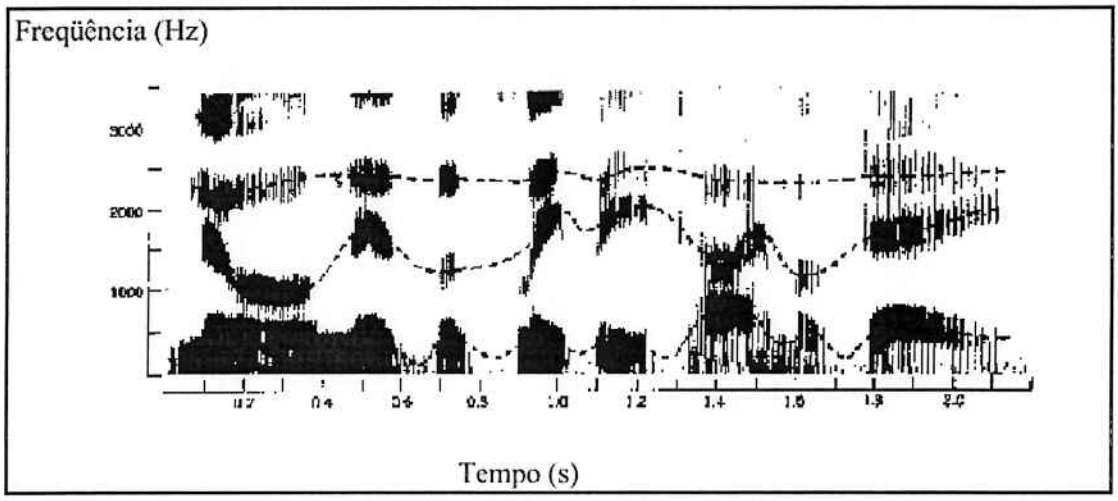


FIGURA 5 - Espectrograma da frase "Noon is the sleepy time of the day", com indicação das três primeiras formantes. (FLANAGAN et al., 1970)

2.3 – O Fonema

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. 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.

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.

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 Brasileiro 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.* Para distingüí-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 [djia].

Cada idioma tem seus próprios fonemas, que são elementos fônicos dotados de função representativa no sistema. A Língua Portuguesa tem 26 fonemas segmentais (19 consoantes e 7 vogais) e um fonema supra-segmental, o *acent*o, que não é um segmento e sim uma qualidade que se superpõe a certos segmentos. Formas como ***d**ívida* e *divid*a; ***s**ábia, sa**b**ia* e *sa**b**iá* opõem-se 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. No entanto, não há uma correspondência exata entre número de letras e o número de fonemas nos idiomas. Alguns exemplos:

- 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 na 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 *regue*, na qual o **u** é insonoro, para 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].*

Há um sistema ortográfico que rege 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 n o . 5.765 de 18 de dezembro de 1971, recentemente tem-se discutido 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).

No entanto, ainda segundo alguns autores, esse sistema integrado letra-fonema parece ser inviável, pois em um País com a dimensão 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.

Citam também que, quanto mais um idioma desenvolve-se, mais o sistema ortográfico afasta-se do sistema fonológico, como ocorre com os idiomas Inglês e Francês. Ainda com relação à simplificação abordada anteriormente, a representação do som [s] sempre pela letra "s" e do som [z] sempre pela letra "z" esbarra na questão das palavras homófonas como *coser/cozer, expiar/espiar, cessão/sessão/seção*, além de palavras como *aterrisar* e *subsídios*, para as quais existem normalmente duas pronúncias, *aterri[s]ar* e *aterri[z]ar*, *sub[s]ídios* e *sub[z]ídios*.

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

2.4 - O Alfabeto Fonético

Para simbolizar na escrita a pronúncia real de um som utiliza-se 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.

A Tabela 1 a seguir 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, que foi adotado neste trabalho, utilizando apenas um único caractere.

TABELA 1 - Alfabetos fonéticos para a Língua Portuguesa.

Símbolos IPA	Símbolos com 1 ou 2 caracteres (CAMPOS, 1980)	Símbolos com 1 caractere	Exemplos
a	A	a	**pá**, gato
e	E	e	**vê**, medo
ɛ	EH	é	**pé**, ferro
i	I	i	**vi**r, bico
o	O	o	**avô**, morro
ø	OH	ó	**avó**, cola
u	U	u	**tu**, bambú
ʍ	AN	ã	**lã**, cama
m	M	m	**mar**, amigo
n	N	n	**n**ada, cano
ŋ	NH	ñ	vi**nh**a, cami**nh**o
b	B	b	**b**ravo, am**b**os
p	P	p	**p**ai, ca**p**rino
d	D	d	**d**ar, an**d**ar
t	T	t	**t**u, can**t**o
g	G	g	fran**g**o, a**g**rado
k	C	k	**c**asa, **qu**e
f	F	f	**f**ilho, a**f**iar
v	V	v	**v**inho, u**v**a
s	S	s	**s**aber, pos**s**o
z	Z	z	ba**z**ar, ca**s**a
ʃ	X	x	**ch**over, **x**arope
ʒ	J	j	**j**á, **j**arra
l	L	l	**l**ado, ve**l**udo
l	L	u	al**t**o, fuzi**l**
ʎ	LH	L	fi**lh**o, pi**lh**a
r	R	r	ca**r**o, co**r**es
r	R	h	mar, carta
ɾ	RR	R	ca**rr**o, **r**oda

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

No entanto, um ditongo pode ser considerado como junção de duas vogais de menor duração com transições suaves entre as suas freqüências formantes. Portanto, no nosso trabalho não foi feita distinção entre o /i/ e o /u/ vogais ou semivogais.

Em posição final de sílaba ou palavra, a consoante "l" pode ser pronunciada como [u] ou [w], como em ***alto*** e *Bras**il***. Por essa razão, quando encontrado nessas posições o "l" será associado ao símbolo [u]. E nessa mesma situação a letra "r" será associada ao símbolo [h], conforme indicado anteriormente na Tabela 1. E é comum utilizar um apóstrofe (') para indicar a sílaba ou vogal tônica na transcrição, com em *['bo-la]* ou *[b'ola]*.

2.5 - CLASSIFICAÇÃO DOS SONS

A classificação dos sons da Língua Portuguesa foi discutida em detalhes por diversos autores.

Apesar de algumas divergências, é de consenso entre os autores a existência de duas classes de sons, as vogais e as consoantes, as quais serão destacadas a seguir.

2.5.1 - 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.

- Classificação quanto à Região de Articulação

Diz respeito ao ponto ou parte em que se dá o contato ou aproximação dos orgãos que cooperam para a produção dos fonemas, no caso das vogais, a língua e o pálato. 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], [é], [i]) levanta-se gradualmente a parte anterior da língua em direção ao pálato 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.

- Classificação quanto ao 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].

Depende da posição da úvula durante a passagem de ar pelo trato vocálico. Quando a corrente sonora é impedida de ressoar 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 acopladadas à cavidade bucal através do abaixamento da úvula, parte da corrente sonora ressoa na cavidade nasal, produzindo as *vogais nasais* ([ã], [en], [in], [õ], [un]).

- Classificação quanto à 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 idoma Português por um reforço da energia expiratória. As vogais que se encontram em sílabas não acentuadas denominam-se *átonas*. No idoma Português normal do Brasil, as vogais [é] e [ó] não aparecem em posição átona, assim como as vogais nasais.

CAMPOS (1980) mostrou que do ponto de vista da fonética acústica não há razão para considerar [en], [in], [õ], e [un] fonemas distintos, pois seus

espectros apresentam uma parte inicial idêntica ao das vogais que o iniciam, seguidos de uma parte muito semelhante a todos eles, que caracteriza o [m] final desses sons. Por isso, podem ser tratados como o encontro de dois fonemas, com transições suaves entre duas configurações do trato vocálico.

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]*.

2.5.2 - Classificação das Consoantes

São 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.

- Classificação quanto ao 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 constritivas*.

No segundo caso, dependendo de como a corrente expiratória escapa, as consoantes 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 pálato ([L]).

- Classificação quanto ao 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]), lingüodentais ([t], [d], [s], [z]), alveolares ([l], [r], [n]), palatais ([x], [j], [L], [ñ]) ou velares ([k], [g], [R]).

- Classificação quanto à 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.

- Classificação quanto ao 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]).

3. PROBLEMAS EM SINTESE DE FALA

3.1 Processamento De Texto

3.1.1 Normalização De Texto

O componente de "normalização de texto" converte de texto-para-fala qualquer entrada texto em uma série de palavras faladas. Trivialmente, normalização de texto converte uma cadeia como "João foi para casa." para umas séries de palavras, "joão", "foi", "para", "casa", junto com um marcador que indica que um período aconteceu. Porém, isto se torna mais complicado quando cadeias como "João foi para casa a 150 km/h" onde "150 km/h" é convertido a "cento e cinqüenta quilômetros por hora". Aqui é como trabalha a normalização de texto:

Primeiro, normalização de texto isola palavras no texto. A maior parte disto é tão trivial quanto procurar uma sucessão de caracteres alfabéticos, permitindo uma apóstrofe ocasional e hífen.

Normalização de texto então procura por números, horas, datas, e outras representações simbólicas. Estes são analisados e convertidos a palavras. (Exemplo: são convertidos "R$54.32" a "cinqüenta quatro reais e trinta dois centavos".) Alguém precisa codificar as regras para a conversão destes símbolos em palavras, desde que eles diferem dependendo do idioma e contexto.

Então, são convertidas abreviações, como "cm" para "centímetros", e "R." para "rua". Os normalizadores usarão um banco de dados que contém as abreviações e para o que elas são expandidos. Algumas das expansões dependem do contexto de palavras circunvizinhas.

O normalizador de texto poderia executar outras transformações de texto como endereços de internet. "http://www.usp.br" normalmente é falado como "w w w ponto uspi ponto br."

O normalizador terá regras que ditam se a pontuação causará uma palavra a ser falada ou se será silencio. (Exemplo: não são falados pontos ao término de orações, mas um ponto em um endereço de Internet é falado como "ponto".)

