UNIVERSIDADE ESTADUAL DE CAMPINAS
FACULDADE DE ENGENHARIA ELÉTRICA

MÓDULO FRONTAL

PARA UM SISTEMA DE RECONHECIMENTO AUTOMÁTICO DE VOZ

POR: MAURÍLIO NUNES VIEIRA
Engenheiro Eletricista (PUCMG - 1987)

ORIENTADOR: PROF. DR. FÁBIO VIOLARO
Professor MS4 do Departamento de Comunicações da
Faculdade de Engenharia Elétrica da UNICAMP

Este exemplar corresponde à redação final da Tese defendida por Maurílio Nunes Vieira e aprovada pela Comissão Julgadora em 19/12/89

Fábio Violaro 19/01/90

Dissertação apresentada à Faculdade de Engenharia Elétrica da UNICAMP como requisito parcial para a obtenção do título de Mestre em Engenharia Elétrica.

CAMPINAS, dezembro de 1989

Este trabalho contou com o apoio financeiro do CNPq e do CPqD/TELEBRAS, através do convênio UNICAMP/TELEBRAS 208/87.

Agradeço

ao Professor Doutor Fábio Violaro, pela orientação segura e pela confiança depositada na realização deste trabalho;

à Professora Doutora Eleonora Albano, do Instituto de Estudos da Linguagem da UNICAMP, pelas lições de Fonética Acústica e de Psicolingüística;

aos Professores Doutores Dalton Soares Arantes, Amauri Lopes, João Baptista Tadanobu Yabu-uti e José Geraldo Chiquito, do Departamento de Comunicações da UNICAMP, pelas diversas formas de incentivo.

A meus pais, Nilson e Ébia.

Para Elizabet.

SUMÁRIO

Este trabalho descreve o desenvolvimento do software para o Módulo Frontal de um Sistema de Reconhecimento Automático de Voz para operação na faixa de 0-4 kHz. O Módulo Frontal, ou Processador Acústico, é responsável pela extração de traços para a caracterização dos diversos sons da fala.

O sinal de voz sofre uma filtragem passa-baixas com corte em 3,4 kHz, é amostrado a 8,0 kHz e quantizado em 12 bits. As análises são feitas em quadros de 25 ms, deslocados a um passo de 5 ms, obtendo-se uma série de parâmetros, como o número de cruzamentos por zero, o período de pitch para os intervalos sonoros, a energia em diversas faixas de freqüência do espectro LPC (Linear Predictive Coding) e a freqüência, amplitude e largura de faixa dos três primeiros formantes.

O quadro é classificado em uma dentre sete categorias: silêncio, fricativo surdo, fricativo sonoro, oclusão sonora, vocálico, coarticulação ou indefinido. Esta última categoria inclui segmentos que não podem ser confiavelmente classificados em nenhuma das outras categorias. A classificação é independente do locutor.

ÍNDICE

CAPÍTULO 1

INTRODUÇÃO

1.1 A RIQUEZA DA FALA

Ao observarmos o diálogo entre duas pessoas, rápida e seguramente descobriremos o sexo e a faixa etária dos indivíduos. Saberemos se a língua que está sendo falada é de nosso conhecimento, e ainda obteremos várias outras informações a partir da voz, gestos, movimentos corporais e expressões faciais das pessoas.

Se nos privarmos dos gestos, movimentos labiais e de todos os outros aspectos não acústicos da comunicação, ainda assim teremos inúmeras informações extraídas unicamente a partir da voz. Estas informações podem ser separadas, para efeito de estudo, em dois grupos: de um lado, informações relativas a uma mensagem primária, compreendendo as palavras e as frases e, de outro, mensagens secundárias, que são informações associadas à pessoa do locutor, seu grupo sócio-cultural, sotaque, estado emocional, estado de saúde, e uma infinidade de outras características do indivíduo, transportadas pela voz.

Desde o início da vida aprendemos a decodificar estas mensagens; com poucos meses um bebê já é capaz de identificar a voz materna e demonstrar contentamento ao ouvir uma voz afetiva. O aprendizado da língua evolui por toda a vida, à medida que aprendemos a detectar matizes mais sutis na fala, como a ironia e a tristeza.

1.2 A MÁQUINA, A VOZ E O HOMEM

Os primeiros esforços para a construção de máquinas falantes datam do final do século XVIII, quando foram elaborados curiosos engenhos acústicos que produziam sons semelhantes à voz e eram "tocados" à maneira de um instrumento musical.

A busca por representações eficientes para a transmissão digital da voz a baixas taxas e para seu armazenamento econômico iniciou-se na Engenharia de Comunicações, a partir da década de 40, e rapidamente recebeu a atenção de toda a comunidade científica [1].

Sistemas para a síntese da fala humana já evoluíram a ponto de existirem protótipos em laboratórios capazes de ler um texto sem que se perceba facilmente a artificialidade da fala [2].

Os esforços para a construção de máquinas capazes de realizar alguma forma de reconhecimento de voz, contudo, têm alcançado resultados mais modestos.

A busca por parâmetros acústicos que variem de indivíduo para indivíduo, de forma a possibilitar a identificação das pessoas pela voz tem sido realizada nos *Sistemas de Reconhecimento de Locutor*, em duas linhas principais [3]:

a) *Sistemas para Verificação do Locutor* [4], com a finalidade de decidir se um determinado trecho de fala pertence ou não a um suposto indivíduo;
b) *Sistemas para Identificação do Locutor* [5], que devem associar um determinado enunciado a um elemento pertencente a um grupo limitado e conhecido de pessoas.

A identificação dos indivíduos por suas vozes é um problema científico de importância social devido às suas implicações legais, tendo em vista a possibilidade de utilização da voz em casos policiais com o mesmo propósito que hoje é dado às impressões digitais [6].

Outra aplicação imediata do problema de Reconhecimento de Locutor está no controle do acesso a algum ambiente pelo uso

de senha verbal.

Nos *Sistemas de Reconhecimento de Voz* são procuradas invariâncias no sinal acústico que possibilitem a discriminação de palavras, sílabas, fonemas, ou quaisquer outras unidades que se queira reconhecer.

As dificuldades na determinação de invariâncias (no caso do reconhecimento de voz) ou variâncias (no reconhecimento de locutor) decorrem da constatação de que as características acústicas das diversas mensagens conduzidas pela voz não estão agrupadas em conjuntos facilmente identificáveis.

O descobrimento de correlatos acústicos associados aos fenômenos perceptuais da fala requer a realização de experimentos com a participação de especialistas de várias áreas.

1.3 ORGANIZAÇÃO DA DISSERTAÇÃO

Após esta ligeira introdução a alguns dos aspectos de interesse corrente na análise de voz voltada à comunicação homem/máquina, no Capítulo 2 são apresentados os Sistemas de Reconhecimento de Voz descritos na literatura e analisadas as questões que orientaram este trabalho.

No Capítulo 3 são discutidos os temas da produção e percepção da fala, e introduzidos os modelos e conceitos a serem utilizados nas análises posteriores.

O Capítulo 4 é dedicado à parte experimental do trabalho, isto é, a elaboração do software para o Módulo Frontal, sendo apresentados os aspectos práticos do desenvolvimento e teste dos algoritmos.

No 5º e último capítulo é feita uma avaliação dos resultados e das perspectivas que estão surgindo no contexto de Reconhecimento Automático de Voz.

Salvo indicação em contrário, todas as formas de onda, espectros e gráficos estatísticos apresentados no trabalho foram obtidos através dos programas desenvolvidos para esta dissertação. A digitalização do sinal de voz e a geração de

arquivos de voz para o desenvolvimento e teste dos algoritmos foi feita no "Sistema de Análise e Processamento Digital de Voz A" [7] do Laboratório de Comunicações Digitais do Departamento de Comunicações da FEE/UNICAMP.

1.4 SÍMBOLOS, ABREVIATURAS E CONVENÇÕES

t	tempo (s);
f	freqüência (Hz);
$\Omega = 2\pi f$	freqüência angular (rad/s);
T	período de amostragem (s);
f_s	freqüência de amostragem (Hz);
$\omega = 2\pi \cdot f/f_s$	freqüência normalizada do espectro do sinal digitalizado (rad);
$\overset{\mathcal{F}}{\longleftrightarrow}$	Transformada de Fourier;
$\overset{\mathcal{Z}}{\longleftrightarrow}$	Transformada $\mathcal{Z}$;
$*$	operação de convolução;
$v(t)$	sinal analógico de voz;
$\{v_n\}$	seqüência ou conjunto de amostras resultantes da digitalização do sinal v(t);
v_n ou $v(n)$	valor de uma amostra particular de $\{v_n\}$;
$\{F_k\}$	seqüência ou conjunto de amostras do espectro de $F(e^{j\omega})$;
F_k	valor de uma amostra particular de $\{F_k\}$;
AMDF	Average Magnitude Difference Function;
DFT	Discrete Fourier Transform;
FFT	Fast Fourier Transform;
LPC	Linear Predictive Coding;
f0	freqüência fundamental de excitação;
F1, F2, F3	primeiro, segundo e terceiro formantes, respectivamente;
/ /	as duas barras indicam a pronúncia de fonemas ou palavras. A palavra "voz", por exemplo, pode ser representada por /vós/.

1.5 REFERÊNCIAS

[1] J. L. Flanagan, *"Voices of Men and Machines"*, Journal of the Acoustical Society of America, vol. 51, pp 1375-1387 (1972);

[2] D. H. Klatt, *"Review of Text-to-Speech Conversion for English"*, Journal of the Acoustical Society of America, vol. 82, pp 737-793 (1987);

[3] D. O'Shaughnessy, *"Speech Communication: Human and Machine"*, Addison-Wesley Publishing Company, cap 11 (1987);

[4] A. E. Rosenberg, *"Automatic Speaker Verification: A Review"*, Proceedings of the IEEE, vol. 64, pp 475-487 (1976);

[5] B. S. Atal, *"Automatic Recognition of Speakers from Their Voices"*, Proceedings of the IEEE, vol. 64, pp 460-475 (1976);

[6] R. H. Bolt et al, *"Speaker Identification by Speech Spectrograms: A Scientist's View of Its Reliability for Legal Purposes"*, in "Human Communication: A Unified View", E. E. David Jr. & P. B. Denes, ed., McGraw-Hill (1972);

[7] F. Violaro, *"Nova Versão do Sistema de Análise e Processamento Digital de Voz, SAPDV-A"*, Anais do 7º Simpósio Brasileiro de Telecomunicações, pp 50-53 - Florianópolis, S. C. (1989).

CAPÍTULO 2

RECONHECIMENTO AUTOMÁTICO DE VOZ:

FUNDAMENTOS, EVOLUÇÃO E PERSPECTIVAS

2.1 INTRODUÇÃO

No capítulo anterior foi feita uma ligeira discussão sobre algumas maneiras de utilização prática das informações conduzidas pela voz. A multidimensionalidade e a simultaneidade dos efeitos acústicos e articulatórios que ocorrem na produção da fala, assim como o desconhecimento das maneiras utilizadas pelo homem para decodificá-la, são fatores que tornam a busca e a compreensão dos parâmetros associados às diversas mensagens presentes no sinal de voz uma tarefa extremamente complexa.

Circuitos ou algoritmos computacionais que possibilitam a construção de máquinas com a capacidade de "ouvir" a voz humana, detectar a pronúncia de palavras específicas ou até mesmo "compreender" o significado de frases construídas a partir de um vocabulário limitado têm sido tratados sob o título geral de Sistemas para Reconhecimento Automático de Voz (SRAV).

Neste capítulo será feita uma apresentação do jargão utilizado nesta área, das técnicas de uso mais corriqueiro na construção de sistemas práticos e uma exposição histórica da evolução dos SRAV.

2.2 PRIMÓRDIOS

O primeiro trabalho descrevendo uma máquina que podia, de alguma forma, reconhecer com certo sucesso a pronúncia de determinadas palavras data de 1952 [1]. Trata-se de um sistema capaz de distinguir os dez dígitos da língua inglesa falados ao telefone, atingindo percentuais de acerto de quase 100% quando devidamente ajustado para um indivíduo. Entretanto, a eficiência caía para até 50% quando o sistema, uma vez calibrado para uma pessoa, era utilizado por outra.

Muitos trabalhos sucederam-se nos anos 60 [2], a nível de laboratório, em torno de reconhecimento de palavras, sílabas, letras e fonemas isolados, em função da descoberta de algumas propriedades da voz através do Espectrógrafo [3] e das novas facilidades oferecidas pelos computadores digitais.

Formaram-se duas linhas de atividade: de um lado, sistemas que procuravam distinguir um conjunto relativamente alto de unidades, geralmente entre 30 e 50 palavras, falados por apenas uma pessoa e, de outro, tentativas de reconhecimento de poucas unidades, como os dígitos e algumas vogais, para um grupo de 5 a 25 pessoas. Esta divisão em sistemas *dependentes do locutor*, com grande vocabulário, e *sistemas multilocutores*, com pequeno vocabulário, decorreu da incrível dificuldade do reconhecimento de grandes vocabulários independentemente da pessoa que fala. Os sistemas multilocutores são normalmente denominados *independentes do locutor* embora, em termos absolutos, não o sejam.

Uma outra característica dos primeiros sistemas é a necessidade da pronúncia das palavras com uma ligeira pausa entre si, de forma a facilitar a localização do início e fim de cada palavra. Isto se deve aos efeitos da coarticulação, onde o final da pronúncia da maioria das palavras, numa fala natural, é alterado ou se funde com o início da palavra seguinte, gerando padrões acústicos bem diferentes dos padrões das palavras pronunciadas isoladamente. Como exemplo, o número "32" é normalmente pronunciado como /trinteidois/, transformando o /a/

final de /trinta/ e o /e/ no ditongo /ei/.

Os efeitos da coarticulação estão, ainda hoje, entre os problemas mais difíceis de serem resolvidos. Desta dificuldade surgiu uma divisão dos Sistemas de Reconhecimento de Voz no que diz respeito à forma de pronunciar as palavras:

a) os sistemas de reconhecimento de *palavras isoladas* ou de *fala discreta*, como descrito no parágrafo anterior, exigem um pequeno intervalo (cerca de 300 ms) entre a pronúncia de duas palavras;
b) os sistemas para *fala natural* ou *fala contínua*, permitem a pronúncia como em uma conversa normal;
c) nos sistemas para *fala conectada*, não há a necessidade de pausas, mas é exigida uma pronúncia bem clara de cada palavra.

Esta última forma de pronúncia é uma posição intermediária entre a primeira (fala discreta) e a segunda (fala contínua), representando um compromisso entre confiabilidade no reconhecimento e conforto na pronúncia. Deve ser ressaltado que o reconhecimento de fala natural é bastante elaborado e requer um profundo conhecimento de lingüística.

2.3 RECONHECIMENTO DE PALAVRAS ISOLADAS

Até agora falou-se de "sistemas" para reconhecimento de voz sem a discussão de suas características internas. A figura 2.1 é um exemplo de sistema que emprega as técnicas tradicionais.

Embora alguns sistemas mais antigos tenham sido implementados integralmente com processamento analógico de sinais, as discussões serão feitas admitindo-se que o sinal de voz é analisado em um computador digital.

O sistema da figura 2.1, é um exemplo da aplicação das técnicas básicas de *reconhecimento de padrões*: é extraído um conjunto de características para cada palavra a ser reconhecida e gerado o conjunto de padrões de referência; isto é feito durante a *fase de treinamento*. Quando o sistema é colocado no *modo de*

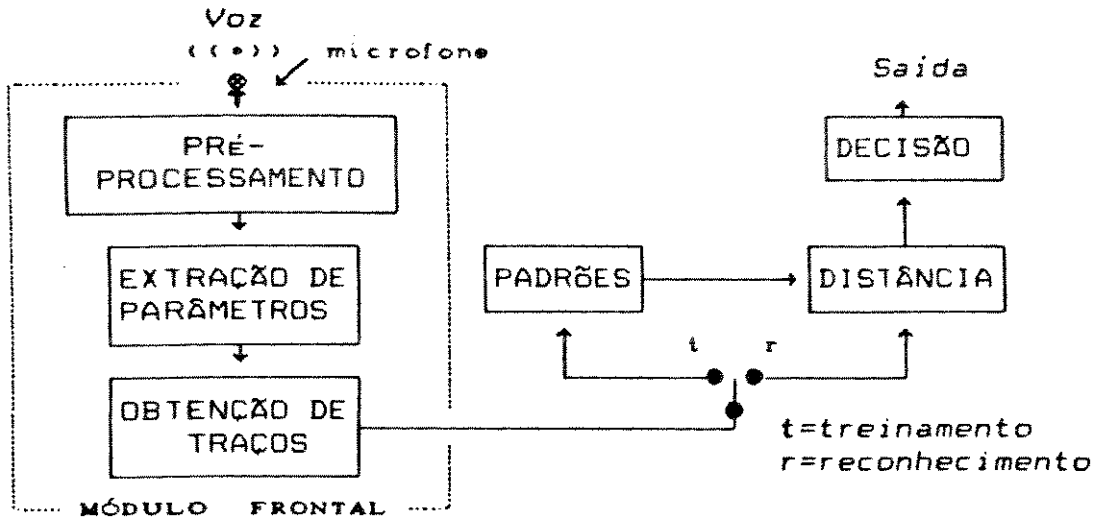


FIG.2.1 - Diagrama em blocos de um Sistema de Reconhecimento de Voz utilizando técnicas clássicas de comparação de padrões.

operação ou reconhecimento, os padrões obtidos para a palavra de teste são comparados com os padrões de referência de todas as outras palavras, escolhendo-se como resposta o padrão que mais se assemelha ao de entrada.

Uma grande dificuldade na realização do reconhecimento de voz com a comparação de padrões decorre das variações na velocidade da fala, prolongando ou encurtando palavras e sílabas. Formas de contornar este e outros problemas serão descritas nos comentários da figura 2.1.