As regras variarão em complexidade que depende da máquina. Alguns normalizadores de texto são projetados para controlar convenções de e-mail como "Você * * * VAI * * * ao encontro. : - ("

Uma vez que o texto foi normalizado e foi simplificado em uma seqüência de palavras, é passado para o próximo módulo, disambiguação de homógrafo.

3.1.2 Disambiguação De Homógrafo

A próxima fase de texto-para-fala é chamada " disambiguação de homógrafo." Freqüentemente não é por si só uma fase, mas é combinado com o componente de normalização de texto ou com o de pronúncia. Estudaremos disambiguação de homógrafo em separado desde que não se ajusta completamente em nenhum dos outros componentes.

Máquinas de texto-para-fala usam uma variedade de técnicas para determinar as pronúncias. O mais robusto é tentar entender o que o texto está falando aproximadamente e decidir qual é o significado mais apropriado ao contexto. Uma vez sabido o significado , normalmente é fácil de propor a pronúncia certa.

Uma vez que os homógrafos foram disambigualizados, as palavras são enviadas à próxima fase para ser pronunciada.

Mas no português há a vantagem da acentuação, que serve como um diferenciador de palavras. Mas ainda temos palavras como gosto (substantivo) que tem pronúncia diferente do gosto (primeira pessoa do singular do presente do verbo gostar).

3.2 Pronúncia

O módulo de pronúncia aceita o texto, e produz uma seqüência de fonemas, igual à vista em um dicionário.

Para se ter a pronúncia de uma palavra, a máquina deve primeiro olhar em seu dicionário léxico. Se a palavra não se encontra, deve-se então reverter a uma lista de regras léxicas.

Regras de letra-para-som deduzem a pronúncia de uma palavra do texto. Eles são um tipo de inverso das regras de ortografia que foi ensinado na escola. Há várias técnicas para identificar a pronúncia, mas o algoritmo descrito aqui é facilmente implementado.

As regras de letra-para-som são treinadas em um dicionário léxico de pronúncias. O dicionário léxico armazena a palavra e a pronúncia, como:

Cachorro k ah x oh rr oh

Um algoritmo é usado para segmentar a palavra e descobrir qual som a letra "produz." Você pode ver que aquele "c" dentro de "cachorro" produz o fonema de "k", o "a" produz o fonema de "ah". Claro que, em outra palavra as letras individuais produzem fonemas diferentes. O "o" em "ótica" produzirá o fonema de "ó."

Para pronunciar "cachorro", as regras de letra-para-som tentam entender o som do fonema de "c" primeiro. Olha pela tabela de exceção por uma palavra que começa com "c" seguido por um "a", quando acha faz com que o som de "c" seja "c". Em

seguida, procura nas exceções um "a" cercado por um "c" e "c", quando achado produz um "ah". O resto das letras são controladas da mesma maneira.

Esta técnica pode pronunciar qualquer palavra, até mesmo se não estivesse no treinamento fixado, e faz uma suposição muito razoável da pronúncia, às vezes melhor que os humanos. Alguns algoritmos de letra-para-som tentam primeiro adivinhar de que idioma veio a palavra , e então usa jogos diferentes de regras para pronunciar cada idioma diferente.

Uma vez que as pronúncias foram geradas, estas são passados para a fase de prosódia.

3.3 Prosódia

Prosódia é o tom, velocidade, e volume com que são faladas as sílabas, palavras, frases, e orações. Texto-para-fala sem prosódia soa robotizada , e com texto-para-fala com prosódia ruim ,parece que está bêbado.

A técnica que as máquinas usam para sintetizar prosódia varia, mas há algumas técnicas gerais.

Primeiro, a máquina identifica o começo e o fim das orações. Em português, o tom tenderá a cair perto do fim de uma declaração, e sobe para uma pergunta. Igualmente, volume e rampa de velocidade da fala que sobe quando o texto-para-fala começa a falar, e cai na última palavra quando parar. São colocadas pausas entre orações.

Máquinas também identificam diferenças entre frase, como frases de substantivo e frases de verbo. Estas terão características semelhantes a orações, mas serão menos pronunciadas. A máquina pode determinar os limites de frase usando a

informação de parte-de-fala gerada durante o disambiguação de homógrafo. São colocadas pausas entre frases ou onde vírgulas acontecem.

Algoritmos tentam determinar quais palavras na oração são mais importante ao significado, e estas são enfatizadas. Palavras enfatizadas são mais altas, mais longas, e terão mais variação de tom. Palavras que são sem importância, são desenfatizadas. Em uma oração como a "João e Paulo caminharam para a loja", o padrão de ênfase poderia ser o "JOÃO e PAULO caminharam para a LOJA." Quanto mais a máquina de texto-para-fala "entender" o que está sendo falado, melhor será a ênfase.

Então, a prosódia dentro de uma palavra é determinada. Normalmente o tom e o volume são aumentados para dar ênfase as sílabas.

O tom, cronometragem, e informação de volume da oração, da frase, e palavra são combinados para produzir juntos a fala final. A produção do módulo de prosódia é uma lista de fonemas com o tom, duração, e volume para cada fonema.

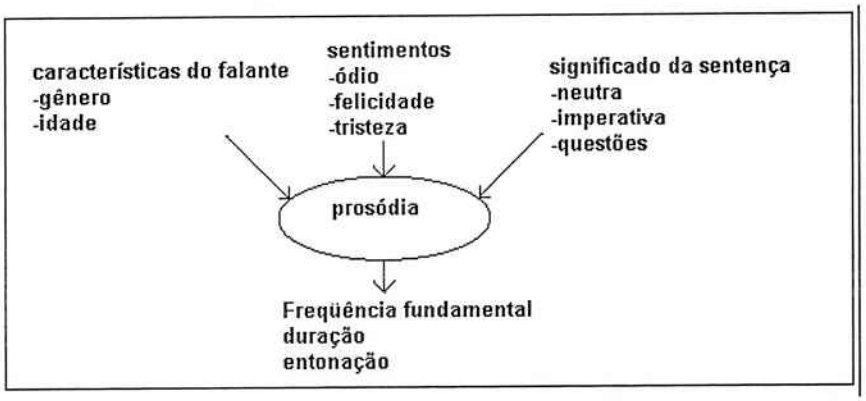


FIGURA 6 - fatores que influenciam a prosódia

4. MÉTODOS, TÉCNICA E ALGORITMOS

4.1 Síntese Por Formantes

Síntese por formante, que é o modelo de freqüência de polo do sinal da fala ou função de transferencia baseado no trato vocal.

- Freqüência fundamental (F0)
- quociente de excitação aberto (OQ)
- Grau da voz em excitação (VO)
- Freqüência dos formantes e amplitudes (F1...F3 and A1...A3)
- Freqüência de um ressonador de baixa freqüência (FN)
- intensidade de region de baixa- e alta-freqüência (ALF, AHF)

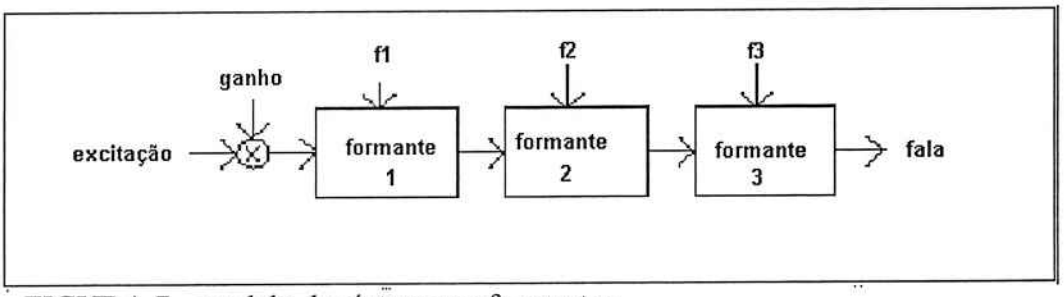


FIGURA 7 - modelo de síntese por formantes

Alguém que realmente merece ser mencionado neste caso é Dennis Klatt. Seu sintetizador de voz é amplamente utilizado em vários pacotes comerciais.

4.2 Síntese Por Concatenação

Síntese por concatenação, que usa trechos da fala natural gravada de diferentes comprimentos. É neste tipo de sintetizador que nos basearemos para desenvolver nosso programa, uma vez que não temos conhecimento muito profundo em DSP. Sendo assim a concatenação de fonemas pré-gravados uma idéia mais fácil de ser

implementada. Por outro lado, concatenação requer que os segmentos sejam processado para se alterar o tom, e a duração.

4.2.1 Método PSOLA

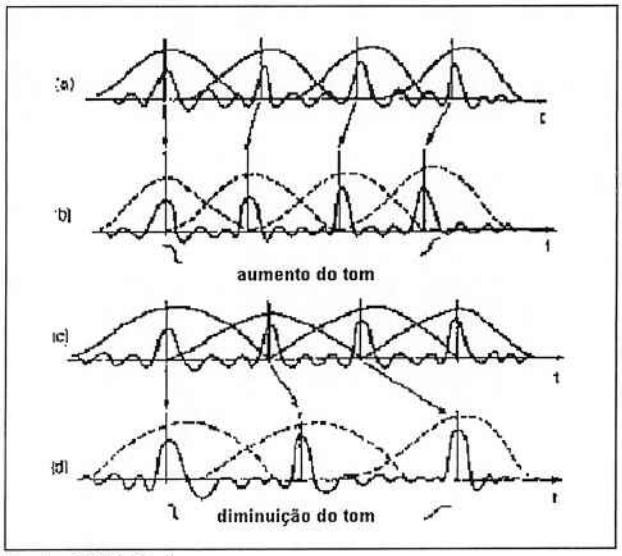


FIGURA 8 - método PSOLA

Os segmentos de voz são "janelados" (por exemplo Hanning) e depois concatenados. O janelamento serve para reduzirmos a introdução de sinais em espectros de freqüências mais altas quando cortamos um pedaço do sinal. Assim o janelamento atenua o sinal nas bordas da janela por uma cossenóide. Mas para que não percamos nenhuma parte do sinal, pois o estamos atenuando, as janelas devem ter uma sobreposição. O método PSOLA consiste em colarmos novamente as janelas cortadas, só que mais próximas ou mais distantes para a mudança do tom. Ou manter a distância mas repetir várias vezes o mesmo sinal para aumentar o tempo do sinal.