2.3.1 CAPTAÇÃO DA VOZ

No início do processamento, o sinal acústico é transformado em uma grandeza elétrica, através de um microfone. Uma característica desejável do microfone é sua alta diretividade, de forma a reduzir os efeitos de ruidos de fundo e as conseqüências de possíveis efeitos acústicos indesejáveis, como o eco e a reverberação. Para evitar choques mecânicos e variação na distância entre a boca e o transdutor, em decorrência

manuseio, é comum o uso de microfones presos à cabeça.

Ao se utilizar o sinal de voz através da rede telefônica, deve ser considerado que a faixa é limitada às freqüências de 300-3400 Hz, e o sinal é corrompido por ruídos de comutação, eco, distorção de amplitude e fase, translação de freqüência e diafonia, por exemplo. Os efeitos de cada um desses fatores no desempenho dos sistemas, contudo, não são conhecidos [4], [11].

2.3.2 PRÉ-PROCESSAMENTO

O *pré-processamento* inclui uma série de procedimentos para preparar adequadamente o sinal de voz para análises posteriores.

Algumas vezes realizam-se, antes da digitalização, várias filtragens analógicas com o objetivo de compensar variações que possam ocorrer durante a aquisição do sinal, como por exemplo:

a) alteração da distância entre a boca e o microfone, corrigida com um controle automático de ganho com uma constante de tempo adequada;
b) distorções causadas pela linha telefônica ou outro meio de transmissão, compensada com um filtro inverso.

Para o processamento em computadores digitais, há a necessidade de se fazer uma conversão A/D adequada. O sinal é amostrado a uma taxa entre 6,4 kHz e 20,0 kHz e quantizado linearmente entre 12 e 16 bits [5]. A largura de faixa do sinal digitalizado é extremamente importante para a confiabilidade do sistema, uma vez que a eliminação das freqüências acima de 5 kHz causará muita dificuldade na análise de determinadas consoantes (/f/, /s/, por exemplo).

(a) (b)

FIG.2.2 - Segmentos de voz selecionados a) pela janela retangular; b) pela janela de Hamming. Estes exemplos correspondem a 25 ms da vogal /a/, amostrada a 8 kHz.

2.3.3 EXTRAÇÃO DE PARÂMETROS

Inicialmente o sinal é segmentado através de um janelamento temporal (figura 2.2). De forma geral, utilizam-se janelas de Hamming ou janelas Retangulares com uma duração entre 10 e 30 ms para análises a *curto prazo*, uma vez que os movimentos articulatórios podem ser considerados quase estacionários durante estes intervalos.

O janelamento pode ser realizado de duas maneiras básicas: dividindo-se a duração da palavra em um número fixo de intervalos ou utilizando-se intervalos de igual duração. No primeiro caso, o comprimento do intervalo mudará em função da duração de cada palavra enquanto, no segundo, a quantidade de segmentos é que será função da duração de cada palavra.

Uma peculiariedade dos sistemas de palavras isoladas é a necessidade, logo no começo do processamento, da *detecção das extremidades*, isto é, os instantes do início e fim da palavra.

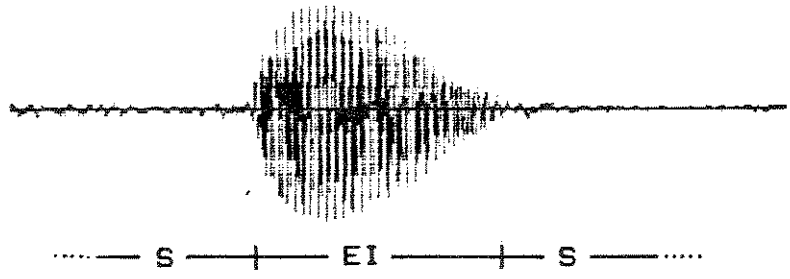


FIG.2.3 - Dificuldade de localização do início e fim da palavra /seis/.

Trata-se de um problema aparentemente simples (contornável apenas com a medida da variação da energia total) que, todavia, apresenta situações delicadas. Na palavra /seis/, conforme a figura 2.3, poderão surgir dificuldades na localização das extremidades, pois o /s/ pode ser confundido com o silêncio.

Após o janelamento, são estimados parâmetros com o objetivo de reduzir as redundâncias e eliminar informações desnecessárias do sinal. Como exemplo, as freqüências superiores a 4 kHz são eliminadas na representação de vogais, por possuirem informações que variam bastante com o locutor, não contribuindo para o reconhecimento do fonema; as vogais são caracterizadas apenas pelos três primeiros picos da envoltória do espectro a curto prazo, normalmente localizados abaixo de 4 kHz. Estes picos são denominados *formantes*, como será visto no capítulo 3.

A hipótese de se tentar fazer o reconhecimento pela análise da forma de onda seria inviável, a começar pelo grande número de combinações possíveis. Como exemplo, se o sinal for digitalizado, com amostragem a 8 kHz e quantização em 12 bits (2^{12} = 4096 níveis), o número de seqüências possíveis para cada milisegundo de voz seria de $4096^{8} \cong 79 \times 10^{27}$. Além disso, uma simples mudança na fase dos harmônicos, não percebida pelo ouvido, causa mudanças enormes na forma de onda.

Dentre os parâmetros de uso mais comum na análise de voz, podem ser citados [5]:

a) parâmetros de natureza temporal, como medidas do número de cruzamentos por zero;
b) parâmetros espectrais, obtidos a partir da Tranformada Discreta de Fourier ou com técnicas de Predição Linear;
c) parâmetros cepstrais, calculados com técnicas de Processamento Homomórfico ou a partir dos coeficientes LPC [3].

Todos estes parâmetros, com excessão dos cepstrais, serão utilizados no capítulo 4.

2.3.4 TRAÇOS

Traços são características mais sofisticadas associadas à voz, podem ter um caráter supra-segmental, extendendo-se por várias janelas, e são obtidos a partir da análise conjunta de vários parâmetros.

Sendo uma operação de redução de redundâncias, a obtenção de traços é delicada, podendo ocasionar erros irreparáveis. Por isso, em sistemas mais simples, os padrões geralmente são elaborados apenas com os parâmetros citados na seção anterior.

As características de *turbulência* de algumas consoantes (como em /s/, /ch/), de *nasalidade*, ou ainda o resultado das decisões surdo/sonoro, são traços comumente empregados em SRAV. Outros exemplos de traços são as categorias fonéticas (fricativo surdo, fricativo sonoro, vocálico, coarticulação e oclusão sonora) utilizadas neste trabalho para a classificação dos segmentos de voz.

Os blocos descritos até aqui, responsáveis pelo processamento do sinal desde a aquisição da voz, até a extração de pistas ou traços, são comumente denominados *Módulo Frontal*. O Módulo Frontal é responsável pela eliminação das supostas variabilidades e redundâncias da fala, desnecessárias ao reconhecimento de voz. Alguns sistemas mais sofisticados também incluem no Módulo Frontal processos de *normalização de locutor*, onde os parâmetros sofrem um escalonamento em função do sexo ou idade da pessoa.

2.3.5 PADRÕES

Em SRAV dependendes do locutor, a aquisição de padrões de referência é feita pelo usuário, colocando o sistema no modo de treinamento. Em sistemas que operam independentemente do locutor, os padrões de referência são estabelecidos a priori a partir da análise das características das vozes dos diversos usuários do sistema.

O número de segmentos utilizados para a análise do sinal de voz é limitado pelo tamanho do vocabulário e pelas características acústicas das palavras. Como exemplo, para distinguir as palavras /sim/ e /não/, um único segmento (toda a palavra) seria suficiente, pois /sim/ apresenta uma maior concentração de energia nas altas freqüências. Caso as palavras a serem reconhecidas sejam os dígitos, o número de segmentos analisados em cada palavra deverá ser maior, para distinguir a pronúncia de pares confusos, como /dois/ e /oito/ ou /treis/ e /seis/. Quando for possível a escolha do vocabulário, devem ser evitadas palavras que gerem confusão.

2.3.6 ALINHAMENTO DE TEMPO

Vamos imaginar que no sistema para reconhecimento de palavras isoladas da figura 2.1, cada palavra seja representada por uma seqüência de S segmentos, com P parâmetros por segmento.

Quando o sistema é colocado no modo reconhecimento, é realizada a análise do sinal em sua entrada até constatar o início de uma palavra; em seguida, é detectada a posição final e calculada a sua duração, T. Se a análise for efetuada com um número fixo de quadros, as amostras são então agrupadas em S intervalos de duração T/S.

A divisão da duração da palavra em um número fixo de segmentos pode ser prática quando o vocabulário é muito pequeno. Com vocabulários mais extensos, o número de segmentos deve ser maior, optando-se pela análise em intervalos iguais seguida de uma *normalização ou alinhamento de tempo* entre a palavra de teste e a de referência.

Em sua forma mais simples, conhecida por alinhamento *linear* de tempo (figura 2.4), a correspondência entre os quadros de teste e de referência é feita a partir de uma reta cuja inclinação é função do número total de segmentos de teste, S'. O padrão de teste i é comparado com o padrão de referência j, onde

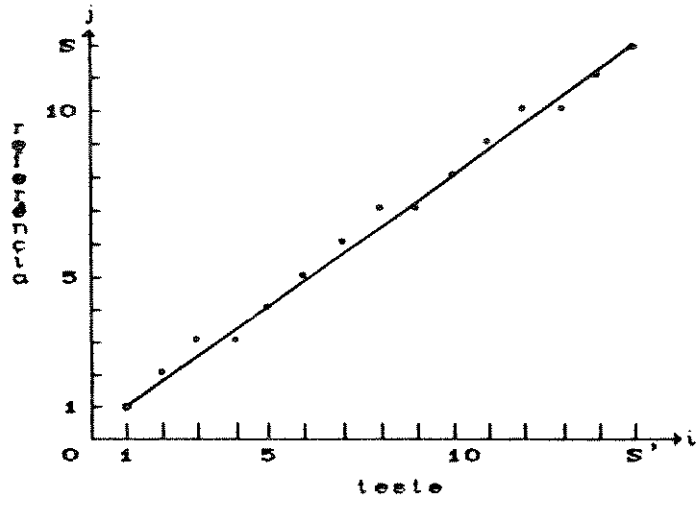


FIG. 2.4 - Alinhamento Linear de Tempo

$$j = Q(i \cdot S/S') \qquad (2.1)$$

$$1 \leq j \leq S$$
$$1 \leq i \leq S'$$

e Q denota a operação de quantização. Obviamente, são usados S níveis de quantização, quantidade equivalente ao número de padrões de referência por palavra.

O alinhamento linear de tempo não é capaz de compensar as oscilações na duração de cada sílaba ou fonema, confrontando trechos diferentes da palavra. Uma maneira mais eficiente e complexa consiste na realização do *alinhamento dinâmico de tempo* (figura 2.5). Os pontos inicial e final também são pré-fixados, mas a trajetória seguida a cada novo segmento de teste é definida de forma a minimizar alguma *distância acumulada* entre referência e teste. Normalmente são impostas restrições às trajetórias, de forma a otimizar o processo. Como exemplo de restrição, na comparação do quadro de teste i com a referência j, tem-se:

$$j_i - j_{i-1} = \begin{cases} 0,\ 1,\ 2, & \text{se } j_{i-1} \neq j_{i-2} \\ 1,\ 2, & \text{se } j_{i-1} = j_{i-2} \end{cases} \qquad (2.2)$$

$$1 \leq j \leq S$$
$$1 \leq i \leq S'$$

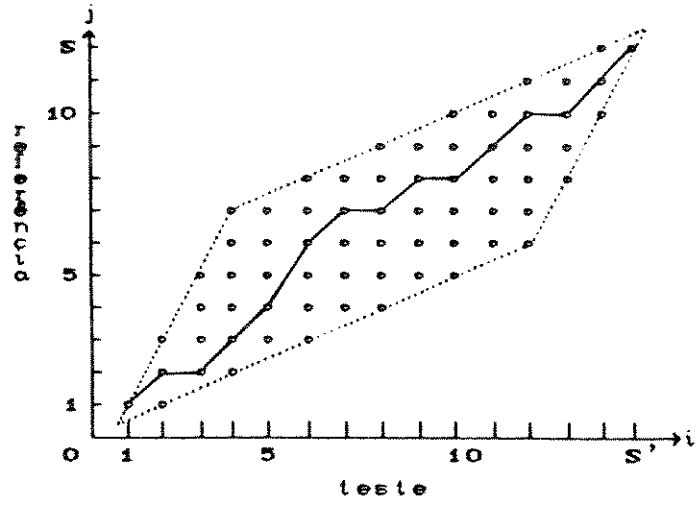


FIG.2.5 - Alinhamento Dinâmico de tempo

limitando a possibilidade de ocorrência da função de alinhamento à região interna do paralelogramo da figura 2.5.

2.3.7 DISTÂNCIAS

De posse dos padrões de teste procede-se à determinação de qual é o padrão de referência que mais lhe aproxima.

Aos padrões de cada palavra podem ser associadas matrizes de dimensões SxP. As matrizes de referência e de teste da k-ésima palavra são definidas por:

$$R_k = [r_1\ r_2\ \ldots\ r_S] \quad e \quad T_k = [t_1\ t_2\ \ldots\ t_{S'}] \qquad (2.3a)$$

onde

$$r_j^T = [r_1\ r_2\ \ldots\ r_P] \quad (1 \le j \le S) \qquad e \qquad t_i^T = [t_1\ t_2\ \ldots\ t_P] \quad (1 \le i \le S') \qquad (2.3b)$$

são vetores correspondentes às representações das palavras de referência e teste através dos parâmetros r_j e t_i,

respectivamente. A relação entre i e j é dada pela equação 2.1 (alinhamento linear de tempo) ou pela equação 2.2 (alinhamento dinâmico de tempo).

Para a comparação entre as matrizes de referência e teste podem ser calculadas várias *medidas de distância*.

A Distância Euclidiana entre as matrizes R_k e T_k é a mais popular, sendo definida por

$$d(R_k,T_k) = \sqrt{\sum_{i=1}^{S'} \sum_{l=1}^{P} (r_{jl} - t_{il})^2} \qquad (2.4)$$

onde i e j estão relacionados pelas equações 2.1 ou 2.2, S' é o número de vetores de teste e P é o número de parâmetros utilizados.

Outras distâncias comuns em reconhecimento de voz são a Distância de Mahalanobis (originária da teoria estatística de decisão), a Distância de Itakura-Saito e sua simplificação, a Razão Log-Probabilidade (específicas para parâmetros LPC) [6], [7]. Deve ser ressaltado que a eficiência de cada distância está intimamente relacionada com os tipos de parâmetros envolvidos.

2.3.8 DECISÃO

Como foi visto anteriormente, o reconhecimento é feito determinando-se o padrão de referência que produz a menor distância ao padrão de teste. Podem ocorrer situações em que uma entrada está praticamente eqüidistante de dois ou mais padrões, impossibilitando uma decisão robusta. Para contornar tais casos, é comum a inclusão de *limiares de decisão*, de tal forma que uma referência somente será escolhida como saída se *i*) a medida de sua distância à entrada não ultrapassar determinado *limiar máximo* e *ii*) a diferença entre as duas menores distâncias for superior a um *limiar mínimo*. Quando o sistema não puder dar uma decisão confiável, poderá ser solicitada uma nova pronúncia da palavra (*rejeição*) ou adiada a decisão, até que novas palavras

estabeleçam um contexto que possibilite uma decisão confiável. Neste último caso, será necessária a inclusão de outras fontes de conhecimento lingüístico.

Até o momento, supôs-se a existência de apenas um padrão para cada palavra, decidindo-se pela referência que resulta na menor distância; esta estratégia é denominada *regra do vizinho mais próximo*. Alguns sistemas, como os multiusuários, possuem mais de um padrão de referência para cada palavra, utilizando regras mais elaboradas para a decisão [7].

2.3.9 SAÍDAS

Tendo em vista que os percentuais de acerto dificilmente atingem 100%, o uso prático do reconhecimento de voz em algum sistema de comando ou controle somente deve ser introduzido em situações onde erros podem ser tolerados ou corrigidos sem maiores conseqüências. Aplicações em *controle verbal* com um vocabulário reduzido são viáveis para a entrada de dados em um computador, auxílio a pessoas incapacitadas fisicamente e encaminhamento de chamadas telefônicas, por exemplo.

Sistemas que possibilitam a *conversão da fala em texto* também são viáveis, mas a implementação é um tanto complicada, exigindo vários níveis de análise.

2.4 MAIS UM POUCO DE HISTÓRIA

Em novembro de 1971, iniciou-se um programa norte-americano de pesquisas conhecido por Projeto ARPA (Advanced Research Projects Agency of the Department of Defense) [8], que representou um marco na evolução dos SRAV. Foi estabelecida uma série de objetivos que foram perseguidos por vários grupos durante os 5 anos de duração do projeto. As metas eram muito ambiciosas comparadas com os sistemas de reconhecimento de palavras isoladas existentes nos laboratórios nesta época.

Resumidamente, desejava-se um sistema com as seguintes especificações:

a) aceitar a fala conectada de vários indivíduos num ambiente silencioso;

b) realizar a compreensão da fala a partir de frases montadas com um vocabuláro de 1000 palavras, tolerando-se até 10% de erros.

Por *compreensão da fala* entenda-se que os sistemas deveriam realizar determinadas tarefas especificadas pelas frases, isto é, importava a análise global das entradas, tolerando-se erros em algumas palavras, desde que eles não alterassem o significado das sentenças.

No início do projeto foi tentada a adaptação das técnicas de reconhecimento de palavras isoladas da década de 60, com pouco sucesso. Os trabalhos anteriores foram realizados, na sua maior parte, por engenheiros e físicos que tinham completo domínio sobre as máquinas e técnicas computacionais, mas muito pouco conhecimento dos fenômenos lingüísticos e contextuais envolvidos na compreensão da fala. De fato, os sistemas de reconhecimento descritos na seção 2.3 são meros detectores de sinais, apresentando resultados semelhantes se usados para sons de outra natureza. Baseavam-se em suposições equivocadas de que a fala é uma seqüência linear e invariante de eventos acústicos, dando igual importância a cada um deles no processo de decisão.

Para o desenvolvimento do projeto, novas capacidades foram inseridas nas máquinas, inspirando-se na maneira de produzir e ouvir a fala.

2.5 SISTEMAS PARA RECONHECIMENTO E COMPREENSÃO DE FALA CONTÍNUA

Durante os cinco anos do Projeto ARPA, houve a aproximação de engenheiros e lingüistas, recorrendo-se a uma série de trabalhos feitos nas décadas anteriores [9], [10], em torno de teorias e modelos de produção e percepção da fala, que

até então não eram utilizados nos SRAV.

A figura 2.6 é uma representação suscinta dos princípios utilizados nos sistemas do Projeto ARPA, destacando-se a introdução do Módulo Lingüístico, que está dividido em duas partes.

A parte inferior é responsável pela *análise* dos parâmetros e traços extraídos no módulo frontal, com a finalidade de obter alguma forma de *representação fonética* da fala, valendo-se de diversos tipos de regras, como:

a) *regras fonológicas* para modelar as transformações da escrita em fala. Como exemplo, o /s/ no final de palavra, seguido por vogal, normalmente é transformado em /z/: "as outras" é pronunciado /azoutras/;

b) *regras fonotáticas* representando as restrições do idioma à ocorrência de determinadas combinações fonéticas. No português, por exemplo, não existe a seqüência /ft/ no início de palavras.

Na parte inferior do Módulo Lingüístico (figura 2.6) também são extraídas informações *prosódicas*, isto é, dados relativos à intonação e acentuação das frases e palavras.

A parte superior realiza a *interpretação* dos dados. A representação fonética obtida em outros módulos é analisada,

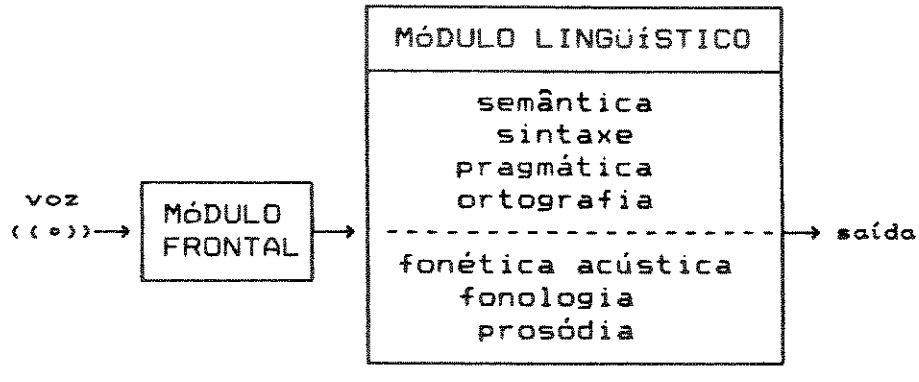


FIG 2.6 - Sistema de Reconhecimento de Voz utilizando fontes de conhecimento lingüístico.

criando-se hipóteses sobre as possíveis seqüências de palavras que a teria produzido. Para tal, são utilizadas regras e estratégias particulares do idioma e do vocabulário, através do conhecimento lingüístico:

a) a *semântica* é responsável pelos problemas de significado das palavras e sentenças, que podem ter uma acepção diferente, dependendo do contexto;
b) a *pragmática* procura exatamente a identificação do contexto estabelecido pelas palavras;
c) a *sintaxe* avalia a estrutura gramatical das frases;
d) a *ortografia* fornece maneiras para a transformação de uma seqüência de símbolos fonéticos em palavras corretamente escritas; esta passagem pode ser feita por consulta a uma tabela de correspondência fonético/ortográfica ou pela aplicação das regras de ortografia específicas do idioma.

Na implementação desses sistemas surgiram várias estratégias de controle para a solução dos problemas de interação entre os módulos e no tratamento de decisões conflitantes de diferentes fontes de conhecimento. Os *frames* (quadros), estruturas de partilhamento de dados utilizadas em Inteligência Artificial, tiveram sua origem no sistema Hearsay-II [12], [13].

Ao final do projeto ARPA, em novembro de 1976, apenas quatro sistemas foram apresentados (tabela 2.1). O Sistema Harpy, único a atender às especificações, realizava uma etapa de

SISTEMA	COMPREENSÃO
Harpy	95%
Hearsay-II	74%
Hwin	44%
SDC	24%

TAB.2.1 Comparação entre os sistemas do Projeto ARPA [8].

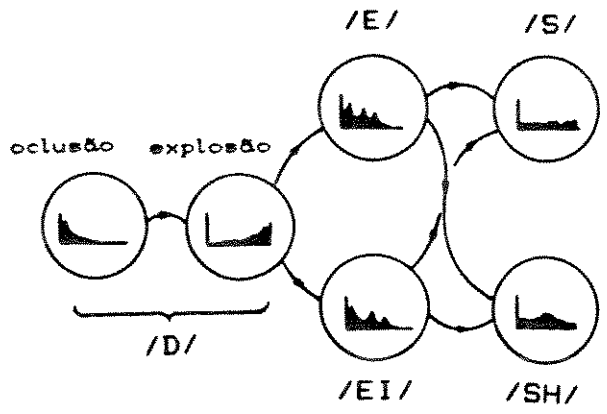


FIG.2.7 - Utilização da estrutura do Sistema Harpy para representar as possíveis pronúncias do número 10.

treinamento para cada usuário (leitura de 20 sentenças) e mapeava todos os segmentos de 10 ms em um conjunto de 98 espectros. Foi montada uma rede com 15000 estados, onde cada estado (ou nó) era representado por um dos 98 espectros; alguns nós eram assinalados com limites de duração, efetuando-se, desta forma, uma normalização de tempo. As ramificações foram reduzidas, incluindo na rede (figura 2.7) apenas as seqüências acústicas das frases aceitáveis. Na análise de uma sentença de entrada, determinava-se a seqüência de estados da rede que melhor representava a entrada; em cada nó era escolhida uma trajetória, optando-se pela melhor alternativa local. Observe-se que as considerações de ordem lingüística estão ocultas nas transições da rede.

Se a avaliação desses cinco anos de pesquisas for feita apenas em função da tabela 2.1, concluir-se-á que os resultados foram bastante modestos. Todavia, a contribuição foi imensa à medida que se conscientizou da importância da pesquisa básica dos fenômenos da fala, assim como da necessidade de um ferramental matemático poderoso para processar grande quantidade de informação.

Outro aspecto importante dos anos 70 foi o início da fabricação de pequenos sistemas comerciais para reconhecimento de voz, que lentamente saíram dos laboratórios e ganharam espaço no mercado americano. As características da quase totalidade desses

QUANTO AO LOCUTOR	TIPO DE PRONÚNCIA	VOCABULÁRIO MÉDIO	PRECISÃO (%)
dependente	isolada	100	95
independente	isolada	15	95

TAB. 2.2 - Características básicas dos primeiros **SRAV** comerciais. Os percentuais de acerto são apenas uma indicação grosseira, devendo ser cuidadosamente avaliados em função do vocabulário, do ruído e da regularidade da pronúncia.

produtos (tabela 2.2), eram [14]:

a) reconhecimento de palavras isoladas;

b) dependência com o locutor;

c) vocabulários médio de 100 palavras;

d) precisão em torno de 95% com baixos níveis de ruído.

2.6 ATUALIDADE

Uma característica do Reconhecimento de Voz nos dias de hoje é a participação ativa de grandes empresas nas pesquisas. Os interesses se dividem em:

a) conversão de fala em texto para grandes vocabulários e apenas um usuário (20000 palavras ditadas em ambiente de pouco ruído para a confecção de cartas e documentos). Nesta linha destacam-se as atividades da IBM, com seus sistemas experimentais Tangora 5000 e Tangora 20000 [15] [20];

b) reconhecimento de dígitos para encaminhamento de ligações telefônicas e serviço automático de auxílio à lista (fala natural, independente do locutor). A AT&T Bell Laboratories estuda o reconhecimento de dígitos através de ligações telefônicas, em condições normais de ruído, fala natural e de maneira realmente independente do locutor [16].

Há também o interesse em combinar reconhecimento e

síntese de voz, de forma a construir uma máquina capaz de realizar a tradução automática entre idiomas.

O uso popular do reconhecimento de voz vai tornando-se possível à medida que os algoritmos evoluem, o custo dos microprocessadores e memórias é reduzido, e circuitos integrados dedicados são construídos para o cálculo de parâmetros espectrais, coeficientes LPC e realização do Alinhamento Dinâmico de Tempo. Os modelos comerciais típicos apresentam um vocabulário entre 100 e 1000 palavras dependentes do locutor [6], [15].

O desenvolvimento e aplicação de modelos matemáticos à fala vem obtendo um grande sucesso, destacando-se a representação por Cadeias de Markov Ocultas (fig 2.8). Estes modelos, utilizados pela IBM e pela AT&T Bell Laboratories, são semelhantes aos empregados no Sistema Harpy, inserindo restrições e probabilidades às transições. Podem ser utilizados para a representação de frases, palavras, fonemas ou quaisquer outras sub-unidades. O termo *oculto* foi adotado porque os modelos são obtidos matematicamente a partir da análise probabilística da voz, e não por uma representação explícita dos fenômenos da produção e percepção da fala [17].

Finalmente, o uso de Redes Neurais em Reconhecimento de Voz está dando seus primeiros passos e os resultados são bastante promissores. Uma característica interessante dessas estruturas é a capacidade de aprendizagem automática, possibilitando a descoberta de relações não suspeitadas entre os parâmetros acústicos [18].

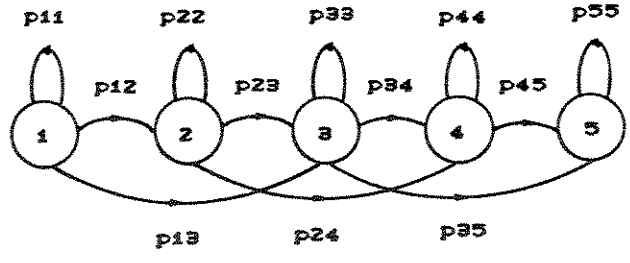


FIG.2.8 - Cadeia de Markov de cinco estados, com restrição nas transições. Cada estado representa um conjunto de espectros e os números em cada arco indicam a probabilidade da transição.

2.7 DISCUSSÃO

Os vários problemas que um Sistema de Reconhecimento de Voz deve solucionar, possibilitando seu uso prático, podem ser resumidos na tabela 2.3. Independente da forma utilizada para a estruturação dos conhecimentos de ordem lingüística, isto é, se as decisões serão tomadas a partir da comparação de padrões, regras explícitas ou outras formas de representar o conhecimento, a solução de cada um dos pontos críticos depende de um bom Módulo Frontal.

A utilização de Cadeias de Markov ou Inteligência Artificial após o processamento acústico, pode ser otimizada se se introduzirem restrições nos modelos, baseando-se em conhecimentos de fonética acústica, de resultados de testes de percepção de fala e da compreensão do movimento articulatório.

O próximo capítulo discutirá os aspectos da produção e percepção da fala, introduzindo os modelos que foram utilizados em nossos experimentos.

1. Sensibilidade ao ruído ambiente
2. Variação fonético-acústica
3. Segmentação adequada do sinal
4. Normalização de tempo
5. Normalização de locutor
6. Forma de representação do léxico
7. Variações fonológicas
8. Detecção e correção de erros
9. Utilização da prosódia

TAB.2.3 - Problemas que devem ser considerados na implementação de um Sistema de Reconhecimento Automático de voz [19].

2.8 REFERÊNCIAS

[1] K. H. Davis et al,"*Automatic Recogniton of Spoken Digits*", Journal of the Acoustical Society of America, vol 24 (6), pp 637-642 (1952);

[2] S. R. Hyde, "*Automatic Speech Recognition: A Critical Survey and Discussion of the Literature*", em "Human Communication: A Unified View", E. E. David Jr. & P. B. Denes, Eds. McGraw Hill, NY (1972);

[3] W. Koenig, H. K. Dunn and L. Y. Lacy, "*The Sound Spectrograph*", Journal of the Acoustical Society of America, vol 17, pp 19-49 (1946);

[4] Geoff Bristow, ed., "*Electronic Speech Recognition: Techniques, Technnology & Applications*", McGraw Hill, N.Y. (1986);

[5] L. R. Rabiner & R. W. Shafer, "*Digital Processing of Speech Signals*", Prentice Hall, Inc., N.J. (1978);

[6] D. O'Shaughnessy, "*Speech Communication: Human and Machine*", Addison-Wesley Publishing Company (1987);

[7] L. R. Rabiner & S. E. Levinson, "*Isolated and Connected Word Recognition - Theory and Selected Applications*", IEEE Transactions on Communications, vol 29, (5), pp 621-659, (1981);

[8] D. H. Klatt, "*Review of the ARPA Speech Understanding Project*", Journal of the Acoustical Society of America, vol 62, pp 1345-1366 (1977);

[9] N. R. Dixon & T. B. Martin, ed., "*Automatic Speech and Speaker Recognition*", IEEE Press (1979);

[10] Gunnar Fant, ed., "*Proceedings of the Speech Communication Seminar - Vol.3: Speech Perception and Automatic Recognition*", Almqvist & Wiksell International, Stockholm (1974);

[11] D. R. Reddy, "*Speech Recognition by Machine: A Review*", Proceedings of the IEEE, vol 64, pp 501-531 (1976);

[12] V. R. Lesser et al, "*Organization of the Hearsay II Speech Understanding System*", IEEE Transactions on Acoustic, Speech and Signal Processing, vol 23, pp 11-24 (1975);

[13] E. Rich, "*Inteligência Artificial*", McGraw Hill, SP (1988);

[14] J. P. Cater, "*Electronically Hearing: Computer Speech Recognition*" (cap 9), SAMS, Indiana (1984);

[15] A. Averbuch et al, "*An IBM PC Based Large-Vocabulary Isolated-Utterance Speech Recognizer*", IBM Research Report, (1986);

[16] J. G. Wilpon, "*A Study on the Ability to Automatically Recognize Telephone-Quality Speech From Large Customer Populations*", ATT&T Technical Journal, vol 64 (2), pp 423-451 (1985);

[17] S. E, Levinson, "*Structural Methods in Automatic Speech Recognition*", Proceedings of the IEEE, vol 73 (11), pp 1625-1650 (1985);

[18] J. L. Elman & D.Zipser, "*Learning the Hidden Structure of Speech*", Journal of the Acoustical Society of America, vol 83 (4), pp 1615-1626 (1988);

[19] D. H. Klatt, "*Speech Perception: a Model of Acoustic-Phonetic Analysis and Lexical Access*", Journal of Phonetics, vol 7, pp 279-312 (1979);

[20] F. Jelinek, "*The Development of an Experimental Discrete Dictation Recognizer*", Proceedings of the IEEE, vol. 73 (11), pp 1616-1624 (1985).

CAPÍTULO 3

A CADEIA DA FALA:

MODELOS PARA A PRODUÇÃO E PERCEPÇÃO DO SINAL DE VOZ

3.1 INTRODUÇÃO

Para uma análise das diversas etapas envolvidas na comunicação pela voz, vamos recorrer à representação simplificada da figura 3.1 [1].

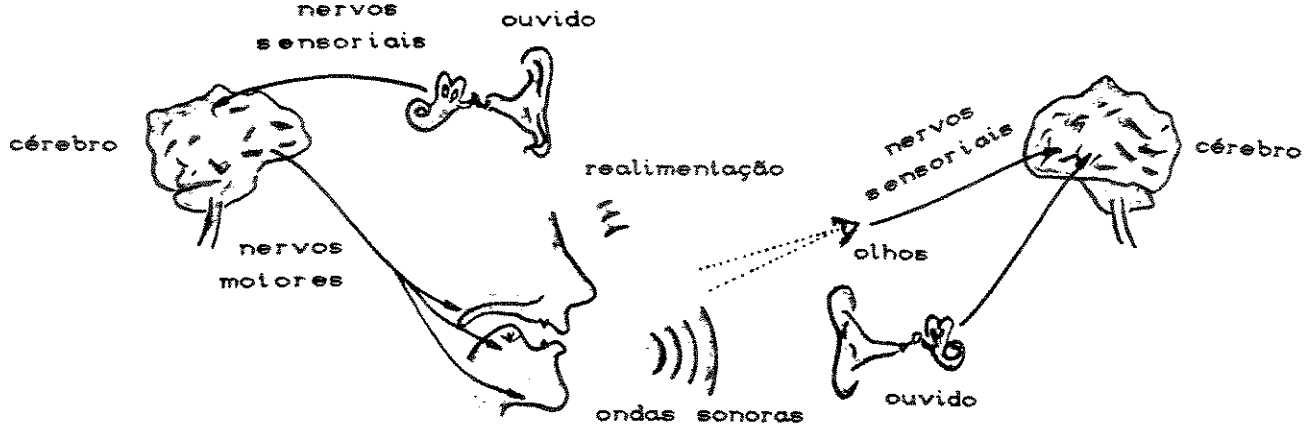


FIG.3.1 - Cadeia da Fala, representando os diversos níveis do processo de comunicação [1]. Note-se que o locutor ouve o que diz. Esta realimentação permite um controle eficiente na produção dos movimentos articulatórios. Eventualmente, o ouvinte pode estar vendo os gestos articulatórios do locutor.