5. APLICAÇÕES DE FALA SINTÉTICA

5.1 Aplicações Para Cegos

É uma ferramenta muito útil pois possibilita que cegos tenham acesso ao computador. Arquivos, Internet, e-mail, textos,...

5.2 Aplicações Para Mudos

Possibilita que os mudos utilizem o sintetizador de fala para se expressar. Num telefone, numa palestra, ou até mesmo em casa

5.3 Aplicações Educacionais

Poderia ser de grande utilidade na ajuda a deficientes estudar um texto, ou uma criança que aprende a falar.

5.4 Aplicações Para Telecomunicações E Para Multimídia

Possibilita também uma melhor interface em jogos de computador, interação de qualquer um com a máquina. Tornando assim o computador em uma maquina mais natural e pessoal.

6. O NOSSO PROGRAMA

Para facilitar a implementação do programa ele foi dividido em módulos que foram implementados passo a passo. De modo geral nosso programa utiliza difones pré gravados que tem a prosódia implementados pelo algoritmo PSOLA.

6.1 Os Módulos

- **Processamento de texto:**

Onde há a conversão de um texto em uma cadeia de fonemas com suas características (tom, duração, amplitude)

- Normalização De Texto

É a fase onde todo o texto é verificado. Caso aja alguma abreviação essa é substituida por sua escrita por extenso. Não foi abordado o tratamento de números. Estes merecem um tratamento especial e dependem da analise sintática para que se possa determinar se é um numero de telefone onde deve ser pronunciado numero por numero ou uma quantia de dinheiro.

- Divisão silábica

A divisão silabica só é necessária para controle de prosódia. Pois determinando as sílabas tônicas depende da divisão silábica. O algoritmo que utilizamos não é capaz de determinar hiatos não acentuados.

- Determinação de sílaba tônica

Para o português é extremamente fácil se já se tem a divisão silábica, pois se baseia única e exclusivamente nas regras de acentuação. Sem as quais seria extremamente difícil determinar a entonação sem a analise sintática.

• Pronúncia

A transcrição de letras para fonema é feita por um algorítimo muito simples. Primeiramente iriamos implementar as regras de transcrição em um arquivo texto para que fosse facil sua edição, mas decidimos não fazer assim por questão de segurança. Então o algoritmo lê a letra, as letras anteriores, posteriores e decide qual fonema deve ser posto.

• Prosódia

A prosódia é um aspecto muito importante de qualquer sintetizador de fala. É ela que garante maior inteligibilidade ao programa. Como nosso programa não faz analise sintática, não podemos determinar a estrutura da frase não determinando a prosódia da frase. Mas foi atribuida entonação à palavra.

• **Sintese da voz:**

Onde a cadeia de fonemas é transformada em um arquivo de som (wave).

• Segmentação

A segmentação é responsável por determinar, a partir da cadeia de fonemas e dos difones gravados, quais são os difones que devem ser postos para ser "colados" juntos antes de ser pronunciados.

• Duração

Depende do fonema e da prosódia. É a duração de cada fonema na fala. É calculado um valor diferente para cada fonema.

• Tom

Depende, como a duração, do fonema e da prosódia. O tom é escolhido de forma a não alterar muito o sinal de voz original. A tecnica que utilizamos não permite uma variação maior a 10%.

• Concatenação

De posse dos segmentos, da duração e da entonação podemos então juntar os segmentos em apenas um para ser enviado à placa de som. Para juntar os segmentos foi utilizada uma janela de Von Han de forma a reduzir a introdução de sinal nas frequencias mais altas. Permitindo assim uma transição mais suave.

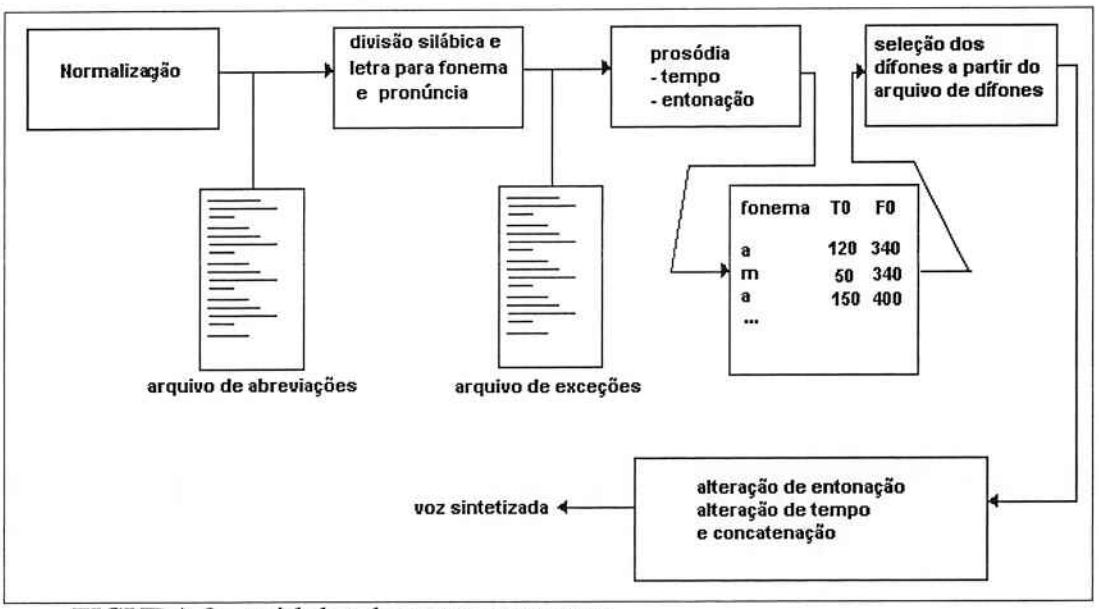


FIGURA 9 - módulos do nosso programa

Fases da implementação (sintese da voz)

❑ 1ª fase (preparação da base de fonemas)

• Geramos todas as combinações de fonemas em difones que precisavamos para nosso programa trabalhar adequadamente. Ex. (consoantes-consoantes-vogais, vogais-vogais, consoantes-vogais,).

• Criamos as palavras para serem gravadas e depois selecionados os difones para serem isolados. Ver tabela 1.

• Cortamos os difones e gravamos separadamente.

• Normalizar em amplitude baseando-se na vogal. Para isso foi preciso elaborar um aplicativo de manipulação de wave's para organizar o banco de fonemas.

• Para a implementação deste aplicativo, utilizamos java. Cada difone a ser gravado é uma instancia de um objeto. Contendo o sinal de voz gravado, a posição de divisão entre os fonemas que compoem o difone e a marcação dos períodos. Com esse aplicativo foram precisas oito horas para gerar uma biblioteca de quatrocentos e cinquenta e um difones. Nos produtos disponíveis no mercado geralmente são precisos quatro dias.

❑ 2ª fase (processamento de texto)

• Entrada de texto: o produto final se parece com um editor de texto bem simples. Onde é digitado o texto. Porém o objetivo maior e futuro é a geração de uma ferramenta de sintese de voz para ser posto em qualquer programa.

• Para o processamento são analisadas frases isoladamente. Cada frase tem sua pronúncia determinada isoladamente das outras frases. Sempre havendo uma pausa entre as frases. Mas primeiro o programa lê todas as palavras do texto armazenando-as em uma cadeia.

• Agora então, se analisara para saber se existe alguma abreviação comparando a frase selecionada com um arquivo de abreviações. A abreviação será substituída pela sua transcrição por extenso.

• Podemos agora comparar as palavras com um arquivo de exceções. Se as palavras existem nesse arquivo, já se seleciona sua pronúncia armazenada e esta palavra não passa mais pelas outras fases de processamento de texto.

• Neste ponto passamos a trabalhar nas palavras isoladamente uma por uma. Primeiro por meio de regras definimos a divisão silábica. As regras se baseiam em um autômato finito.

• A partir de regras de acentuação, podemos determinar qual a sílaba tônica. As palavras que não se encaixam nas regras devem ser tratadas no arquivo de exceções. Como por exemplo, palavras inglesas ou italianas (pizza).

• Sabendo-se a sílaba tônica, podemos definir a entonação e o tempo dado a cada fonema da palavra. Como no exemplo a seguir. Cada ponto no gráfico representa um fonema. O tempo obedece a uma curva equivalente.