De acordo com a figura 3.1, quando o locutor deseja transmitir alguma mensagem, as palavras são selecionadas e organizadas segundo a gramática de sua língua. Esta primeira etapa de transformação se passa no nível linguístico da produção da fala. Em seguida, o cérebro comanda o envio de estímulos nervosos para a realização do movimento coordenado dos músculos; a atividade neuromuscular, correspondente ao nível fisiológico da cadeia da fala, resulta na produção de ondas de pressão sonora que se propagam pelo ar até o ouvinte. A voz, nível acústico do processo de comunicação, é então analisada no ouvido, transformada novamente em impulsos nervosos e, finalmente, a mensagem é recuperada no cérebro do ouvinte, através de uma complexa cadeia associativa, cujo funcionamento é ainda pouco compreendido.

Este capítulo tem por objetivo discutir a produção do sinal acústico e alguns aspectos da sua percepção, com o intuito de introduzir modelos para a análise prática da voz. Uma abordagem matemática rigorosa dos processos físicos pode ser encontrada em [2] e [10].

3.2 FISIOLOGIA DA PRODUÇÃO DA FALA

Todos os sons da língua portuguesa são produzidos com o ar sendo expelido dos pulmões. Algumas línguas, como o árabe, têm fonemas produzidos durante a inspiração. Para gerar o som desejado, o locutor exerce uma série de controles sobre o aparelho fonador, representado na figura 3.2, produzindo a configuração articulatória e a excitação apropriadas. A configuração articulatória é determinada pelo posicionamento dos *articuladores*, como a língua, lábios, maxilar, dentes ou úvula.

O trato vocal é o conjunto de cavidades que se estendem desde as cordas vocais até os lábios, tendo um comprimento médio de 17 cm para um homem adulto. Seu formato é determinado pelo posicionamento dos órgãos articuladores.

Na produção de consoantes nasais ou vogais nasalizadas,

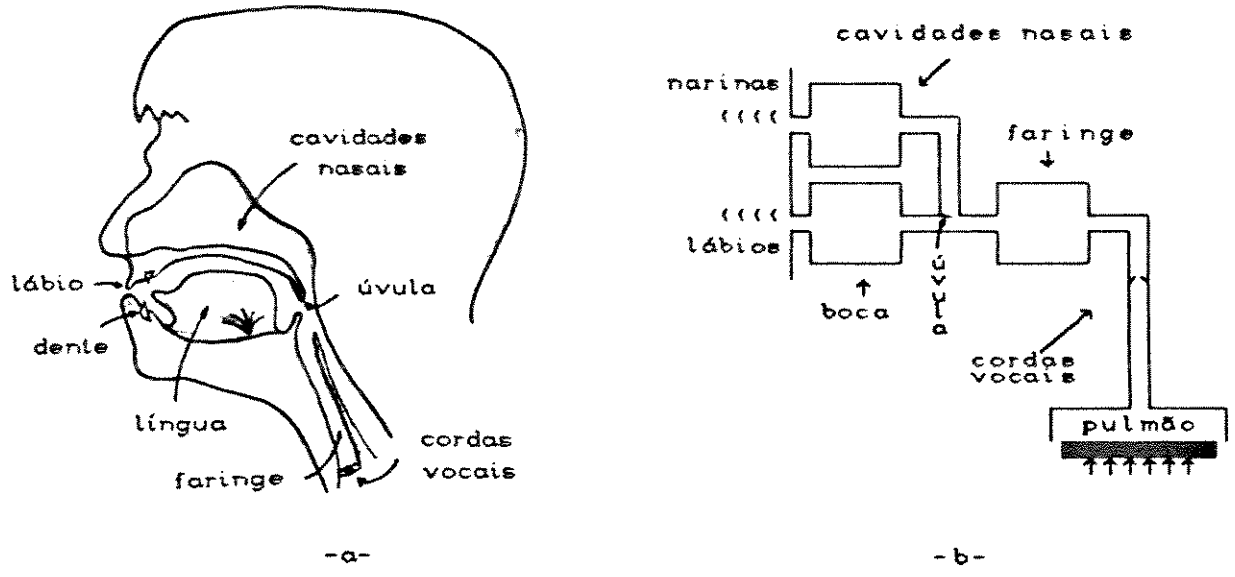


FIG. 3.2 - Aparelho fonador: a) anatomia; b) modelo acústico.

o fluxo de ar é total ou parcialmente dirigido ao trato nasal, sob o controle da úvula. Observe-se que o formato do trato nasal, não pode ser alterado voluntariamente pelo locutor.

Após a filtragem, determinada pela conformação do aparelho fonador, o fluxo de ar injetado pelos pulmões é acoplado ao ambiente externo por intermédio dos orifícios dos lábios e/ou narinas.

3.3 CARACTERÍSTICAS DE ALGUNS SONS DA FALA [4], [11]

A simples passagem do ar pelo aparelho fonador, como no processo natural da respiração, não produz sons característicos de fala; para sua produção, é necessário que o aparelho fonador seja excitado de maneiras específicas. Na língua portuguesa há três formas básicas de excitação, resultando nos sons *sonoros*, *fricativos* e *explosivos*, que serão discutidos a seguir.

3.3.1 SONS SONOROS

O fluxo de ar vindo dos pulmões é controlado pela abertura e fechamento das cordas vocais ou, mais apropriadamente, *dobras vocais* (figura 3.3), ligamentos semelhantes a dois lábios que podem ser tensionados e aproximados sob o controle do locutor. A abertura entre as dobras é denominada glote. Estando a glote totalmente fechada, o fluxo de ar originário dos pulmões é interrompido e a pressão sub-glótica aumenta até que as dobras vocais sejam separadas, liberando o ar pressionado, gerando um pulso de ar de curta duração. Com o escoamento do ar, a pressão glótica é reduzida, permitindo uma nova aproximação das cordas vocais. O processo se repete de uma forma quase periódica. A *freqüência média desses pulsos é denominada freqüência fundamental de excitação*, f0 e o *período de pitch*, P, é definido por

$$P = 1/f0 \qquad (3.1)$$

Tendo em vista a pequena abertura da glote em relação às cavidades superiores do aparelho fonador, considera-se que a vazão glótica não é influenciada pelos movimentos dos

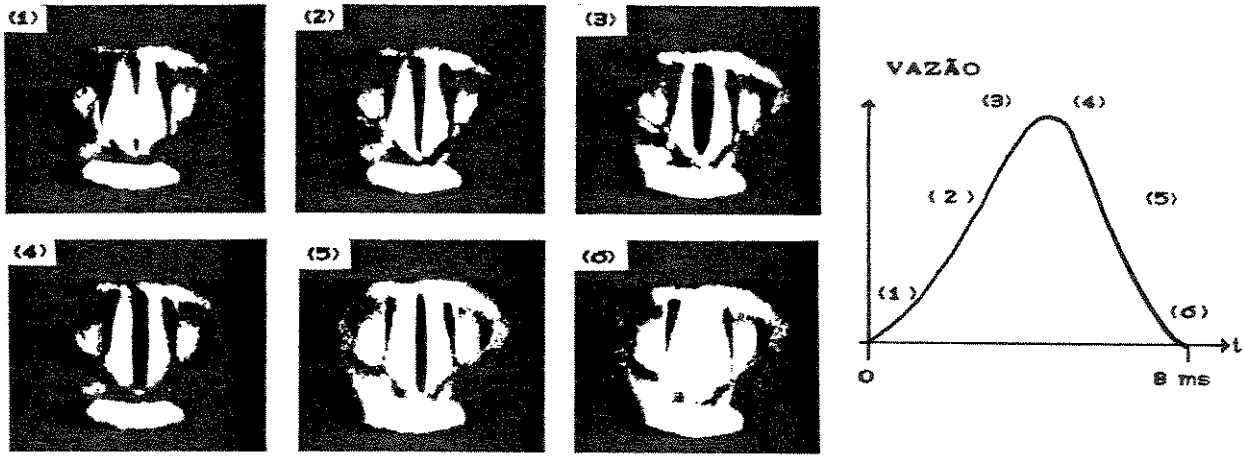


FIG.3.3 - a) Ciclo de vibração das cordas vocais [2]; b) vazão de ar correspondente aos vários instantes de abertura da glote.

articuladores. Em outras palavras, o sistema glotal pode ser visto como uma fonte de corrente de alta impedância acoplada ao trato vocal. Fazendo-se uma analogia com a eletricidade, a pressão corresponde à tensão e a vazão à corrente.

As vogais (figura 3.4), cujo grau de nasalização é determinado pelo abaixamento da úvula, são exemplos típicos de sons sonoros. Observe-se que os espectros apresentam uma componente de variação "rápida" e quase regular, relativa à excitação, e uma envoltória "lenta", dada pela resposta em freqüência do trato vocal, que modela o espectro da excitação.

Algumas consoantes, como a lateral /l/ e a nasal /m/, também são produzidas com a excitação glotal (figura 3.5).

Todos os sons descritos nessa seção são agrupados na categoria dos sons *vocálicos* ou *ressoantes*, e são caracterizados pela presença da vibração glotal e uma maior concentração de energia na faixa de 200-3500 Hz. Será visto, em seguida, uma segunda forma de excitar o aparelho fonador e os sons resultantes dessa excitação.

3.3.2 FRICATIVOS SURDOS

Na produção dos sons fricativos surdos, ou sibilantes, a glote permance aberta, não havendo vibração das cordas vocais. Entretanto, em algum ponto do trato vocal é realizado um estreitamento, pela aproximação de dois articuladores. Como exemplo, na produção do /f/ (figura 3.6), lábios e dentes são ligeiramente pressionados, resultando em uma passagem estreita para o ar; o fluxo de ar torna-se turbulento nas imediações da constricção, excitando as cavidades do trato vocal. Esta excitação é de baixa intensidade, com um espectro plano na faixa de áudio (ruído branco).

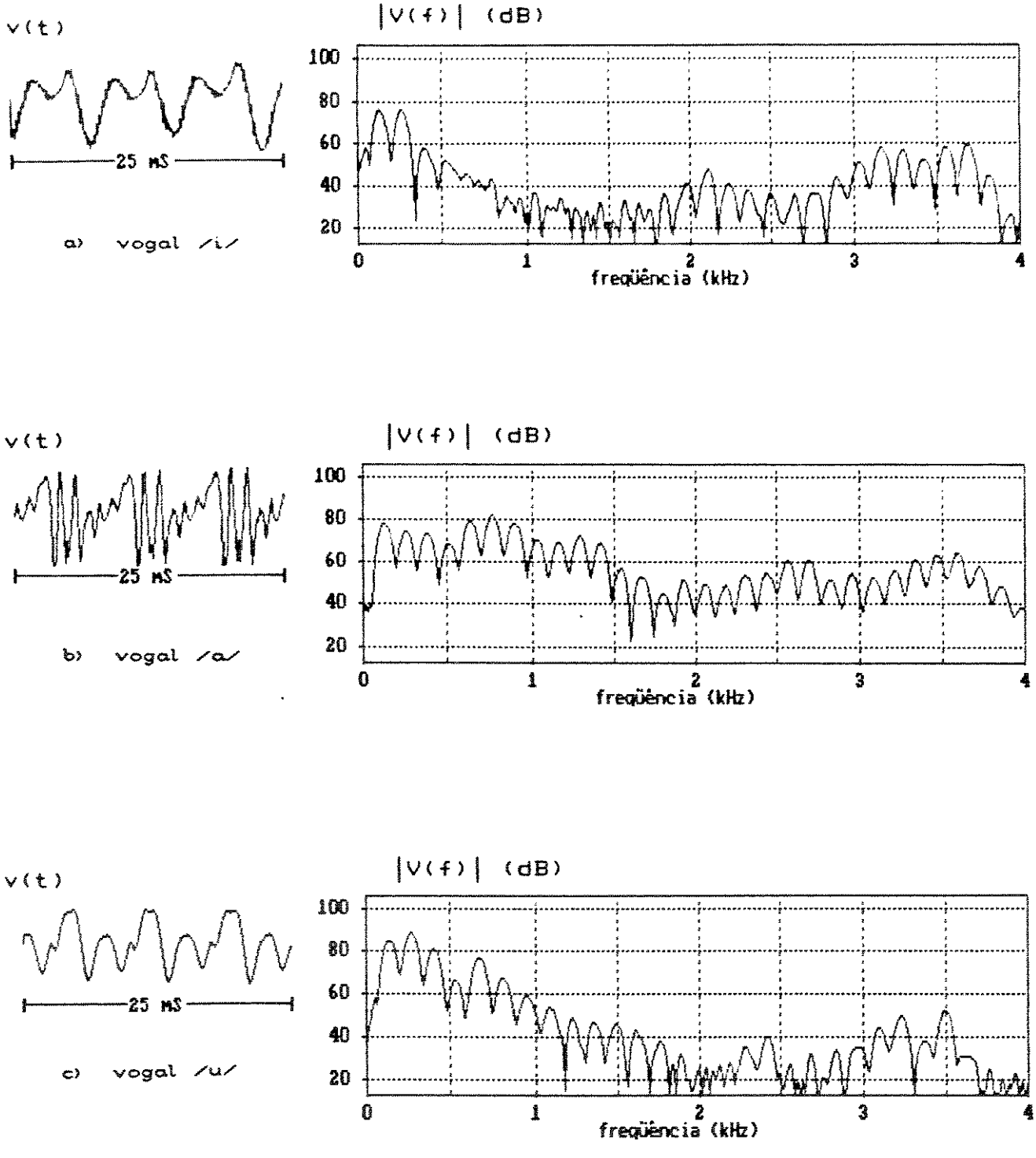


FIG. 3.4 - Forma de onda e espectro a curto prazo de algumas vogais não nasalizadas: a) vogal /i/; b) vogal /a/; c) vogal /u/.

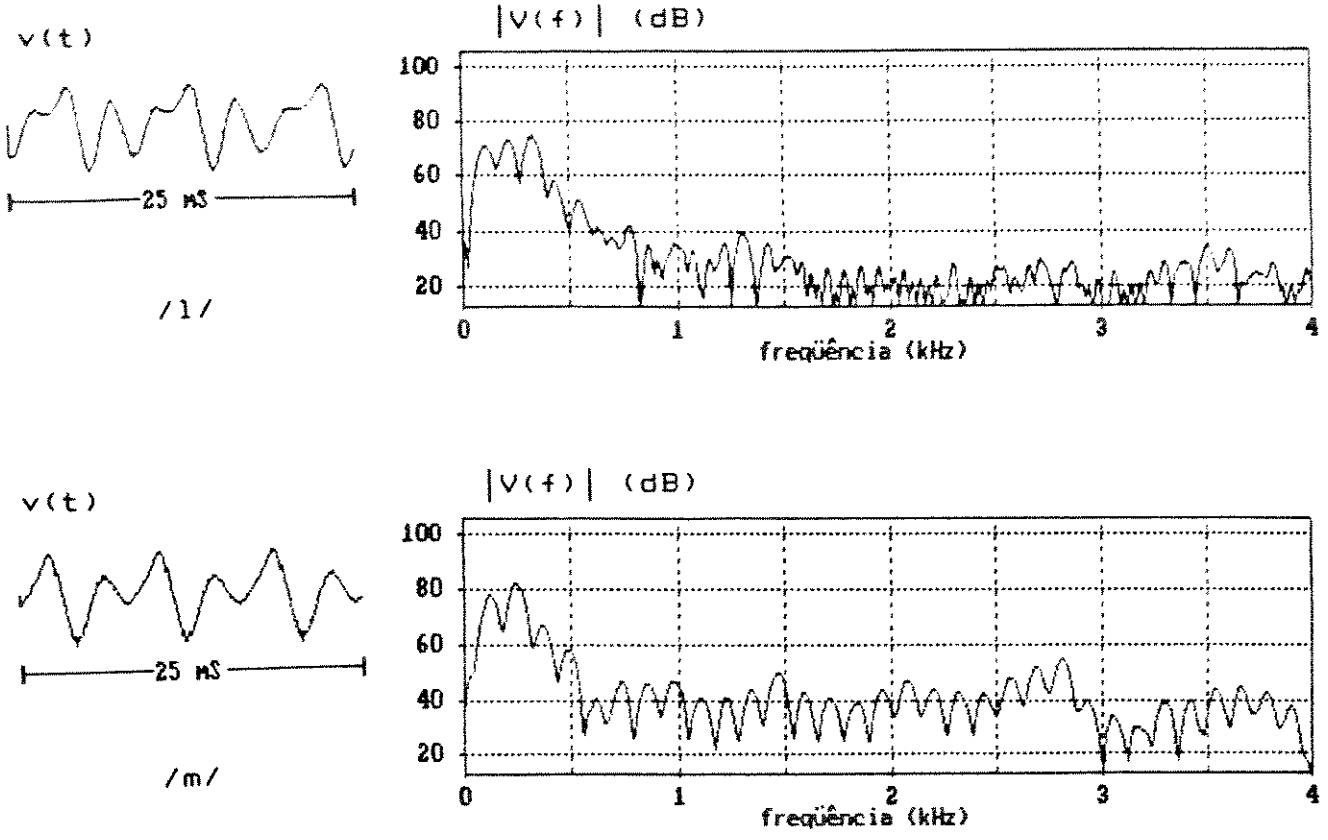


FIG. 3.5 - Forma de onda e espectro das consoantes /l/, em /lá/, e /m/ em /má/.

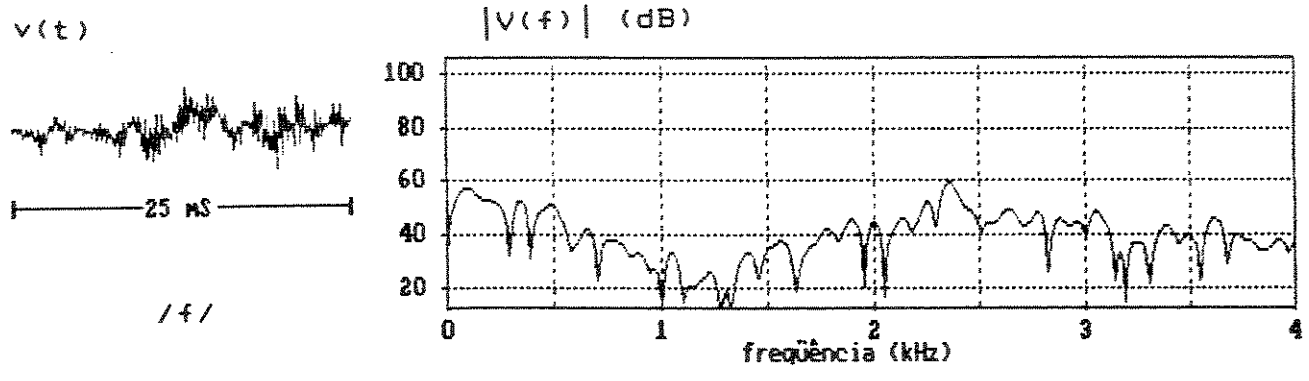


FIG. 3.6 - Forma de onda e espectro do fricativo /f/ em /fá/.

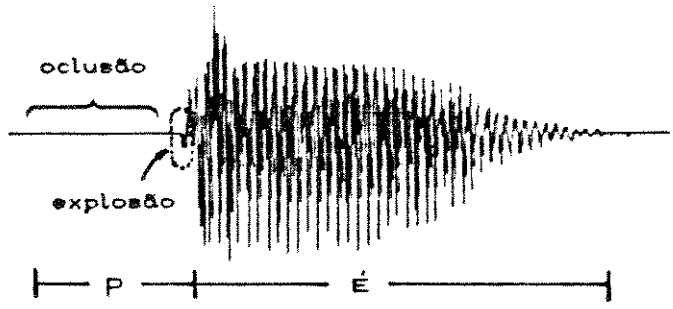


FIG.3.7- Forma de onda da sílaba /pé/, com os detalhes da produção do fonema /p/.

3.3.3 EXPLOSIVOS

A última maneira básica de excitar o trato vocal é realizada como em /p/ (figura 3.7), /t/ ou /k/, onde o ar é totalmente dirigido à boca, que se encontra completamente fechada. Com o aumento da pressão, a oclusão é rompida bruscamente, gerando um pulso que excita o aparelho fonador; a explosão é acompanhada de um movimento rápido dos articuladores em direção à configuração do som seguinte. Esta excitação é também denominada *excitação transitória*.

3.3.4 SONS COM EXCITAÇÃO MISTA

Os sons *fricativos sonoros*, como /j/ (figura 3.8), /v/ e /z/, são produzidos combinando-se a vibração das cordas vogais e a excitação turbulenta. Nos períodos de máxima pressão glótica o escoamento pela obstrução torna-se turbulento, gerando o caráter fricativo do som; assim que a pressão glótica cai abaixo de certo valor, extingue-se o escoamento turbulento de ar e as ondas de pressão têm um comportamento mais suave.

Os sons *oclusivos (ou explosivos) sonoros*, /b/ (figura 3.9), /d/ e /g/, são produzidas de forma semelhante aos correspondentes não sonoros, /p/, /t/ e /k/, havendo, todavia, vibração das cordas vocais durante a fase de fechamento da cavidade oral.

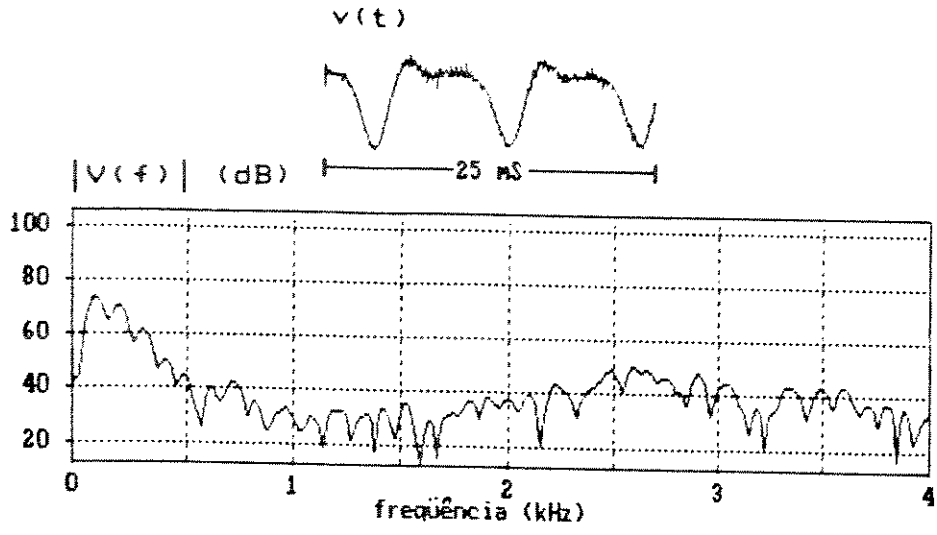


FIG. 3.8 - Forma de onda e espectro do fonema /j/ em /já/

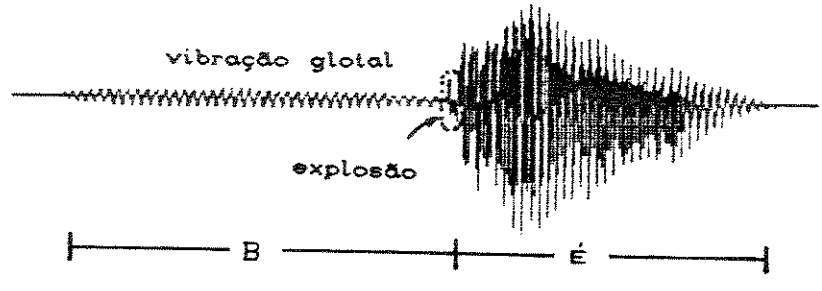


FIG.3.9 - Forma de onda da sílaba /bé/, com os detalhes da produção do fonema /b/.

3.4 MODELOS ANALÓGICOS PARA A PRODUÇÃO DA FALA

Baseando-se nas discussões anteriores e incluindo-se os efeitos da irradiação nos lábios e narinas, que também são considerados independentes da conformação do aparelho fonador, o espectro do sinal de voz, $V(\Omega)$, $\Omega = 2\pi f$ [rad/s], pode ser dado por:

$$V(\Omega) = E(\Omega) \cdot F(\Omega) \cdot T(\Omega) \qquad (3.2)$$

onde

$E(\Omega)$ = espectro da excitação;
$F(\Omega)$ = resposta em freqüência do trato vocal;
$T(\Omega)$ = impedância de carga dos lábios e/ou narinas.

Esta relação está representada de forma equivalente no diagrama em blocos da figura 4.10, que será estudado detalhadamente a seguir.

3.4.1 MODELO DA EXCITAÇÃO

O bloco referente à excitação, $E(\Omega)$, está dividido em duas partes, correspondentes ao modelamento da vibração das

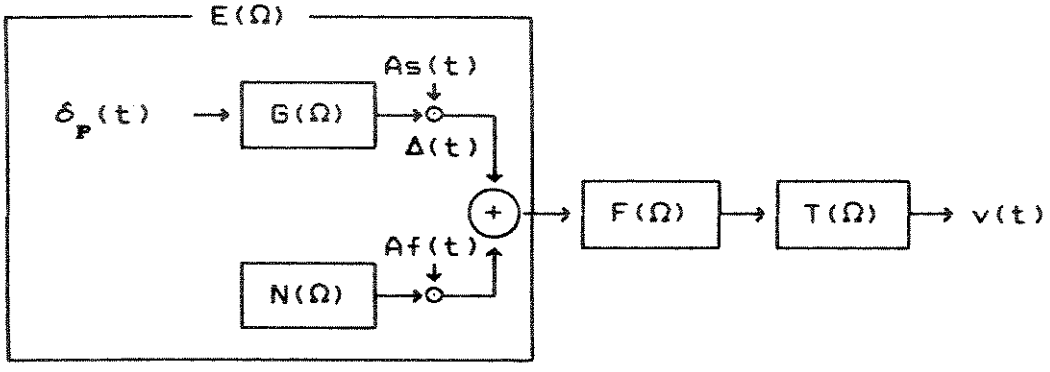


FIG. 3.10 - Modelo Linear para a produção da fala, onde P = período de pitch, Δ(t) = trem de pulsos glotais, As(t) = amplitude dos pulsos glotais, Af(t) = amplitude do ruído, v(t) é o sinal de voz, G(Ω) é um filtro conformador e N(Ω) é uma fonte de ruído branco.

cordas vocais e ao fluxo turbulento. Os fonemas fricativos sonoros são produzidos com a combinação das fontes de excitação sonora e turbulenta, cujas intensidades são controladas por $A_s(t)$ e $A_f(t)$, respectivamente. A explosão dos fonemas oclusivos não é incorporada explicitamente no modelo pois, com boa aproximação, ela pode ser representada por um ruído aleatório de curta duração.

O trem de pulsos glotais, $\Delta(t)$, é modelado pela convolução

$$\Delta(t) = \delta_p(t) * g(t) \qquad (3.3a)$$

onde

$$\delta_p(t) = \sum_{k=0}^{\infty} \delta(t-kP) \qquad (3.3b)$$

é uma seqüência periódica de impulsos, P é o período de pitch e g(t) é a resposta impulsiva de um filtro conformador, $G(\Omega)$, que responde pelo formato do pulso glotal. A forma do pulso glotal assemelha-se a uma onda dente de serra, com um tempo de subida maior que o tempo de descida. Para simplificar as análises, entretanto, será adotado um filtro conformador com uma resposta impulsiva triangular, conforme a figura 3.11, equivalente a uma resposta em freqüência cuja envoltória cai com $1/\Omega^2$ (ou 12 db/oitava):

$$g(t) = \begin{cases} 1 + t/\tau, & -\tau \le t \le 0 \\ 1 - t/\tau, & 0 \le t \le \tau \end{cases} \qquad (3.4a)$$

$$\xleftrightarrow{\mathcal{F}} \qquad G(\Omega) = \tau \cdot \left[\frac{\mathrm{sen}(\Omega\tau/2)}{(\Omega\tau/2)}\right]^2 \qquad (3.4b)$$

Deve ser ressaltado que a naturalidade da fala sintetizada está intimamente relacionada à precisão do modelamento dos pulsos glotais.

A excitação turbulenta, $N(\Omega)$, é simulada por uma fonte de ruído branco, com densidade de probabilidade uniforme, média nula, variância unitária e incorrelata com as outras variáveis do modelo.

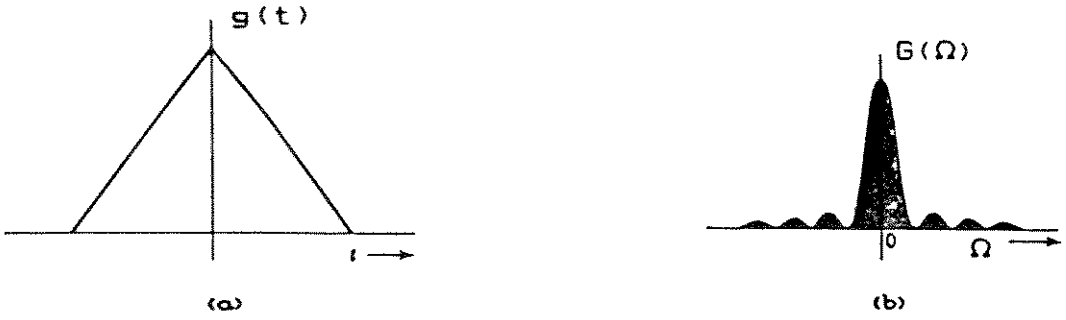


FIG.3.11 Filtro Conformador: a) resposta impulsiva; b) resposta em freqüência.

3.4.2 MODELAMENTO DO APARELHO FONADOR

De forma geral, a função de transferência do aparelho fonador, $F(\Omega)$, é dada por

$$F(\Omega) = \frac{N(\Omega)}{D(\Omega)} \qquad (3.5)$$

onde $D(\Omega)$ e $N(\Omega)$ são polinômios cujas raízes correspondem respectivamente aos pólos e zeros do aparelho fonador. O grau desses polinômios é determinado pelo som a ser modelado. Como exemplo, para fricativos surdos (figura 3.12), as cavidades anteriores à constricção são representadas por um pólo, assim como a própria constricção. As cavidades posteriores são representadas por um zero.

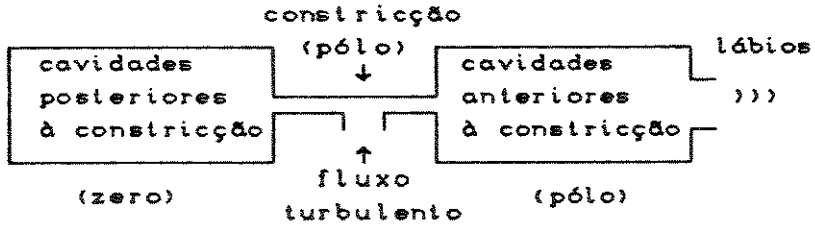


FIG 3.12 - Equivalente acústico para a produção de fricativos surdos. Em fonética articulatória o termo "anterior" refere-se à parte frontal da boca, próxima aos lábios e dentes; o termo "posterior", em oposição, está associado à parte da boca mais próxima da glote.

No modelamento de sons vocálicos nasalizados, $F(\Omega)$ apresenta pólos e zeros, enquanto, para vocálicos não nasalizados, $F(\Omega)$ apresenta apenas pólos, simplificando a análise.

O estudo do modelo contendo somente pólos é importante, pois conduz a resultados simples e eficazes para a análise de voz. Para a elaboração do modelo, o comprimento L do trato vocal é dividido em M tubos cilíndricos, concêntricos, e de igual comprimento, conforme a figura 3.13. Fazendo-se as seguintes considerações:

a) inexistência de perdas nas seções cilíndricas;

b) limitação da análise às freqüências abaixo de 4 kHz, ($\lambda \geq 8$ cm), de forma que os comprimentos de onda das freqüências de interesse sejam bem maiores que o raio médio do trato vocal (2 cm), garantindo a propagação do som apenas na direção axial do aparelho fonador,

as ondas planas de pressão e vazão, no k-ésimo tubo, no instante t, podem ser dadas por [3], [5]:

$$-\frac{\partial p}{\partial x} = \frac{\rho}{A}\frac{\partial u}{\partial t} \qquad (3.6a)$$

$$-\frac{\partial u}{\partial x} = \frac{A}{\rho c^2}\frac{\partial p}{\partial t} \qquad (3.6b)$$

onde:

x = distância, referenciada à extremidade esquerda de cada tubo, sendo $0 \leq x \leq L/M$;

$p = p_k(x,t)$ é a pressão no k-ésimo tubo no instante t, a uma distância x;

$u = u_k(x,t)$ é a vazão no k-ésimo tubo no instante t, a uma distância x;

$A = A_k(t)$ = área da seção transversal do k-ésimo tubo, no instante t;

ρ = densidade do ar no interior do trato vocal;

c = velocidade do ar no interior do trato vocal.

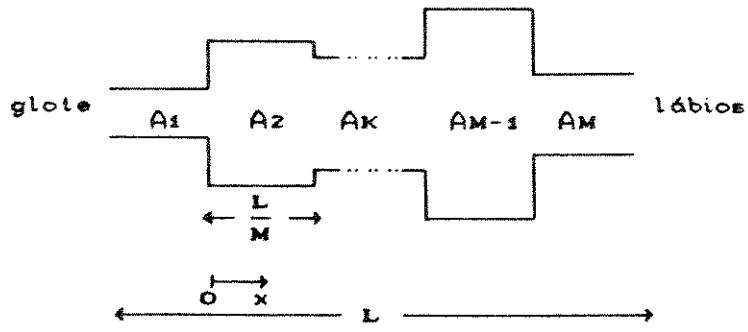


FIG.3.13 - Corte longitudinal de um modelo para o trato vocal contendo M tubos sem perdas.

A solução das equações 3.6 é dada por [3]:

$$u_k(x,t) = u_k^+(t - x/c) - u_k^-(t + x/c) \qquad (3.7a)$$

$$p_k(x,t) = \frac{\rho \cdot c}{A_k}\left[u_k^+(t - x/c) + u_k^-(t + x/c)\right] \qquad (3.7b)$$

onde $u_k^+(t - x/c)$ e $u_k^-(t + x/c)$ podem ser interpretadas como as ondas de vazão incidentes e refletidas, respectivamente, propagando-se dentro do k-ésimo tubo.