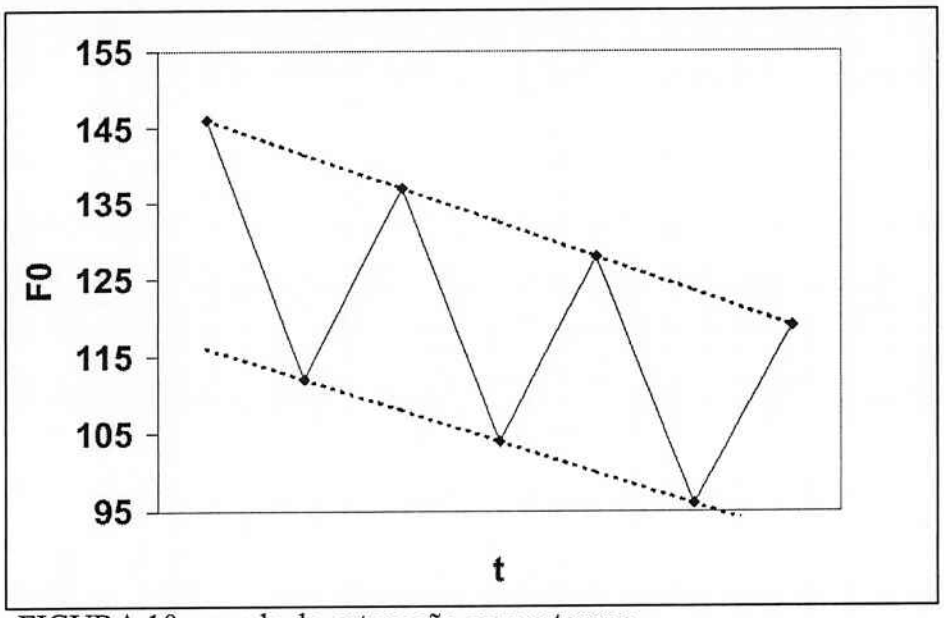
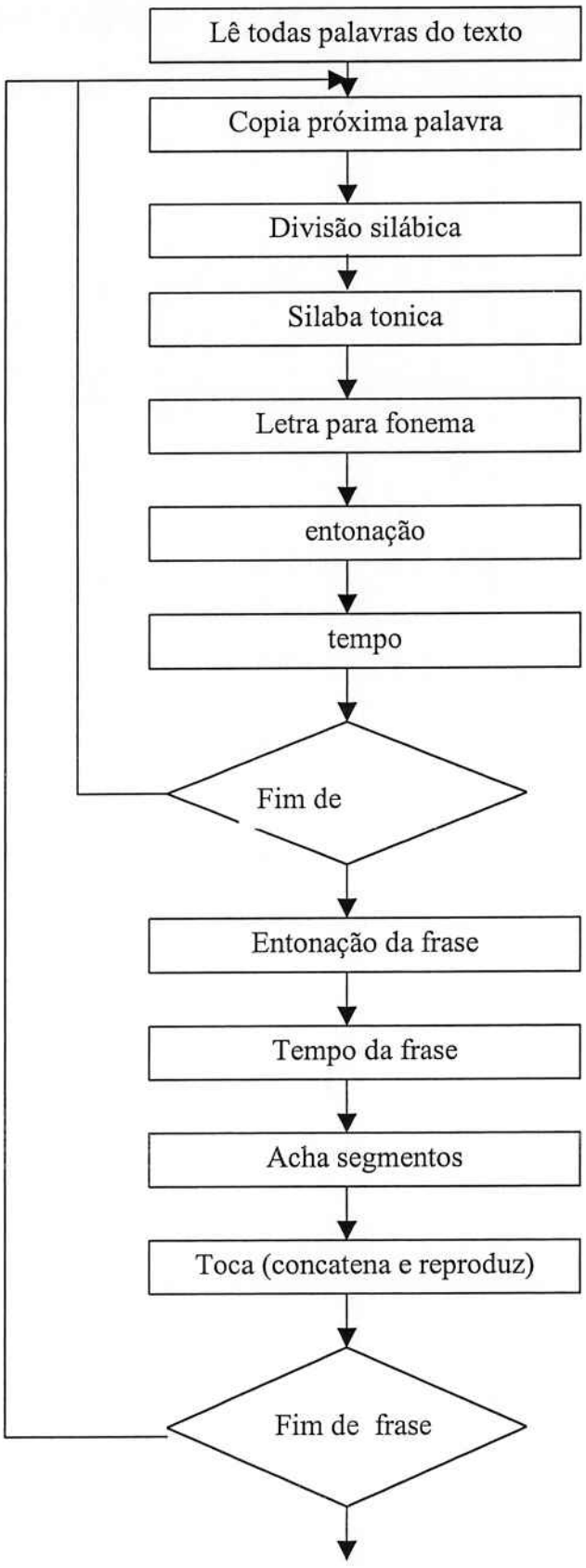


FIGURA 10 - queda da entonação com o tempo

□ 3ª fase (geração da fala)

• Precisamos então olhar para a seqüência de fonemas da frase inteira para poder determinar quais segmentos de quais difones deverão ser usados para que, se colando juntos, possa pronunciar a frase completa.

• De posse da seqüência de difones que serão utilizados, do tempo para cada parte do difone e de sua entonação também podemos aplicar as transformadas de forma a atingirmos o objetivo. Utilizamos um algoritmo chamado PSOLA para tais transformadas. Deverá se observar a continuidade de entonação entre fonemas vizinhos.


Lê todas palavras do texto
Copia próxima palavra
Divisão silábica
Silaba tonica
Letra para fonema
entonação
tempo
Fim de
Entonação da frase
Tempo da frase
Acha segmentos
Toca (concatena e reproduz)
Fim de frase

7. CONCLUSÃO

Este trabalho se baseou unicamente na pesquisa e implementação de técnicas existentes para a síntese de voz. A maior parte do tempo foi destinado a pesquisa. Portanto a parte de implementação foi bastante simplificado. Para a implementação foi utilizado o JAVA 2. O que possibilitou uma agilidade muito grande na codificação.

Certamente há muitas melhorias a ser feitas no programa. Pode-se observar a diferença de qualidade de som quando se troca de biblioteca de voz. Portanto deve-se dar uma atenção especial à gravação da voz. Já em relação a parte de processamento de sinal de voz, estamos utilizando valores absolutos para as transformadas. Isso causa algumas distorções no som quando essas transformadas (PSOLA) ultrapassam 10 %.

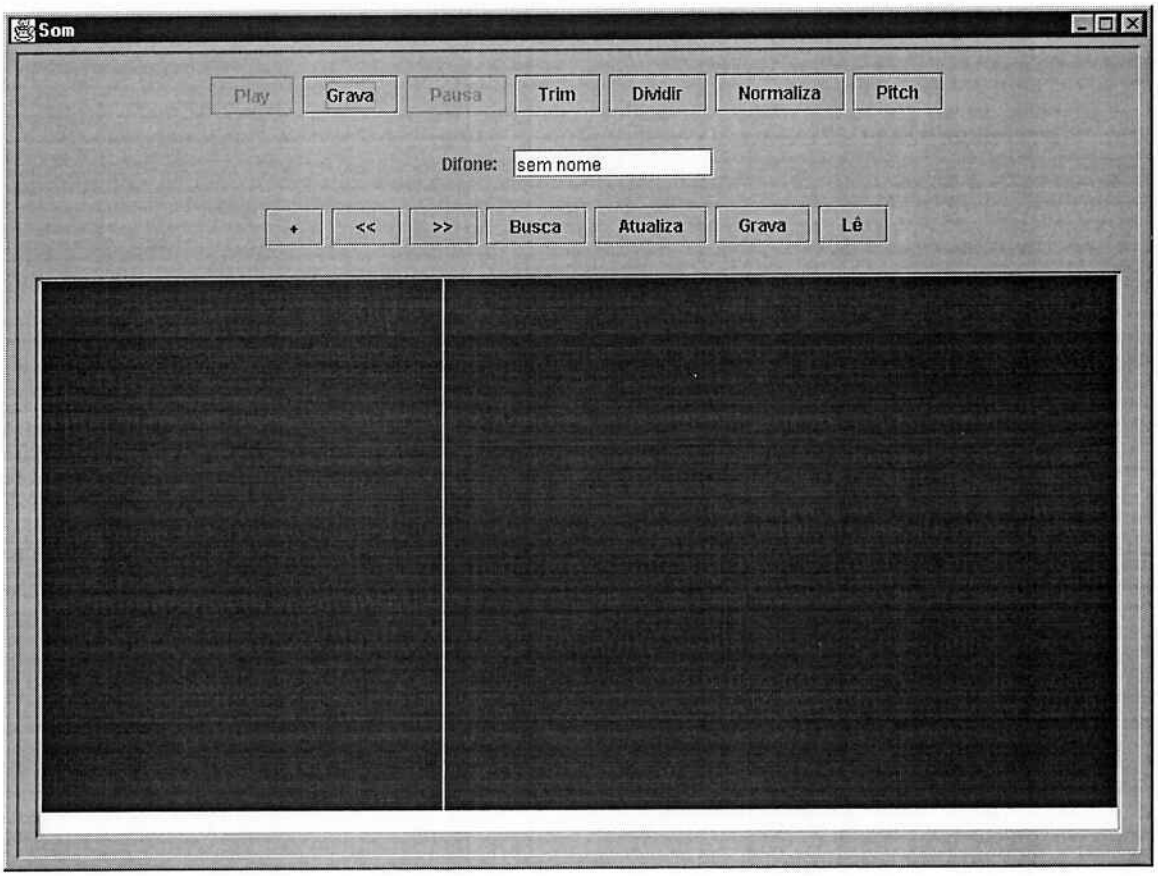
O aplicativo final ficou com qualidade razoável de inteligibilidade. Mas ainda são necessários ajustes nos valores de tempo e entonação dos fonemas, das palavras e das orações. Mas há falta de literatura nesta área relativo ao português.

O ideal para uma boa transcrição de entonação seria fazer uma analise sintática na frase para podermos determinar se é uma pergunta, uma pergunta de sim ou não, uma afirmativa, etc. Mas para isso seria necessário um trabalho muito maior e fugiria ao escopo do trabalho.

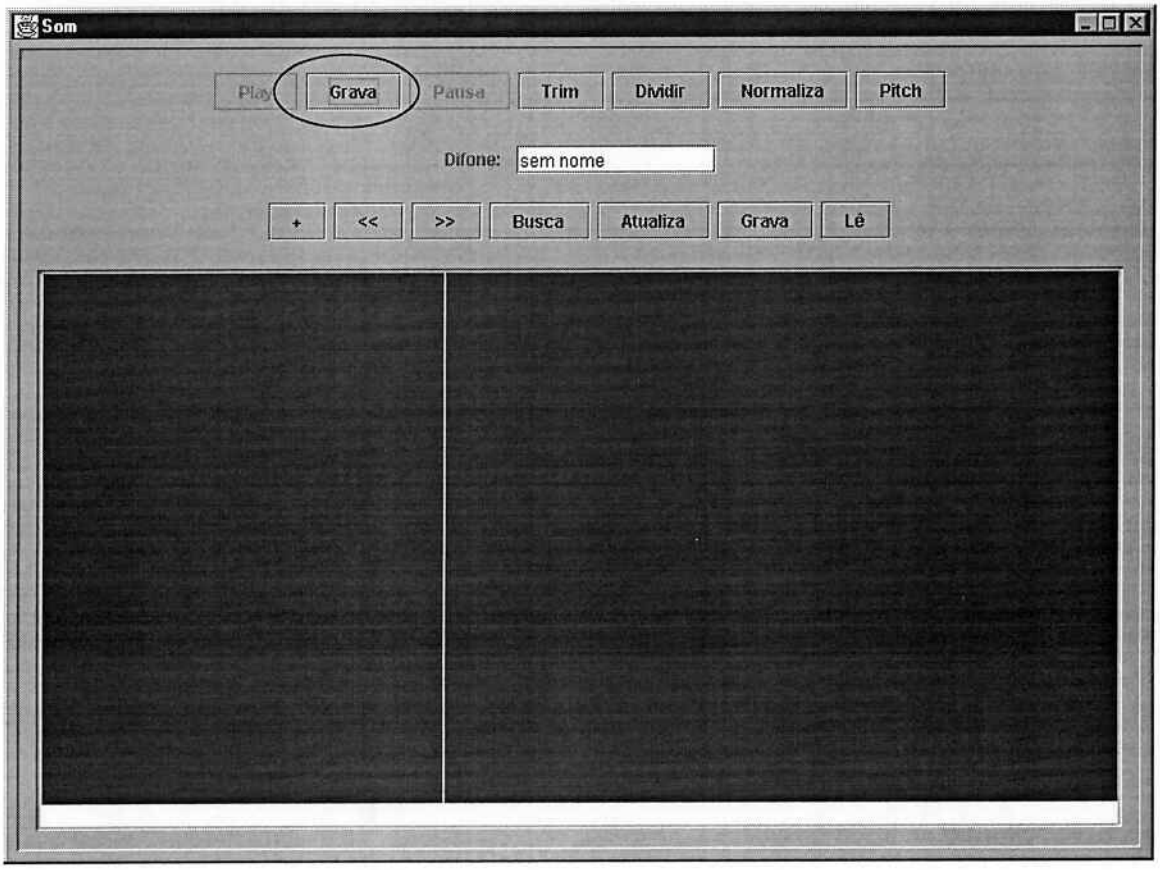
APENDICE I

Manual do programa Jwave.

O Programa Jwave destina-se a produção e edição de bibliotecas de voz para o programa SinteVoz. O aplicativo foi desenvolvido em Java e por isso precisa ter o Java Runtime Enviroment instalado.
Para executar basta fazer um duplo click em Jwave.bat.



Esta é a tela inicial do programa.

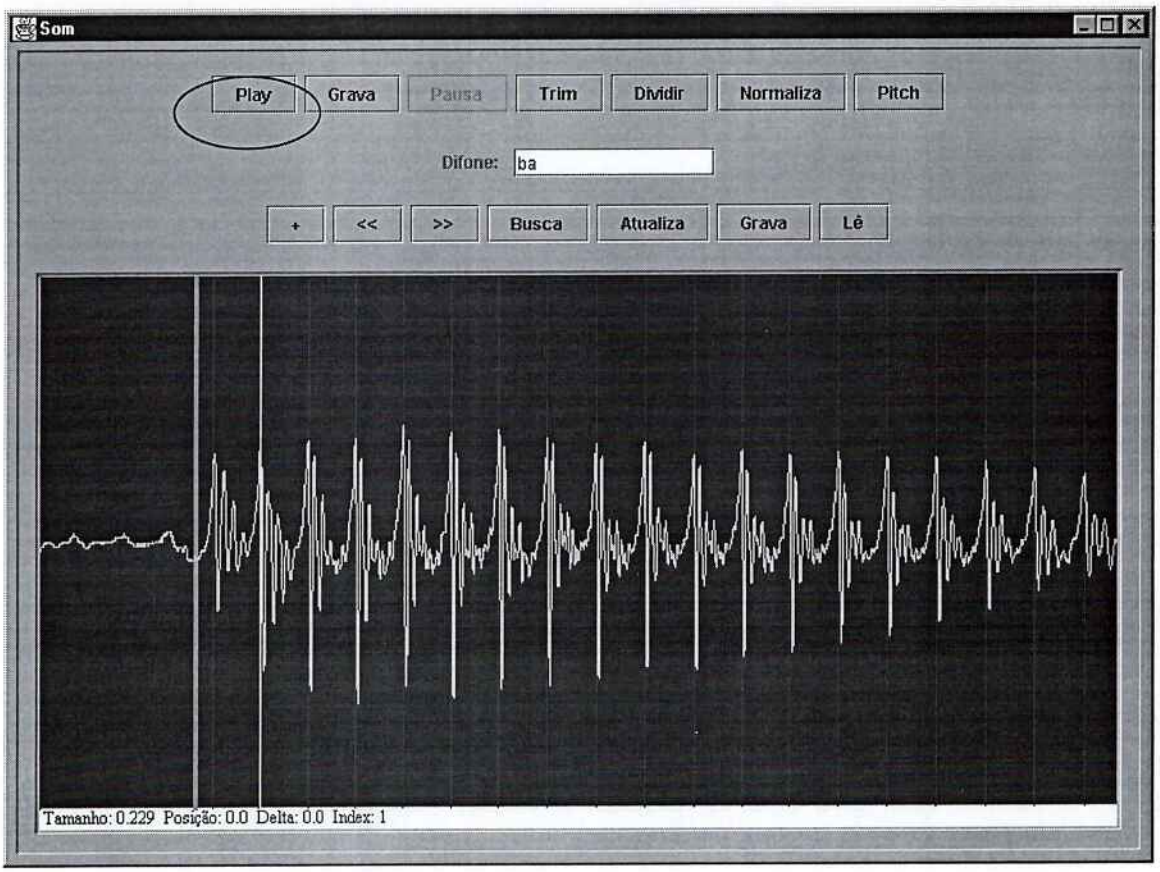


Botão Grava no superior da tela.

O botão **Grava** inicia a captura de som a partir do dispositivo selecionado. Para parar basta pressionar o mesmo botão. Porem este deve apresentar o nome de **Parar**.

É importante observar o relógio que aparecerá no canto inferior esquerdo da tela, na faixa branca. Este relógio indica se o programa está gravando e qual o tempo de gravação.

Uma vez terminada a gravação, o programa mostrara automaticamente a forma de onda no retângulo preto no interior da janela.

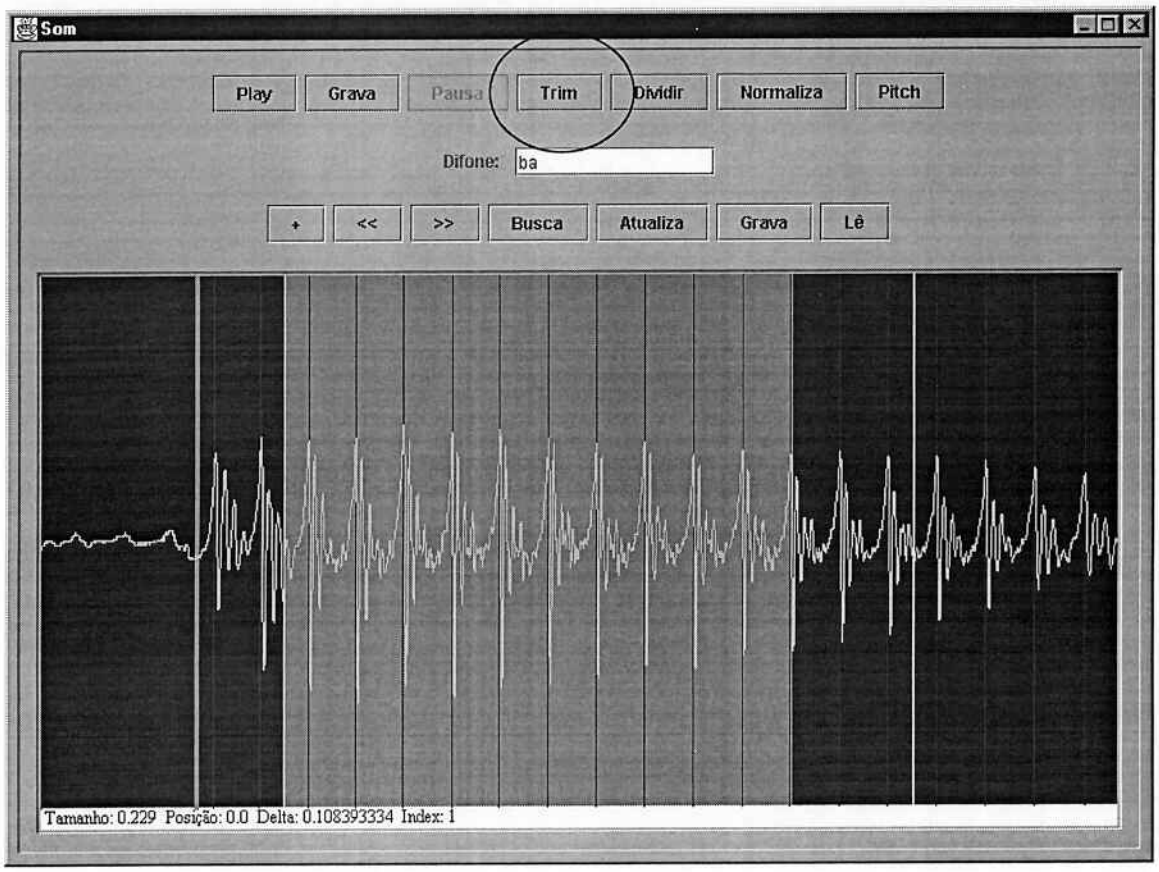


Botão play.

O botão Play é responsável pala reprodução do som. Para parar basta precionar o mesmo botão novamente.