Aplicando-se, nas junções, as condições de continuidade da vazão e da pressão, isto é,

$$p_k(L/M,\ t) = p_{k+1}(0,\ t) \qquad (3.8a)$$

$$u_k(L/M,\ t) = u_{k+1}(0,\ t) \qquad (3.8b)$$

obtém-se [3], a partir de 3.7a,b e 3.8a,b, a expressão para a função de transferência do trato vocal, F(Ω), como:

$$F(\Omega) = \frac{\left[\prod_{k=1}^{M}(1+r_k)\right] \cdot e^{-j\Omega \cdot M \cdot \Delta t}}{D(\Omega)} \qquad (3.9a)$$

onde

$$F(\Omega) = U_M(\Omega)/U_1(\Omega) \qquad (3.9b)$$

é a relação entre a Tansformada de Fourier da vazão nos lábios, $U_M(\Omega)$, e a Transformada de Fourier da vazão na glote, $U_1(\Omega)$, e

$$r_k = \frac{A_{k+1} - A_k}{A_{k+1} + A_k} \qquad (3.9c)$$

é o coeficiente de reflexão entre os tubos k e k+1, k = 1, . . . , M. Foi suposto que o coeficiente de reflexão é unitário na glote, e nulo nos lábios, o que é equivalente à interpretação do sistema glotal como uma fonte de corrente ideal e do meio externo como um tubo de área infinita acoplado aos lábios e/ou narinas. Ainda a respeito da equação 3.9a,

$$\Delta t = L/(M \cdot c) \qquad (3.9d)$$

é o tempo necessário para o som percorrer o comprimento de cada seção e $D(\Omega)$ satisfaz a recursão polinomial

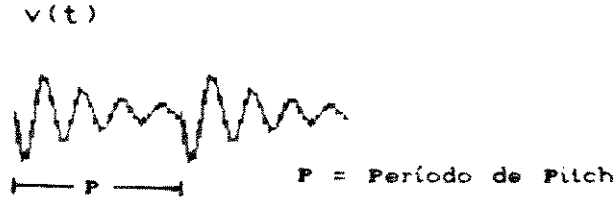
$$D_0(\Omega) = 1 \qquad (3.9e)$$

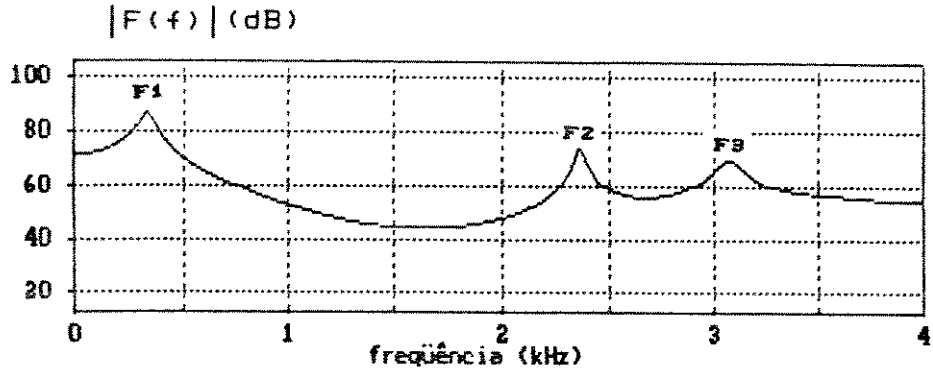
$$D_k(\Omega) = D_{k-1}(\Omega) + D_{k-1}(-\Omega) \cdot r_k \cdot e^{-j.2\Omega.\Delta t.k} \qquad (3.9f)$$

$$D_M(\Omega) = D(\Omega) \qquad (3.9g)$$

O inconveniente deste modelamento é a necessidade do conhecimento prévio da área de cada seção cilíndrica. Seu estudo, contudo, é importante por auxiliar a interpretação dos resultados das técnicas de Predição Linear utilizadas em nosso trabalho.

No estudo de vocálicos, os picos da resposta em freqüência do trato vocal (envoltória do espectro) recebem a denominação de formantes [13] sendo os três primeiros formantes, F1, F2 e F3, utilizados no reconhecimento das vogais. A figura 3.14 fornece uma justificativa para o termo "formante", onde uma vogal foi sintetizada ou "formada" utilizando-se três senóides amortecidas.





$$v(t) = \sum_{i=1}^{3} Ae^{-\sigma(i)t}\,\mathrm{sen}(2\pi f_i t) \qquad (\alpha \le t \le P)$$

$$\xleftrightarrow{\mathcal{F}}$$

$$|F(f)|_{(dB)} \cong 10 \cdot \log \prod_{i=1}^{3} \frac{f_i^2}{(f+f_i)\sqrt{(f-f_i)^2+(B_i/2)^2}}, \quad B_i = \sigma(i)/\pi, \quad B_i \ll f_i$$

FIG.3.14 - Forma de onda e espectro para uma vogal sintetizada com 3 formantes. As aproximações para a expressão da resposta em freqüência são sugeridas em [13].

3.4.3 MODELO DA IRRADIAÇÃO

O mecanismo de irradiação é modelado por uma impedância que transforma as ondas de vazão em ondas de pressão. Considerando-se as aberturas dos lábios e narinas desprezíveis em relação à superfície total da face, que é tratada então como um refletor plano de área infinita, obtém-se a "impedância de carga" das aberturas [2] por

$$T(\Omega) = \frac{j\Omega L R}{R + j\Omega L} \qquad (3.10a)$$

que é equivalente à ligação de uma resistência R em paralelo com uma indutância L, sendo

$$R = \frac{128}{9\pi^2} \qquad (3.10b)$$

e

$$L = \frac{8 \cdot a}{3\pi c} \qquad (3.10c)$$

onde "a" é o raio da circunferência cuja área é igual à área da abertura dos lábios ou narinas, e "c" é a velocidade do som no interior do trato vocal. A abertura pode ter um raio entre 0,5-1,5 cm, e c $\cong$ 35000 cm/s.

Para freqüências inferiores a 4 kHz, a equação 3.10a pode ser simplificada para

$$T(\Omega) = j\Omega L . \qquad (3.11)$$

Desta forma, os lábios (e/ou narinas) são responsáveis por uma inclinação de +6 db/oitava no espectro da voz. Para freqüências abaixo de 4 kHz (λ > 8 cm), a distância entre o nariz e a boca corresponde a uma pequena fração do comprimento de onda. O efeito da irradiação simultânea pode, então, ser aproximado pela superposição linear das ondas de pressão (ou vazão).

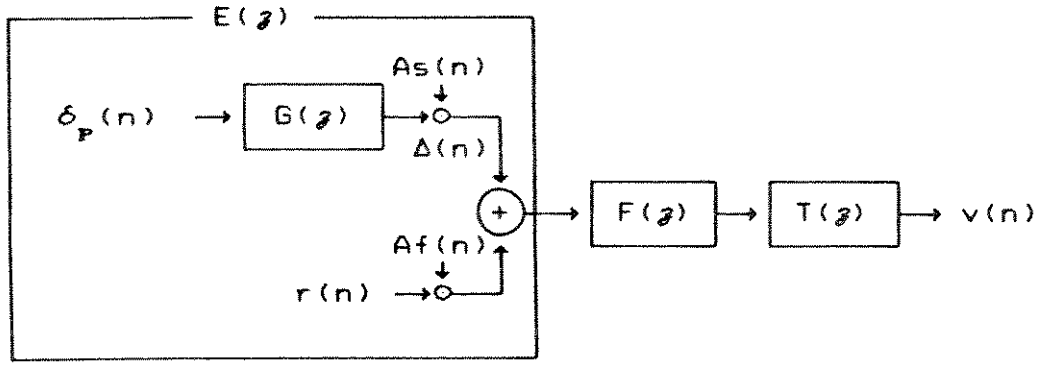


FIG. 3.15 - Modelo Digital para a Produção da Fala.

3.5 MODELO DIGITAL PARA A PRODUÇÃO DA FALA

A discussões anteriores tiveram por objetivo ressaltar os aspectos físicos envolvidos na produção da voz e desenvolver um modelo que servirá de base para um modelo discreto, apropriado para a análise prática em computadores digitais.

Nas discussões seguintes, será admitido que o espectro do sinal de voz, $V(\Omega)$, é limitado às freqüências abaixo de $\Omega_{máx} = 2\pi f_{máx}$ [rad/s] e amostrado na freqüência $f_s \geq 2 \cdot f_{máx}$ [Hz], de acordo com o Teorema da Amostragem. A freqüência angular do sinal discreto, $\omega = 2\pi f/f_s$ [rad], está normalizada em relação à freqüência de amostragem. Com estas considerações, a versão digital da figura 3.10 (Modelo da Produção da Voz) pode ser dada pela figura figura 3.15, onde $z = e^{j\omega}$. No restante da seção são discutidos os aspectos envolvidos nesta discretização.

3.5.1 EXCITAÇÃO

O trem de pulsos glotais, $\Delta(n)$, da figura 3.15 é modelado pela convolução

$$\Delta(n) = \delta_p(n) * g(n) \qquad (3.12a)$$

onde

$$\delta_P(n) = \sum_{k=0}^{\infty} \delta(n-kP) \qquad (3.12b)$$

é uma seqüência de impulsos, espaçados pelo intervalo de P instantes de amostragem. Sendo T o período de amostragem, o período de pitch é dado por P•T. A função g(n) corresponde à resposta impulsiva do filtro glotal (figura 3.16), dada por:

$$g(n) = \begin{cases} n+1 & 0 \le n < N \\ 2N-n-1, & N \le n < 2N \end{cases} \qquad (3.13a)$$

$$\longleftrightarrow \quad G(z) = \left[\frac{1 - z^{-N}}{1 - z^{-1}}\right]^2 \qquad (3.13b)$$

A equação 3.13b pode ser re-escrita da forma:

$$G(z) = \left[\frac{z^{N} - 1}{z^{N-1}\ (z - 1)}\right]^2 \qquad (3.14)$$

donde se vê que o filtro conformador possui 2(N-1) pólos na origem do plano z e (N-1) zeros duplos. Há um cancelamento de pólos e zeros para z = 1.

A amplitude da resposta em freqüência do filtro conformador é obtida fazendo-se $z = e^{j\omega}$ na equação 3.14,

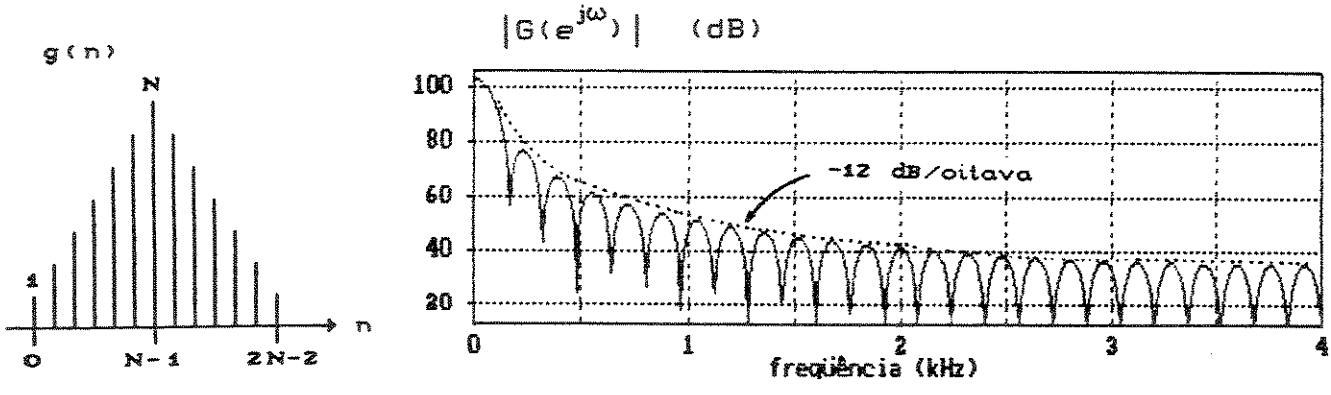


FIG. 3.16 - Filtro conformador: a) resposta impulsiva; b) resposta em freqüência, para N = 50 e fs = 8 kHz.

resultando em:

$$|G(e^{j\omega})| = \left[\frac{\text{sen}(N\omega/2)}{\text{sen}(\omega/2)}\right]^2 \qquad (3.15)$$

A excitação turbulenta é simulada por uma seqüência de números aleatórios, r(n), correspondente a um ruído branco, com densidade de probabilidade uniforme, média nula, variância unitária e incorrelato com as outras variáveis do modelo.

3.5.2 TRATO VOCAL

Na figura 3.13 (representação do trato vocal por M seções cilíndricas concêntricas, sem perdas), uma pertubação impulsiva na entrada do primeiro tubo resultará numa vazão de saída no m-ésimo tubo da forma:

$$u_M(t) = \alpha_o \cdot \delta(t - M\cdot\Delta t) + \sum_{k=1}^{\infty} \alpha_k \cdot \delta(t - M\cdot\Delta t - 2k\cdot\Delta t) \qquad (3.16a)$$

onde o primeiro impulso, de amplitude α_o, não sofre nenhuma reflexão, atingindo a saída em um tempo $M\cdot\Delta t$; os demais impulsos estão defasados de $2\Delta t$, uma vez que este é o tempo necessário para uma frente de onda percorrer cada seção duas vezes. A resposta em freqüência do trato vocal, $F(\Omega)$, é dada, portanto, pela Transformada de Fourier da equação 3.16a (resposta impulsiva do modelo) resultando em

$$F(\Omega) = e^{-j\Omega M\cdot\Delta t} \cdot \sum_{k=o}^{\infty} \alpha_k \cdot e^{-j\Omega k2\cdot\Delta t} \qquad (3.16a)$$

Ignorando-se o termo correspondente ao atraso, $e^{-j\Omega M\cdot\Delta t}$, na equação 3.16a, verifica-se que o termo associado às ressonâncias,

$$\tilde{F}(\Omega) = \sum_{k=o}^{\infty} \alpha_k \cdot e^{-j\Omega k2\cdot\Delta t} \qquad (3.16b)$$

apresenta a periodicidade

$$\tilde{F}(\Omega) = F\left(\Omega + \frac{2\pi l}{z \cdot \Delta t}\right) \qquad (3.16c)$$

onde l é um inteiro. Com base nesta última equação e explorando a periodicidade do espectro de um sinal digitalizado, pode-se obter uma representação discretizada do trato vocal, fazendo-se as seguintes suposições [3]:

a) a excitação do modelo de tubos concêntricos, sem perdas (figura 3.13), é limitada às freqüências abaixo de $\pi/2 \cdot \Delta t$ [rad/s];

b) o sinal de excitação é amostrado com um período $T = 2 \cdot \Delta t$;

c) o trato vocal é representando por um filtro digital dado por

$$u(n) = \begin{cases} \alpha_n & n \geq 0 \\ 0 & n < 0 \end{cases} \qquad (3.17)$$

onde $u(n)$ é obtido a partir de $u_M(t)$ (equação 3.16a), desconsiderando-se o atraso $M \cdot \Delta t$.

Desta forma, a versão digitalizada do trato vocal, $F(z)$, é obtida fazendo-se $z = e^{j \cdot 2\Omega \cdot \Delta t}$ nas equações 3.9, resultando em

$$F(z) = \frac{\left[\prod_{k=1}^{M} (1+r_k)\right] \cdot z^{-M/2}}{D(z)} \qquad (3.18a)$$

onde r_k é o k-ésimo coeficiente de reflexão, dado por 3.9c e $D(z)$ satisfaz a recursão

$$D_0(z) = 1 \qquad (3.18b)$$

$$D_k(z) = D_{k-1}(z) + D_{k-1}(z^{-1}) \cdot r_k \cdot z^{-k} \qquad (3.18c)$$

$$D_M(z) = D(z) \qquad (3.18d)$$

Note-se que, desconsiderando-se o atraso $z^{-M/2}$, o numerador da

equação 3.18a é equivalente a um número real, determinado unicamente pelos coeficientes de reflexão entre os tubos, este número pode ser interpretado como um fator de ganho. Quanto ao denominador, $D(z)$, vê-se que ele corresponde a um polinômio em z^{-1}, de grau M, cujas M raizes são os M pólos de $F(z)$. Portanto, para sons vocálicos não nasalizados, o trato vocal pode ser simulado por um modelo discreto de M pólos. Naturalmente, quanto maior o valor de M, melhor será a aproximação da geometria do trato vocal pelo modelo de tubos. Resta estabelecer a relação entre o número de pólos e a freqüência de amostragem.

Esta relação é obtida substituindo-se o valor de $\Delta t = L/Mc$ (equação 3.9c) em $T = 1/f_s = 2 \cdot \Delta t$, resultando em

$$f_s = \frac{Mc}{2L}, \qquad (3.19a)$$

Esta equação estabelece que o valor da freqüência de amostragem cresce linearmente com o número de seções cilíndricas do modelo. Caso a freqüência de amostragem seja pré-fixada, o número de seções pode ser estabelecido em função de f_s. Fazendo-se $L \cong 17$ cm e $c \cong 35000$ cm/s na equação 3.19, obtém-se

$$M \cong f_s/1000 \qquad (3.19b)$$

que impõe a necessidade de uma seção cilíndrica, de comprimento L/M, para cada quilohertz da freqüência de amostragem.

3.5.3 IRRADIAÇÃO

Como foi visto na seção 3.4.3, a influência dos lábios ou narinas no espectro da voz pode ser simplificada por um ganho fixo de 6 dB/oitava. O equivalente digital pode ser obtido aplicando-se a transformação bilinear [14]

$$j\Omega = \frac{2}{T}\left[\frac{1 - z^{-1}}{1 + z^{-1}}\right] \qquad (3.20)$$

na equação 3.10, resultando na expressão

$$T(z) = \frac{\left(\frac{2RL}{RT + 2L}\right)(1 - z^{-1})}{1 + \left(\frac{RT - 2L}{RT + 2L}\right) z^{-1}} \qquad (3.21a)$$

que apresenta um zero em $z = 1$ e um pólo real, que é função de L e, portanto, da abertura labial (equação 3.10c). Para $c = 35000$ cm/s, $T = 125\ \mu s$ ($f_s = 8$ kHz) e $0,5\ cm \leq a \leq 1,5\ cm$, o pólo varia entre $z = -0,76$ e $z = -0,42$. Uma simples aproximação para a equação 3.21a, sugerida em [3], é obtida desprezando-se os efeitos do pólo, ou seja,

$$T(z) \cong \left(\frac{2RL}{RT + 2L}\right)(1 - z^{-1}) \qquad (3.21b)$$

onde $\left(\frac{2RL}{RT + 2L}\right)$ pode oscilar entre 0,17 e 0,41, para $0,5\ cm \leq a \leq 1,5\ cm$ e $T = 125\ \mu s$ (ou $f_s = 8$ kHz).

3.5.4 MODELO DIGITAL SIMPLIFICADO PARA A PRODUÇÃO DA FALA

Na seção 3.5.2 ficou demonstrado que na produção de sons vocálicos não nasalizados, o trato vocal pode ser modelado por um filtro contendo apenas pólos, cuja ordem, M, é diretamente proporcional à freqüência de amostragem. Este resultado será aproveitado nesta seção, que tem por objetivos:

a) simplificar o modelo digital de produção da fala da figura 3.15,

b) utilizar o modelo contendo apenas pólos, de forma aproximada, na produção de vocálicos nasalizados e de fricativos.