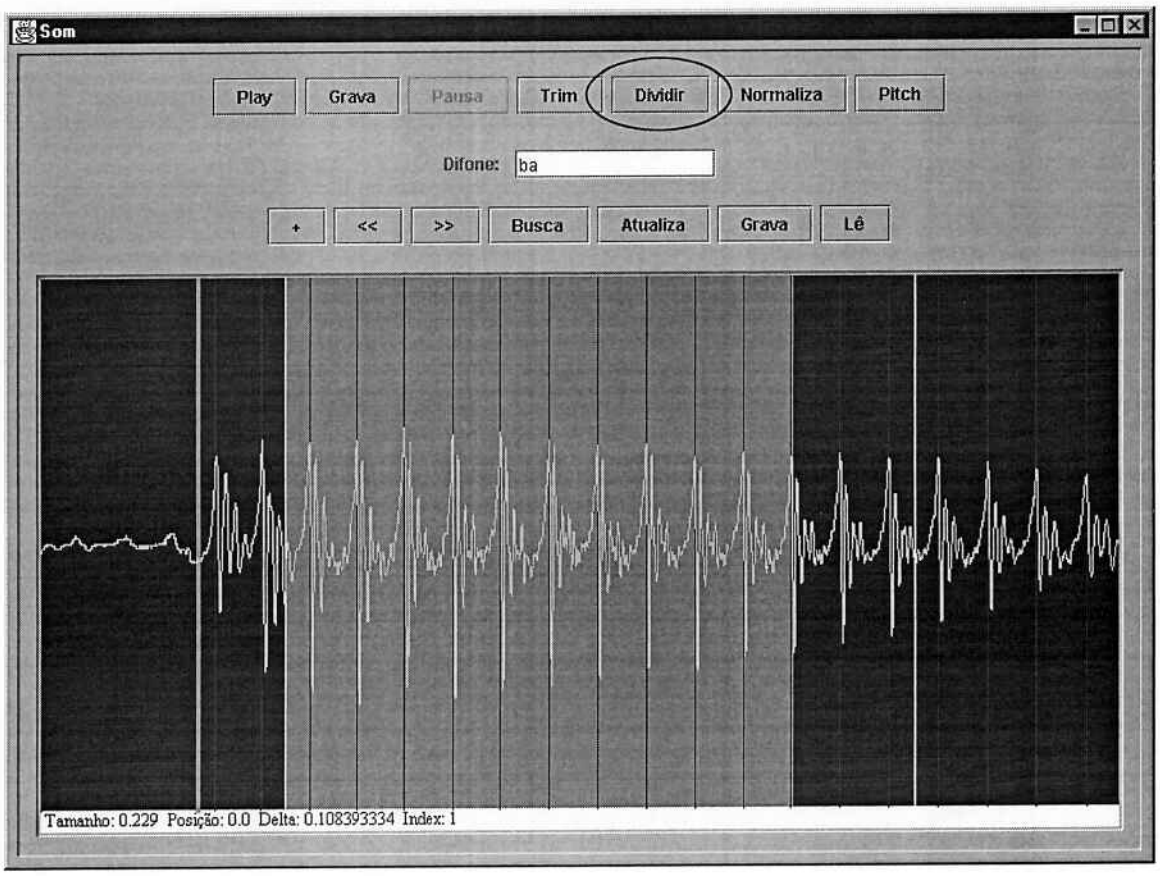
Caso haja uma região selecionada pelo mouse, somente esta região será tocada.

Para selecionar com o mouse, clique no inicio da área desejada. Arraste o mouse até o termino desejado e então se solta o mouse. A região selecionada deve deve ser mostrada com fundo cinza.



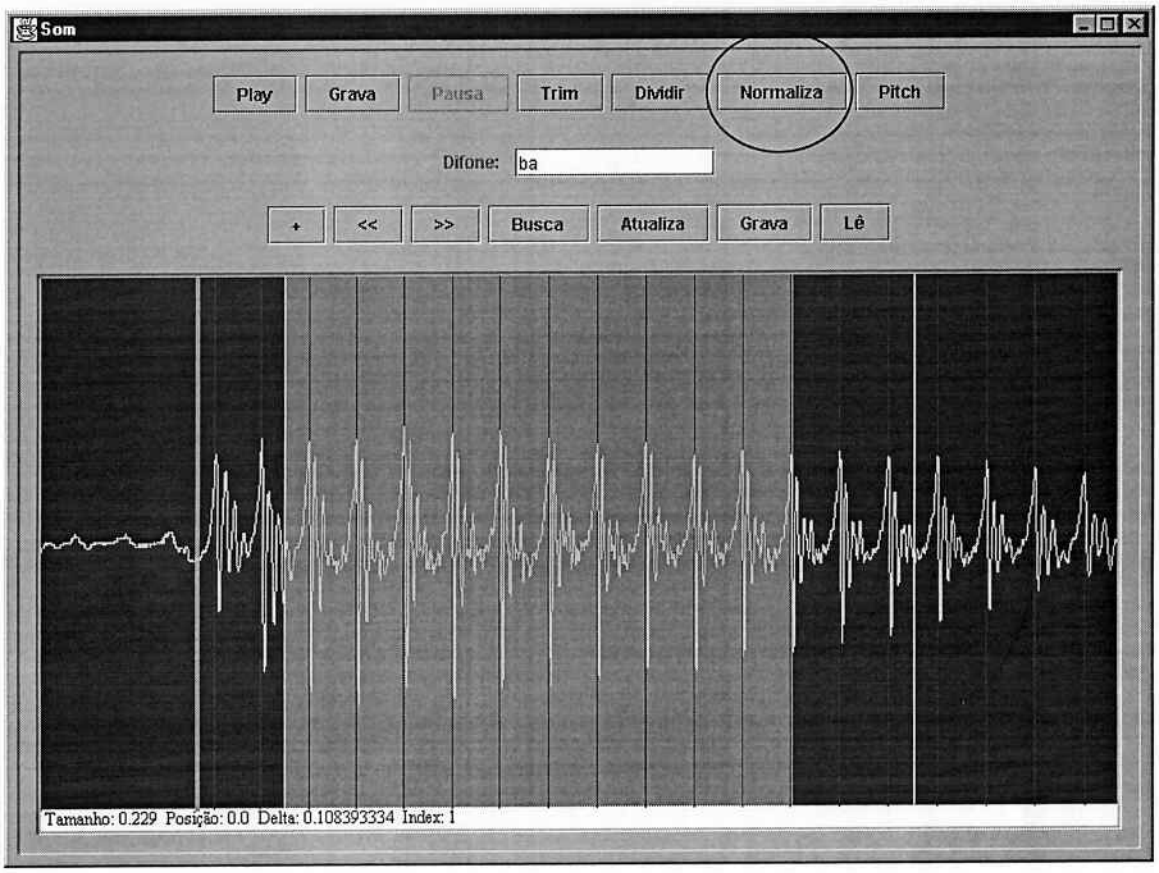
Botão Trim.

O botão trim é encarregado de aparar a região selecionada. Cortando fora a parte que não foi selecionada. Podendo assim ser selecionado apenas o dífone desejado.



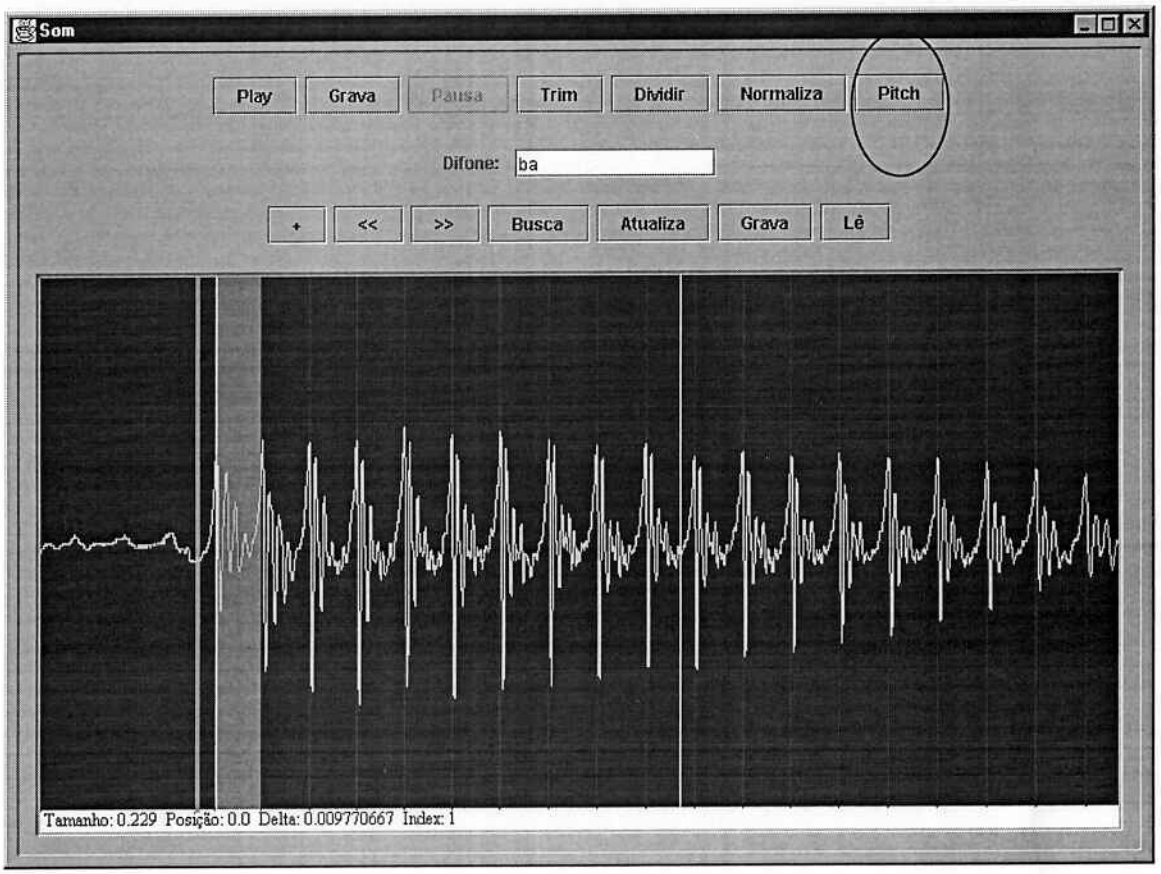
Botão dividir.

Uma vez selecionado e aparado o difone desejado, falta demarcar a divisão entre os fonemas que compõem este difone. Basta clicar com o mouse no local da divisão e então pressionar o Dividir. Eutomaticamente deverá aparecer uma linha verde no local.