O modelo simplificado, apresentado na figura 3.17, é obtido a partir da figura 3.15. A seqüência de voz, v(n), é a

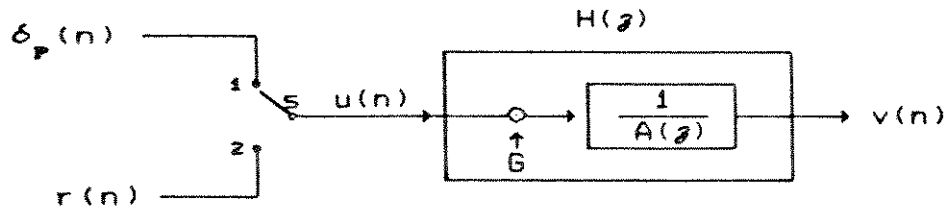


FIG. 3.17 - Modelo Digital Simplificado para a Produção da Fala. A chave "s" seleciona ou a excitação periódica (posição 1), ou a excitação turbulenta (posição 2). O filtro $H(z)$ possui apenas pólos. G determina a amplitude do sinal de voz e $A(z)$ é um polinômio de grau M.

resposta do filtro $H(z)$ à excitação u(n), cuja transformada $\mathcal{Z}$ é $U(z)$. Formalmente,

$$V(z) = U(z) \cdot H(z) = U(z) \frac{G}{1 - \sum_{i=1}^{M} a_i \cdot z^{-i}} \qquad (3.22)$$

onde G é um fator de ganho. Note-se as semelhanças entre as equações 3.22 e 3.18a. Na produção de *sons sonoros*, a chave "s" do modelo simplificado é colocada na posição 1 e u(n) é dada pela seqüência de impulos $\delta_p(n)$ da equação 3.12b. Na produção de *fricativos surdos*, "s" é colocada na posição 2 e u(n) corresponde à seqüência aleatória r(n). Observe-se que o modelo simplificado não prevê a produção de fricativos sonoros.

No modelo simplificado, $H(z)$ deve incorporar os efeitos do filtro conformador do pulso glotal, $G(z)$ (figura 3.16), e da irradiação (equação 3.21b). Entretando, o modelo não possui zeros que respondam pelos zeros do filtro conformador, $G(z)$, e/ou do aparelho fonador, na produção de sons fricativos, consoantes nasais ou vogais nasalizadas. Porém, se o número de pólos for suficientemente elevado, $H(z)$ pode simular razoavelmente o efeito destes zeros. Isto é justificado, lembrando-se que um zero da forma $1 - \beta z^{-1}$, pode ser expresso por

$$1 - \beta z^{-1} = \frac{1}{1 + \sum_{i=1}^{\infty} (\beta z^{-1})^i} \cong \frac{1}{1 + \sum_{i=1}^{Q} (\beta z^{-1})^i} \qquad (3\ 23)$$

onde Q é o número de pólos da aproximação. Desta forma, a ordem M do filtro $H(z)$ deverá ser superior a $f_s/1000$, (equação 3.19b) para que os pólos excedentes respondam pelos efeitos dos zeros do pulso glotal, da irradiação e, eventualmente, do próprio trato vocal.

Até o momento, este capítulo tem estudado em detalhes vários aspectos da *produção* da fala. Nesta última seção mostrou-se que o sinal de voz pode ser razoavelmente modelado pela resposta de um filtro $H(z)$ composto por apenas M pólos. A excitação deste filtro é um trem de impulsos, na produção de sons sonoros, ou uma seqüência de número aleatórios, na produção dos sons fricativos surdos. Dito de outra forma, o sinal de voz pode ser considerado um processo autoregressivo de ordem M. Este resultado, juntamente com as hipóteses de que o sinal de voz é ergódico[1] e portanto estacionário no sentido amplo,[2] permitem que as técnicas de Predição Linear sejam utilizadas no seu estudo. A validade das duas hipóteses levantadas será discutida na próxima seção. O grande mérito da Predição Linear, quando aplicada à *análise* do sinal de voz, reside na possibilidade de estimar os parâmetros do filtro $H(z)$ de forma simples e precisa, *a partir do próprio sinal*.

3.6 PREDIÇÃO LINEAR DO SINAL DE VOZ

A formulação da Predição Linear foi realizada por

---

[1] Um processo estocástico é ergódico quando suas médias estatísticas são iguais suas médias temporais.

[2] Um processo estocástico estacionário no sentido amplo possui uma média constante e uma função de autocorrelação que depende apenas da diferença entre os intervalos de correlação.

diversos métodos, todos levando a resultados semelhantes, como o Método da Autocorrelação, o Método da Covariância, da Filtragem Inversa ou o Método da Máxima Verossimilhança Um estudo dos vários métodos e a comparação entre eles pode ser encontrado em [12]. No nosso trabalho será utilizado o Método da Autocorrelação, que será discutido em seguida

Partindo-se do princípio que a voz é produzida pelo modelo simplificado da figura 3.17, o problema a ser solucionado pode ser resumido na seguinte questão: como obter os parâmetros do filtro $H(z)$, isto é, o ganho G e os coeficientes do polinômio $A(z)$, a partir do próprio sinal de voz?

Para formular o problema, numa análise a curto prazo, inicialmente é selecionado um segmento do sinal de voz através de uma janela de comprimento finito e igual a N (figura 3.18). A escolha adequada do valor de N garante uma boa aproximação às hipóteses de ergodicidade e estacionariedade no sentido amplo, levantadas anteriormente. Devido à inércia dos articuladores, é intuitivo que o sinal de voz possa ser considerado estacionário em intervalos apropriados, de curta duração.

Nas equações seguintes, $\{v_n\}$ corresponderá ao segmento selecionado e ponderado pela janela sendo, portanto, nulo para $n < 0$ e $n > N$. A origem do eixo "n" será estabelecida no início de cada segmento, para simplificar as notações.

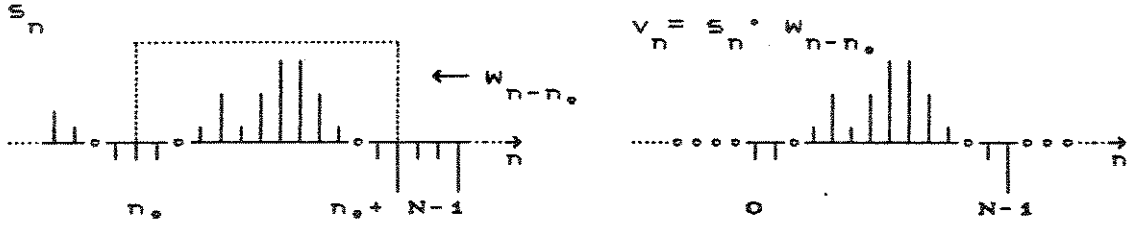


FIG.3.18 - Exemplo de um segmento de voz selecionado a partir da seqüência $\langle s_n \rangle$ através de uma janela retangular. Para simplificar as notações, a origem do eixo n é redefinida a cada segmento selecionado.

Em seguida, a equação 3.22 será escrita no domínio do tempo como

$$v_n = G \cdot u_n + \sum_{i=1}^{M} \alpha_i \cdot v_{n-i} \qquad (3.24)$$

onde (v_n), $0 \leq n < N$, é o segmento do sinal de voz. Esta equação justifica a denominação "modelo auto-regressivo" que normalmente é dada ao modelo simplificado da figura 3.17.

A idéia fundamental da Predição Linear consiste em aproximar cada amostra do sinal de voz pela combinação linear de amostras passadas do sinal. Sendo M o número de amostras passadas utilizadas na combinação linear, pode-se formalizar a aproximação da amostra genérica v_n pela relação

$$\tilde{v}_n = \sum_{i=1}^{M} \alpha_i \cdot v_{n-i} \qquad (3.25)$$

onde $\tilde{v}_n$ é a aproximação de v_n e α_i é o i-ésimo coeficiente da combinação linear; $\tilde{v}_n$ é normalmente denominada a *estimativa* ou *predição* de ordem M da amostra v_n.

O erro de predição de cada amostra, e_n, é definido por

$$e_n = v_n - \tilde{v}_n = v_n - \sum_{i=1}^{M} \alpha_i \cdot v_{n-i} \qquad (3.26)$$

e o erro quadrático, E_n, acumulado em todo o segmento por

$$E_n = \sum_{n=-\infty}^{\infty} e_n^2 \qquad (3.27)$$

Como o segmento de voz é nulo para $n < 0$ e para $n \geq N$, o erro de predição (equação 3.27) é conseqüentemente nulo para $n < 0$ e $n > N+M-1$. Com esta consideração, e substituindo-se a equação 3.26 na equação 3.27, obtém-se:

$$E_n = \sum_{n=0}^{N+M-1} \left(v_n - \sum_{i=1}^{M} \alpha_i \cdot v_{n-i} \right)^2 \qquad (3.28)$$

O conjunto de coeficientes α_i que minimiza E_n é obtido a

partir de

$$\frac{\partial [E_n]}{\partial [\alpha_i]} = 0 \qquad 1 \leq i \leq M \tag{3.29}$$

Com a substituição da equação 3.28 em 3.29 e a realização das M derivadas parciais chega-se ao seguinte sistema de equações lineares:

$$\sum_{k=1}^{M} \alpha_k \cdot R(|i-k|) = R(i) \qquad 1 \leq i \leq M \tag{3.30a}$$

onde

$$R(k) = \sum_{n=0}^{N-k-1} v_n \cdot v_{n+k} \tag{3.30b}$$

é a função de autocorrelação a curto prazo. As equações 3.30, conhecidas por Equações de Yule-Walker, podem ser visualizadas mais facilmente se colocadas na forma matricial:

$$\begin{bmatrix} R(0) & R(1) & R(2) & \cdots & R(M-1) \\ R(1) & R(0) & R(1) & \cdots & R(M-2) \\ R(2) & R(1) & R(0) & \cdots & R(M-3) \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ R(M-1) & R(M-2) & R(M-3) & \cdots & R(0) \end{bmatrix} \cdot \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \cdots \\ \alpha_M \end{bmatrix} = \begin{bmatrix} R(1) \\ R(2) \\ R(3) \\ \cdots \\ R(M) \end{bmatrix} \tag{3.31}$$

Os coeficientes α_i do preditor são determinados através da solução das equações 3.30 (ou 3.31) e utilizados como estimativas dos coeficientes a_i do filtro $H(z)$ da figura 3.17.

Resumindo, os coeficientes de $H(z)$ para um segmento de voz de comprimento N são estimados da seguinte forma:

a) cálculo das autocorrelações a curto prazo através da equação 3.30b;

b) solução do sistema de equações 3.31;

A exploração das simetrias da matriz de autocorrelações permite a elaboração de algoritmos recursivos muito eficientes para solução do sistema. No capítulo 4 será utilizado um algoritmo clássico, conhecido por Algoritmo de Levinson-Durbin [3], [12].

Uma vez estimados os coeficientes do polinômio $A(z)$, resta determinar o ganho, G, que pode ser expresso por [3]

$$G = \left[R(0) - \sum_{k=1}^{M} a_k \cdot R(k) \right]^{1/2} \qquad (3.32)$$

onde R(i) é a função de autocorrelação calculada com atraso i. Esta relação é válida tanto para a excitação periódica quanto para a excitação turbulenta do modelo.

3.7 SISTEMA AUDITIVO E PERCEPÇÃO DA FALA

O objetivo desta seção é a apresentação de alguns aspectos que ocorrem com o ouvinte, no elo final da Cadeia da Fala, e como eles podem ser utilizados em Sistemas de Reconhecimento de Voz.

O processo de decodificação será dividido em *audição* e *percepção*. A audição está relacionada com processos físicos bem conhecidos, responsáveis pela transformação das ondas sonoras em disparos neurais. A percepção diz respeito às formas utilizadas pelo cérebro para a associação dos impulsos neurais originários do sistema auditivo, e é estudada com a realização de testes psicoacústicos. Nestes testes são avaliados os efeitos perceptuais de estímulos sonoros aplicados ao ouvido humano.

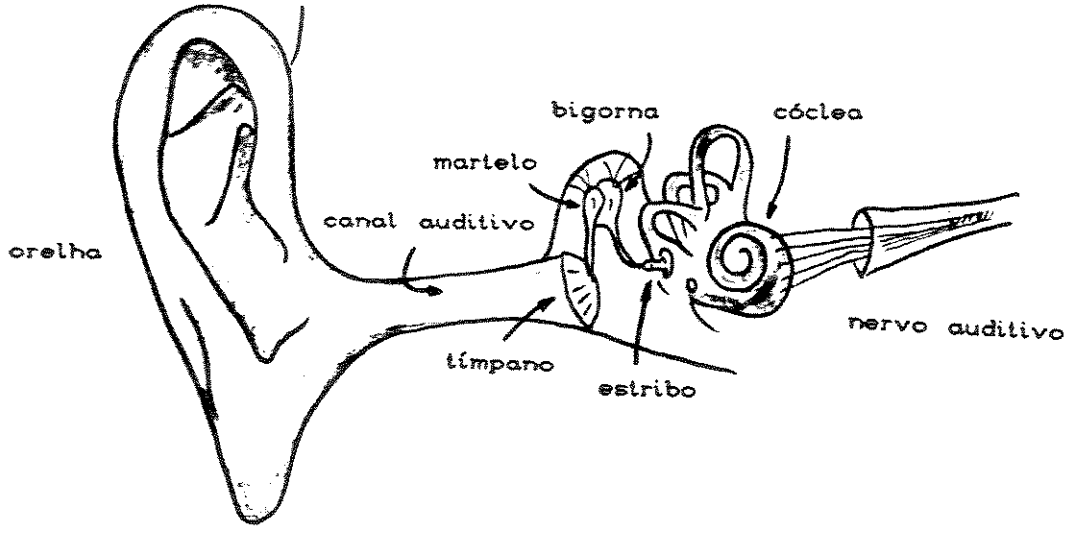


FIG. 3.19 - Sistema Auditivo Periférico

3.7.1 AUDIÇÃO

Conforme a figura 3.19, o sinal acústico é conduzido pelas orelhas para o interior do *canal auditivo*. Sendo aberto em uma extremidade e fechado na outra pelo *tímpano*, o canal auditivo (com ℓ = 27 mm de comprimento e d = 7 mm de diâmetro) comporta-se como um ressoador de quarto de onda, com o primeiro modo de ressonância em:

$$f = \frac{c}{4 \cdot \ell} = \frac{35000}{4 \times 2,7} \cong 3,2 \text{ kHz} \qquad (3.33)$$

Medições revelam um aumento de 5 a 10 dB da pressão sonora no tímpano, em relação à pressão na entrada do canal, na faixa de 3-5 kHz [2]. Nesta região concentra-se a maior parte da energia dos fricativos surdos.

As vibrações do ar no interior do canal provocam o movimento do tímpano. Entre o tímpano e a *cóclea*, duto espiralado que contém o líquido coclear, estão localizados os três ossículos

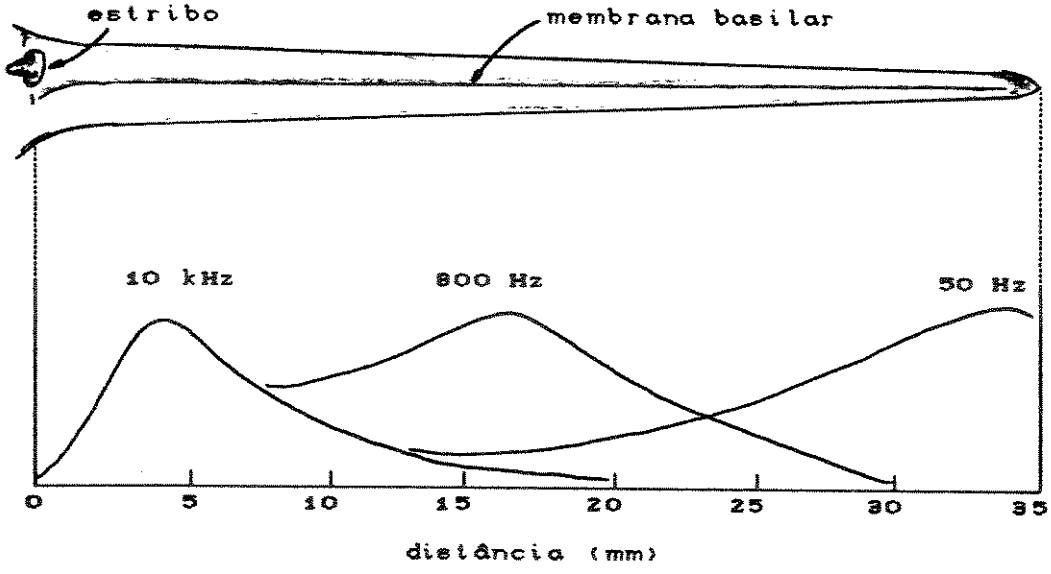


FIG 3.20 - Cóclea "desenrolada" e resposta em freqüência da membrana basilar ao longo de seu comprimento. Ref: "Acústica", L. X. Nepomuceno - Editora Blücher - 1977.

(martelo, bigorna e estribo). Os ossículos atuam como um sistema de alavancas efetuando o acoplamento entre o ar e o líquido coclear, que é posto em movimento em resposta às vibrações do estribo.

No interior da cóclea, ao longo de seus 35 mm de comprimento, ocorre a transformação do movimento do líquido coclear em impulsos nervosos. Nesta transformação destaca-se a *membrana basilar* que sofre deformações em função do movimento do líquido. Cada região da membrana basilar (figura 3.20) comporta-se como um filtro que responde a determinada faixa de freqüências. Estes filtros apresentam um fator de mérito (isto é, a relação entre a freqüência central e a largura de faixa) aproximadamente constante. Portanto, a resolução é melhor nas baixas freqüências.