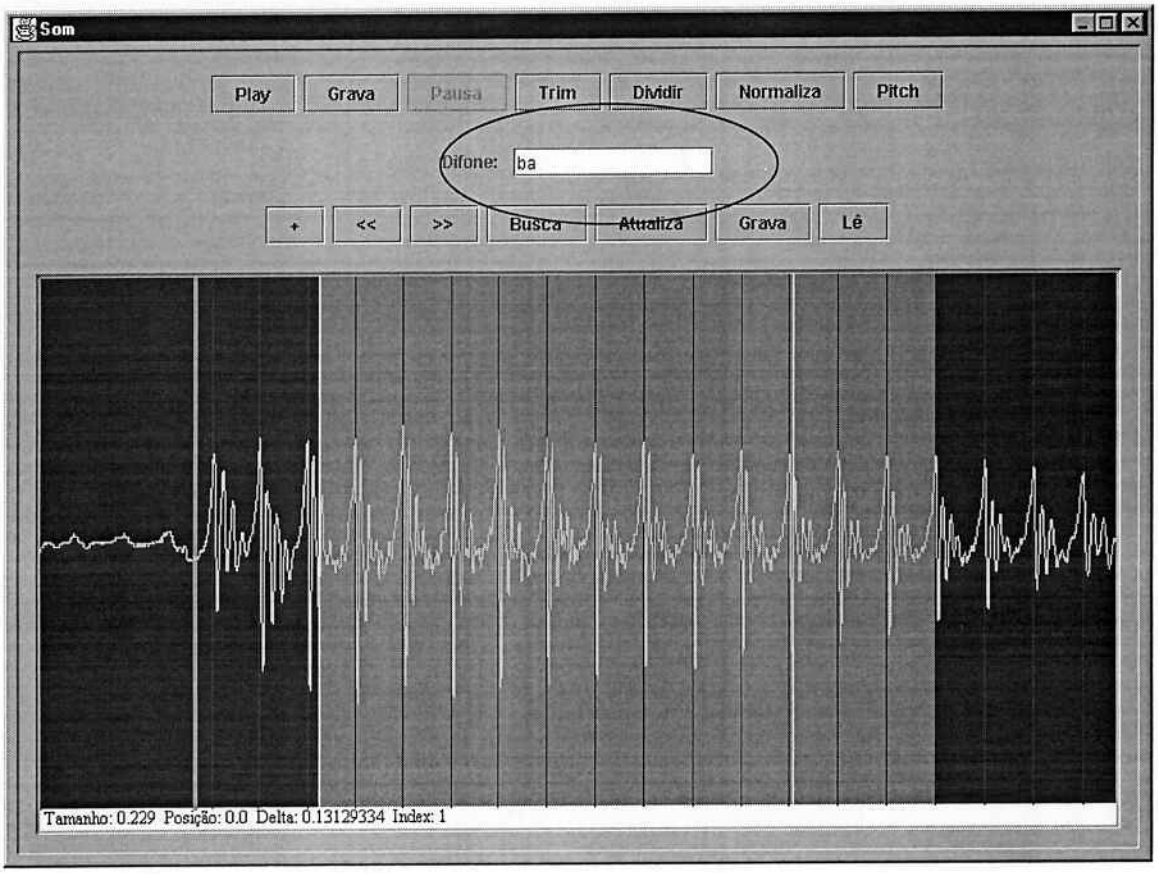
Botão Normalizar.

O difone poderá conter uma vogal. E neste caso seria interessante normalizar todas as vogas da biblioteca para um mesmo valor de amplitude. Então clica-se do lado (esquerdo ou direito da divisão) que contem a vogal e pressiona-se o botão normalizar.



Botão Pitch.

Selecionando um período (de preferência o primeiro) e depois clicando em Pitch tem-se a marcação de todos os períodos. Isso será útil para o programa de síntese para poder mudar o tom e a duração do fonema. Somente as vogais e algumas consoantes precisam ter seus períodos marcados.



A área de texto deverá ser preenchida com o nome do difone que é a própria seqüência dos fonemas que o compõem.

Pode-se então passar para os botões inferiores.

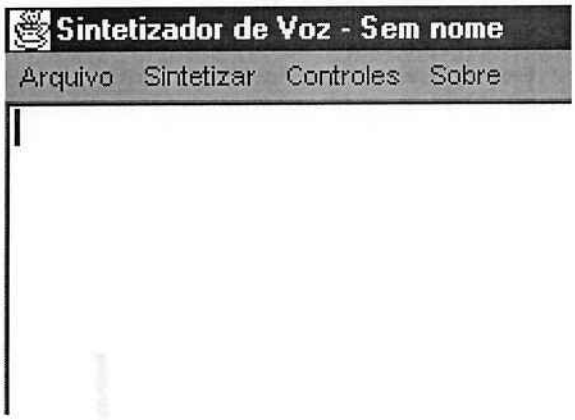
+	adiciona mais este difone a biblioteca de difones.
<<	
>>	permitem avançar e recuar pelos difones já gravados.
Busca	digitando um nome de difone na área de texto e pressionando-se Busca o programa mostra o difone gravado se ele existir.
Atualiza	caso seja necessário fazer alguma mudança no difone, ao final das mudanças basta fazer sua atualização.
Grava	grava a biblioteca em um arquivo chamado de difone.dat que deverá ter o nome mudado para ****.voz. (ex. pedro.voz). Isso evita que se grave por cima de alguma biblioteca.
Lê	Lê o arquivo de nome difone.dat para ser editado. Para isso é preciso renomear a biblioteca de ***.voz para difone.dat.

APENDICE II

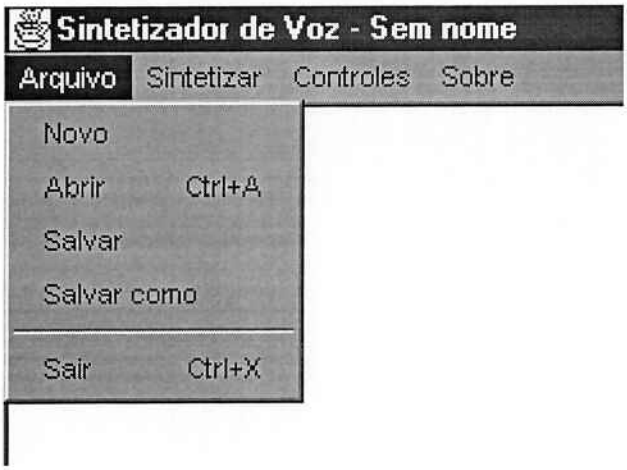
Manual do programa SinteVoz.

O aplicativo destina-se a síntese de voz. É um editor de texto simples onde se tem a opção de reproduzir o texto em forma de fala. O aplicativo foi desenvolvido em Java e por isso precisa ter o Java Runtime Enviroment instalado.
Para executar basta fazer um duplo click em SinteVoz.bat.

Programa SinteVoz. Reproduz o texto de forma falada.



Menu de operação:
Arquivo
Sintetizar
Controles
Sobre



Arquivo:

- Novo

  Inicia um novo arquivo texto.

- Abrir

  Abre a janela para escolher um arquivo para abrir.

- Salvar

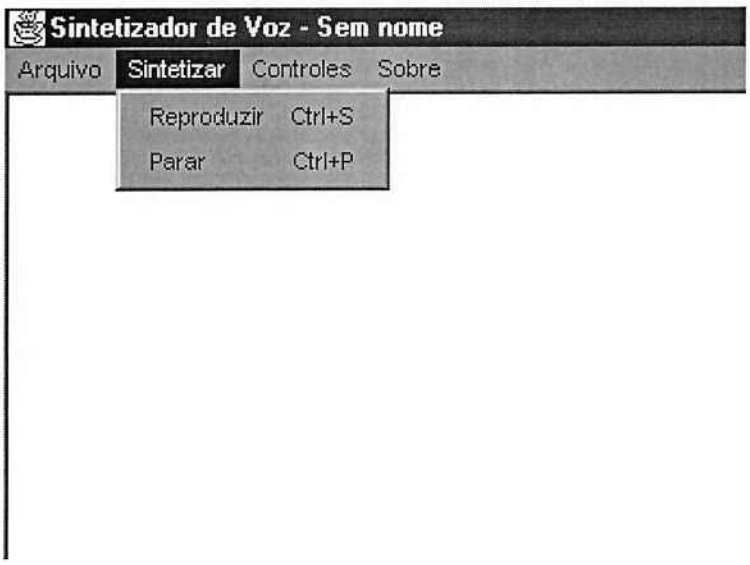
  Grava o arquivo texto

- Salvar como

  Grava o arquivo texto com um novo nome.

- Sair

  Finaliza o programa.



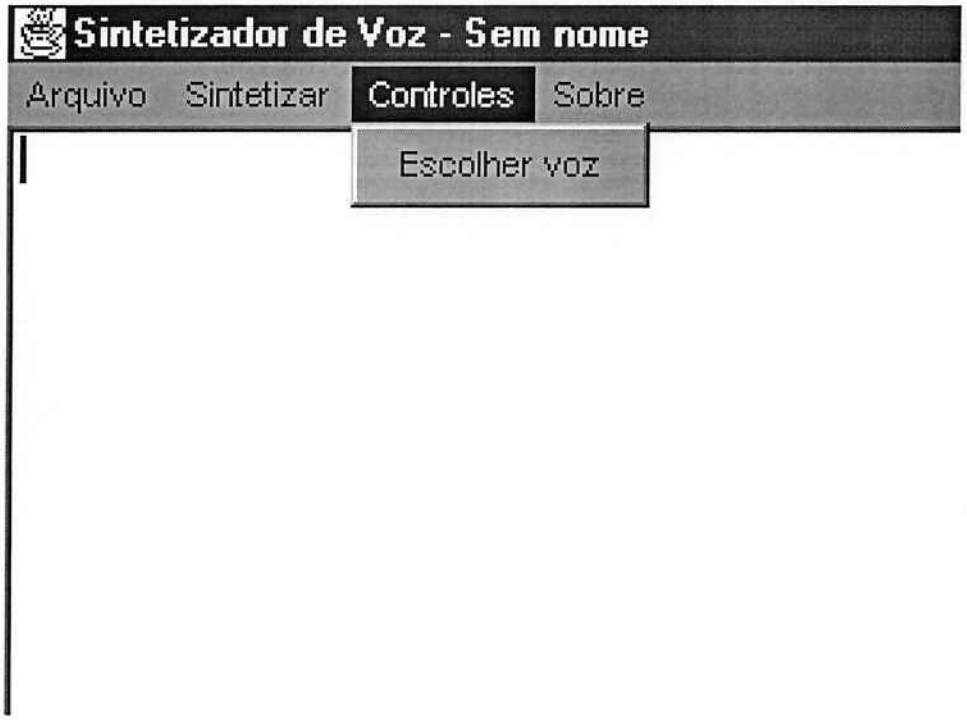
Sintetizar

- Reproduzir

  Inicia a sintese de fala e pronuncia.

- Parar

  Para a sintese de voz.



Controles:

- Escolher voz

  Permite a escolha da biblioteca de voz adequada. Atualmente com uma voz masculina e uma feminina.

Arquivos de configuração:

Execao.txt
Deve conter as exceções não tratadas no programa.

elo #[é]lo#;
cadela #ka[dé]la#;
horas #[ó]ras#;
hora #[ó]ra#;
Tadashi #ta[da]xi#;
pateta #pa[té]ta#;
espero #es[pé]ro#;
muito #[muI]to#;

Como funciona:

A palavra a esquerda deve ser a plavra a ser tratada, e a transcrição a direita deve ser iniciada e finalizada com '#' e a sílaba tonica deve estar entra colchetes. E as letras devem ser a transcrição dos fonemas.

Abrev.txt
Deve conter as abreviações a serem tratadas

depto. departamento;
Depto. departamento;
DEPTO. departamento;
D. dom;
dr. doutor;
Dr. doutor;
DR. doutor;
dra. doutora;

Dra. doutora;

DRA. doutora;

ed. edição;

Como o caso anterior, a esquerda temos a palavra para ser trocada pela palavra da direita.

Limitações:

O texto deve conter necessariamente um ponto final ao fim do texto a ser pronunciado e deve terminar em uma linha nova em branco. ('.' + 'enter')

O texto está limitado no tamanho pois o programa não trata textos muito grandes. O tamanho pode variar dependendo do número de palavras, orações e pontuações.

O programa também não trata palavras com mais de 10 sílabas.

Instalação

Para executar o programa, é preciso ter um computador com Java 2 Runtime Enviroment que pode ser obtido gratuitamente na página da Sun.

www.java.sun.com

O programa SinteVoz pode ser executado a partir do próprio CD ou o seu diretório pode ser copiado para qualquer diretório no computador.

	a	ã	e	é	~e	i	~i	o	ó	õ	u	~u
b	acabado	cabana	abedo	fubeca	fubenta	rebite	abinque	abono	abóbora	abonda	abutre	abunda
c	acaso	recanto	duquesa	moqueca	faquenca	aquilo	requinte	decôro	decoro	aconda	oculta	acunda
d	rodada	rodando	cadete	adega	badenga	pudico	pudim	adorno	adora	adonda	adulta	adunda
f	mofada	mofando	defesa	afeto	bafenga	palafita	lafinta	afônico	refoga	afonda	refuta	afunda
g	regato	uganda	foguete	espaguet	baguenga	preguiça	faguindo	engodo	fogosa	agonga	aguda	agunda
j	rajada	rajando	lajeta	cafajeste	bajenga	fugido	mugindo	tijolo	tijolos	ajonga	ajuda	ajunda
l	salada	falando	roleta	pileque	molenga	palito	falindo	piloto	maloca	ajonga	maluca	alunda
lh	palhaço	falhando	palheta	talheres	balhenga	falhito	palhinha	palhona	palhoça	alhonga	alhures	alhunda
m	camada	amante	camelo	amélia	flamenga	demite	faminto	famoso	amola	amonga	remuda	amunda
n	renato	banana	caneta	panela	fanenta	bonito	faningo	anômalo	denota	anonga	renutre	anunda
ñ	ganhado	banhando	ranheta	munheca	ranhenta	renhida	ganhindo	canhoto	canhota	anhonga	ranhuda	anhunda
p	sapato	apanha	tapete	capela	capenga	cupido	carpindo	apodo	capote	aponga	reputa	apunda
r	barato	aranha	careta	aurélio	arenga	guarita	parindo	garoto	herodes	aronga	biruta	arunda
rr	arraso	arranha	carreta	derrete	perrenga	garrida	arrindo	barroca	derrota	arronga	derruba	arrunda
s	laçada	miçanga	macete	acesso	facenga	possível	fassindo	assopro	paçoca	açonga	açude	açunda
t	ataque	pitanga	cateto	pateta	setenta	batida	fatindo	atoba	patota	atonga	fratura	atunda
v	lavada	avante	avesso	reverso	seventa	revide	avindo	avoba	gaivota	avonga	avuta	avunda
x	fachada	rachando	rochedo	recheque	pixenta	buxixo	achindo	muxoxo	enxota	axonga	machuca	axunda
z	casado	casando	azedo	casebre	fazenda	reside	resinda	amazona	resolve	azonga	resuta	azunda

	a	ã	e	é	~e	i	~i	o	ó	õ	u	~u
bl	ablate	ablante	ableto	abléto	ablento	ablito	ablinto	abloto	ablote	ablonte	abluto	ablunte
cl	aclate	aclante	acleto	acléto	aclento	aclito	aclinto	acloto	aclóte	aclonte	acluto	aclunte
fl	aflate	aflante	afleto	afléto	aflento	aflito	aflinto	afloto	aflóte	aflonte	afluto	aflunte
gl	aglate	aglante	agleto	agléto	aglento	aglito	aglinto	agloto	aglóte	aglonte	agluto	aglunte
pl	aplate	aplante	apleto	apléto	aplento	aplito	aplinto	aploto	aplóte	aplonte	apluto	aplunte
tl	atlate	atlante	atleto	atléto	atlento	atlito	atlinto	atloto	atlóte	atlonte	atluto	atlunte
br	abrate	abrante	abreto	abréto	abrento	abrito	abrinto	abroto	abróte	abronte	abruto	abrunte
cr	acrate	acrante	acreto	acréto	acrento	acrito	acrinto	acroto	acróte	acronte	acruto	acrunte
dr	adrate	adrante	adreto	adréto	adrento	adrito	adrinto	adroto	adróte	adronte	adruto	adrunte
fr	afrate	afrante	afreto	afréto	afrento	afrito	afrinto	afroto	afróte	afronte	afruto	afrunte
gr	agrate	agrante	agreto	agréto	agrento	agrito	agrinto	agroto	agróte	agronte	agruto	agrunte
pr	aprate	aprante	apreto	apréto	aprento	aprito	aprinto	aproto	apróte	apronte	apruto	aprunte
tr	atrate	atrante	atreto	atréto	atrento	atrito	atrinto	atroto	atróte	atronte	atruto	atrunte
vr	avrate	avrante	avreto	avréto	avrento	avrito	avrinto	avroto	avróte	avronte	avruto	avrunte

	a	ã	e	é	~e	i	~i	o	ó	õ	u	~u
s	afaste	afãnste	afeste	aféste	afenste	afiste	afinste	afôste	afóste	afonste	afuste	afunste
z	afazer	afanzer	afezer	afézer	afenzer	afizer	afinzer	afôzer	afózer	afonzer	afuzer	afunzer
j	afajim	afanjer	afejer	aféjer	afenjer	afijer	afinjer	afôjer	afójer	afonjer	afujer	afunjer
ch	afachim	afanchim	afechim	aféchim	afenchim	afichim	afinchim	afôchim	afóchim	afonchim	afuchim	afunchim
f	afafim	afanfim	afefim	aféfim	afenfim	afifim	afinfim	afôfim	afófim	afonfim	afufim	afunfim

	a	ã	e	é	~e	i	~i	o	ó	õ	u	~u
a		ãa	ea	éa	ena	ia	ina	oa	óa	õa	ua	una
ã	aã		eã	éã	enã	iã	inã	oã	óã	õã	uã	unã
e	ae	ãe		ée	ene	ie	ine	oe	óe	õe	ue	une
é	aé	ãé	eé		ené	ié	iné	oé	óé	õé	ué	uné
~e	aen	ã	e	é		i	in	o	ó	õ	u	un
i	ai	ã	e	é	en		in	o	ó	õ	u	un
~i	ain	ã	e	é	en	i		o	ó	õ	u	un
o	ao	ã	e	é	en	i	in		ó	õ	u	un
ó	aó	ã	e	é	en	i	in	o		õ	u	un
õ	aõ	ã	e	é	en	i	in	o	ó		u	un
u	au	ã	e	é	en	i	in	o	ó	õ		un
~u	aun	ã	e	é	en	i	in	o	ó	õ	u	

ditongos crescentes			ditongos decrescentes		
ea	área	orquídea	ãe	mãe	pães
eo	áureo	níveo	ai	mais	pai
ia	concórdia	glória	ãi	cãimbra	plaina
ie	espécie	série	ão	cão, órgão	calam
io	fio	pavio	au	austor	bacalhau
ao	mágoa	páscoa	éi	papéis	quartéis
ua	água	tábua	ei	feira	leite
uã	quando	quanto	~ei	bem	refém
eu	eqüestre	tênue	éu	céu	chapéu
u~e	freqüente	agüento	eu	ateu	sandeu
ui	cuidado	tranqüilo	iu	iugoslavo	tiziu
u~i	pingüim	qüinqüênio	õe	corações	põe
uo	árduo	ingênuo	ói	anzóis	herói
			oi	açoite	boi
			ou	agouro	tesoura
			ui	circuito	gratuito
			~ui	muito	

BIBLIOGRAFIA

CAMPOS,G.L. **Síntese de voz para o Idioma Português.** São Paulo, 1980.Tese (doutorado) - Escola Politécnica, Universidade de São Pulo.

CHBANE,D.T. **Desenvolvimento de sistema para conversão de textos em fonemas no idioma português.** São Paulo, 1994. Dissertação (mestrado), Escola Politécnica, Universidade de São Paulo.

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

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.

GOUVÊA,E.B. **Síntese de voz com qualidade.** São Paulo, 1993. Dissertação (mestrado), Escola Politécnica, Universidade de São Paulo.

GUIMARÃES,F.; GUIMARÃES,M. **Agramática lê o texto.** Ed Moderna, São Paulo, 1997.

JOSÉ NETO,J. **Introdução à Compilaçã**o. __ ed., Rio de Janeiro, Livros Técnicos e Científicos Editora S.A., 1987.

KAPLAN,G.; LERNER,E.J. Realism in Synthetic Speech. **IEEE Spectru**m, v. 22, p. 32-7, Apr. 1985.

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

MAIA,E.M. **No reino da fala.** editora ática, São Paulo, 1991.

SAUSSURRE,F. **Curso de lingüística geral.** Cultrix, São Paulo, 1984.