Terminações nervosas especializadas, acopladas à membrana basilar, detectam seus movimentos, transformando-os em disparos elétricos que são conduzidos ao cérebro.

3.7.2 PERCEPÇÃO DA FALA

A resolução em freqüência do ouvido permite a distinção de uma variação de até 3 Hz na percepção de dois tons não simultâneos localizados nas vizinhanças de 1 kHz. Entretanto, na percepção de tons simultâneos próximos em freqüência, temos a sensação de ouvir apenas o tom de maior intensidade, a menos que a diferença em freqüência entre os tons ultrapasse uma certa distância, denominada *banda crítica* (zc), que é função da localização dos tons na escala de freqüência. Testes psicoacústicos revelam a existência de cerca de 24 bandas críticas na faixa audível (20 Hz a 20 kHz). A correspondência entre freqüência e banda crítica (expressa em "barks") pode ser obtida pela figura 3.21 ou pela tabela 3.1 [8]. A relação é aproximadamente linear até cerca de 500 Hz, e aproximadamente logarítmica, a partir daí.

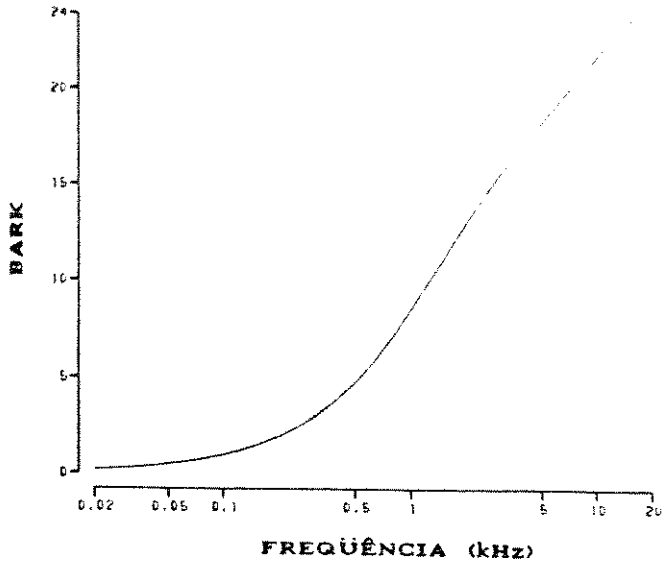


FIG.3.21 - Relação entre freqüência (em kHz) e bandas críticas (em barks) [8].

entre freqüência e banda crítica (expressa em "barks") pode ser obtida pela figura 3.21 ou pela tabela 3.1 [8]. A relação é aproximadamente linear até cerca de 500 Hz, e aproximadamente logarítmica, a partir daí.

O uso de bandas críticas foi introduzido recentemente no reconhecimento de voz (em vogais, particularmente, como será visto no capítulo 5). Para seu uso prático foram propostas várias fórmulas [8], [9], destacando-se as expressões abaixo, que são aproximações às curvas da figura 3.21:

$$z_c = 13 \cdot \tan^{-1}(0{,}76 \cdot f) + 3{,}5 \cdot \tan^{-1}(f/7{,}5)^2 \qquad (3.34)$$

$(20\ \text{Hz} < f < 20\ \text{kHz})$

onde z_c é a banda crítica (em barks), f é a freqüência (em kHz) e o argumento de $\tan^{-1}$ é dado em radianos.

Uma relação mais simples, para freqüências na faixa de 200 Hz a 6800 Hz, consiste em:

$$z_c = \frac{26{,}81}{1 + \frac{1{,}96}{f}} - 0{,}53 \qquad (3.35)$$

sendo a freqüência em kHz e a banda crítica em barks.

Zc	fcen Hz	fcor Hz	bw Hz
		20	
1	50	100	80
2	150	200	100
3	250	300	100
4	350	400	100
5	450	510	110
6	570	630	120
7	700	770	140
8	840	920	150
9	1000	1080	160
10	1170	1270	190
11	1370	1480	210
12	1600	1720	240
13	1850	2000	280
14	2150	2320	320
15	2500	2700	380
16	2900	3150	450
17	3400	3700	550
18	4000	4400	700
19	4800	5300	900
20	5800	6400	1100
21	7000	7700	1300
22	8500	9500	1800
23	10 500	12 000	2500
24	13 500	15 500	3500

TAB.3.1 - Divisão da faixa audível em bandas críticas [7]. Zc é a banda crítica (em bark), fcen é a freqüência central da banda (em Hz), fcor são as freqüências de corte (em Hz), e bw a largura de faixa (em Hz) [8].

A largura da banda crítica pode ser calculada por:

$$CB = 25 + 75 \cdot (1 + 1,4 \cdot f^2)^{0,69} \qquad (3.36)$$

onde CB é a largura da banda crítica (em Hz) e f é a freqüência (em kHz).

Finalizando, serão citados dois fatos da percepção da fala bastante conhecidos e utilizados pela engenharia:

a) a "Lei da Fase", postulando que a percepção de um som depende somente do seu espectro de potência, sendo independente dos ângulos de fase das suas componentes; a título de curiosidade, esta lei foi descoberta pelo mesmo G. S. Ohm [6], autor da "Lei de Ohm" para circuitos elétricos;

b) as duas formas de percepção do pitch: pela existência da energia da freqüência fundamental no espectro, ou através do espaçamento entre os harmônicos (que é igual à freqüência de excitação); é por esse último mecanismo que a intonação da fala é percebida nas

faixa em 300 Hz-3400 Hz, elimina praticamente toda a componente fundamental da freqüência de excitação glótica.

3.8 DISCUSSÃO

A compreensão dos fenômenos físicos da produção da fala é fundamental para a busca de pistas no sinal acústico, associadas aos diferentes sons da fala. Em reconhecimento de voz, igual importância deve ser dada aos processos de produção, audição e percepção da fala [16]. Modelos para o sistema auditivo periférico precisam ser aprimorados, para que se tenha uma representação da voz apropriada para o uso em SRAV. Trabalhos recentes [15], [17], [18], [19], revelam que o uso de técnicas baseadas no comportamento do sistema auditivo, como a transformação em bandas críticas, resultam em melhorias significativas no desempenho dos Sistemas de Reconhecimento de Voz.

3.9 REFERÊNCIAS

[1] P. B. Denes & E. N. Pinson, "*The Speech Chain: The Physics and Biology of Spoken Language*", Anchor Books, N.Y. (1963);

[2] J. Flanagam, "*Speech Analysis Synthesis and Perception*", Academic Press, N.Y. (1965);

[3] L. R. Rabiner & R. W. Schafer, "*Digital Processing of Speech Signals*", Prentice Hall, N.J. (1978);

[4] S. Singh & K. S. Singh, "*Phonetics: Principles and "Practices*", University Park Press, Londres (1976);

[5] D. O'Shaughnessy, "*Speech Communication: Human and Machine*", Addison-Wesley, N.Y. (1987);

[6] M. R. Schroeder, "*Models of Hearing*", Proceedings of the IEEE, vol 63 (9), pp 1332-1350 (1975);

[7] E. Zwicker, *"Subdivision of the Audible Frequency Range into Critical Bands (Frequenzgruppen)"*, The Journal of The Acoustical Society of America, vol 33 (2), p 248 (1961);

[8] E. Zwicker & E. Terhardt, *"Analytical Expressions for Critical-Band Rate and Critical Bandwidth as a Function of Frequency"*, The Journal of the Acoustical Society of America, vol 68 (5), pp 1523-1525 (1980);

[9] H. Traunmüller, *"Paralinguistic Variation and Invariance in the Caracteristic Frequency of Vowels"*, Phonetica, 45, pp 1-29 (1988).

[10] G. Fant, *"Acoustic Theory of Speech Production"*, Mouton, The Hague, Paris (segunda impressão) (1970);

[11] G. Fant, *"Speech Sounds and Features"*, The MIT Press, Massachusetts, (1973);

[12] J. Markel & A. H. Gray Jr., *"Linear Prediction of Speech"*, Springer Verlag, N.Y., (1975);

[13] G. Fant, *"On The Predictability Of Formant Levels and Spectrum Levels from Formant Frequencies"*, em I. Lehist, ed., *"Readings in Acoustic Phonetics"*, pp 44-56, The MIT Press, Massachusetts, (1963);

[14] A. V. Oppenheim & R. W. Schaffer, *"Digital Signal Processing"*, Prentice Hall Inc., Englewood Cliffs, N. J., (1975).

[15] S. Seneff, *"A Computational Model For the Peripheral Auditory System: Application do Speech Recognition Research"*, Proceedings of the International Conference on Acoustic, Speech and Signal Processing, pp 1983-1986, Tóquio, Japão (1986);

[16] K. Stevens, *"The Quantal Nature of Speech: Evidence From Articulatory-Acoustic Data"*, em E. E. David Jr. & P. D. Denes, eds., *"Human Communication: A Unified View"*, McGraw Hill, Inc. (1972);

[17] S. Furui, *"Speaker-Independent Isolated Word Recognition Based on Emphasized Spectral Dynamics"*, Proceedings of the International Conference on Acoustic, Speech and Signal Processing, pp 1991-1994, Tóquio, Japão (1986);

[18] H. Kuwabara, *"Representation of Vowels in Connected Speech Based on Speech Perception Experiments"*, Proceedings of the International Conference on Acoustic, Speech and Signal Processing, pp 19873-1990, Tóquio, Japão (1986);

[19] A. K. Syrdal & H. S. Gopal, *"A Perceptual Model of Vowel Recognition Based on The Auditory Representation of American English Vowels"*, The Journal of The Acoustical Society of America, vol 79 (4) pp 1086-1100 (1986).

CAPÍTULO 4

O MÓDULO FRONTAL

4.1 INTRODUÇÃO

Um Sistema de Reconhecimento Automático de Voz baseado em conhecimentos lingüísticos pode ser dividido em dois blocos principais, sob o ponto de vista do nível do processamento envolvido:

a) o primeiro bloco, de acordo com a figura 4.1, é normalmente denominado Módulo Frontal, ou Módulo Fonético-Acústico, sendo responsável pela análise da forma de onda, e/ou de alguma transformação equivalente do sinal, para a obtenção de uma seqüência de parâmetros característicos do sinal de voz. O processamento realizado no Módulo Frontal corresponde à análise de baixo nível do sistema;

b) o segundo bloco corresponde ao Módulo Lingüístico. Este módulo, encarregado do processamento de alto nível do sistema, tem a função de determinar a

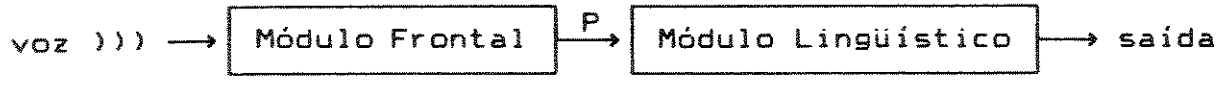


FIG. 4.1 - Sistema de Reconhecimento de Voz. A saída do Módulo Frontal, P, é um conjunto de parâmetros e/ou traços obtidos pela análise a curto prazo do sinal de voz.

provável elocução de entrada, a partir da análise da seqüência de informações que lhe são enviadas pelo Módulo Frontal. O processo de decisão que se desenvolve no Módulo Linguístico requer o conhecimento de aspectos específicos da língua, do vocabulário e, eventualmente, do próprio locutor.

Este capítulo trata especificamente do desenvolvimento e avaliação do software para o Módulo Frontal de um Sistema de Reconhecimento Automático de Voz, operando na faixa de 0-4 kHz. A saída P do nosso Módulo Frontal é composta pela classificação da elocução de entrada segundo categorias fonéticas (silêncio, pausa sonora, coarticulação, fricativo surdo, fricativo sonoro, vocálico ou indefinido) e, adicionalmente, dos valores da freqüência fundamental de excitação, f0 e dos três primeiros formantes, F1, F2, F3. O trabalho, iniciado em agosto de 1987, assemelha-se ao Módulo Frontal de diversos sistemas descritos na literatura [10-13], [26-30]. A grande diferença, contudo, é que a maioria desses sistemas trabalha com um sinal de faixa mais ampla, tipicamente de 0-8 kHz, o que torna a análise significativamente mais simples. Diante da proposta de desenvolver um sistema que operasse na faixa de 0-4 kHz, independente do locutor, foi necessário analisar a fala de inúmeras pessoas, incluindo homens e mulheres, com o intuito de:

a) escolher um conjunto mais adequado de parâmetros;
b) representar cada categoria fonética em função dos valores assumidos por esses parâmetros.

Os conceitos básicos de fonética acústica e de produção da fala que orientaram a escolha e cálculo dos parâmetros foram aqueles discutidos no capítulo 3. O trabalho foi realizado com os recursos disponíveis no Laboratório de Comunicações Digitais do Departamento de Comunicações (DECOM) da Faculdade de Engenharia Elétrica da UNICAMP.

4.2 BLOCOS FUNCIONAIS

O Módulo Frontal, acoplado à estação de trabalho SAPDV-A [1], está representado esquematicamente na figura 4.2. Conforme o diagrama em blocos desta figura, o sinal de voz é captado por um microfone, digitalizado e transferido à memória do microcomputador. Toda esta operação é realizada pelo SAPDV-A. Em seguida, o arquivo de voz é lido em quadros de pequena duração e processado, *em tempo não real*, pelos algoritmos desenvolvidos neste trabalho. Todo o software foi desenvolvido no ambiente de programação Turbo C versão 2.0 (Borland International Inc.), instalado em um microcomputador compatível com o IBM PC AT, equipado com coprocessador numérico. Não houve a preocupação em otimizar os programas quanto a tempo de execução ou quantidade de memória utilizada.

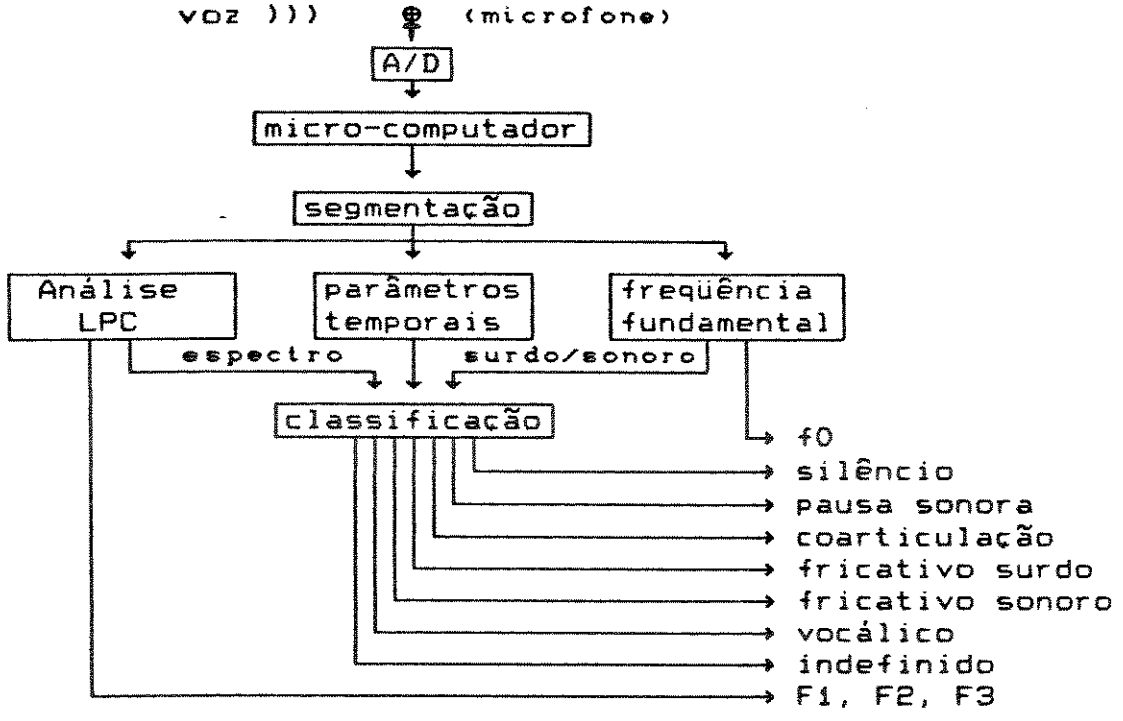


FIG. 4.2 - Módulo Fonético-Acústico. Após a digitalização, são realizadas três linhas de processamento: análise LPC, estimação de parâmetros temporais e da freqüência fundamental.

4.2.1 DIGITALIZAÇÃO DO SINAL DE VOZ

A digitalização (bloco A/D da figura 4.2), como já foi dito, foi realizada no Sistema de Análise e Processamento Digital de Voz, SAPDV-A. Neste equipamento, o sinal analógico é inicialmente limitado à faixa de 0-3400 Hz por um filtro passa-baixas elíptico, com uma atenuação menor que 0,1 dB abaixo de 3,4 kHz, e maior que 34 dB acima de 4,6 kHz. Em seguida, o sinal é amostrado a uma taxa constante de 8 kHz e quantizado linearmente em 12 bits (2^{12} = 4096 níveis). As amostras do sinal de voz são transferidas ao disco rígido do microcomputador através de uma interface GPIB (General Purpose Interface Bus IEE 488). A duração do enunciado a ser digitalizado é limitada apenas pela memória disponível no disco rígido. O SAPDV-A também dispõe de facilidades para a monitoração dos sinais em osciloscópio, assim como recursos para a reprodução sonora dos arquivos digitalizados.

As gravações foram feitas no próprio Laboratório de Comunicações Digitais, que não dispõe de nenhuma forma de isolamento acústico; entretanto, elas foram realizadas na ausência de ruídos provocados por conversas de fundo, operação de equipamentos de ar condicionado ou motores de automóveis, por exemplo. O microfone utilizado (figura 4.3) apresenta uma alta

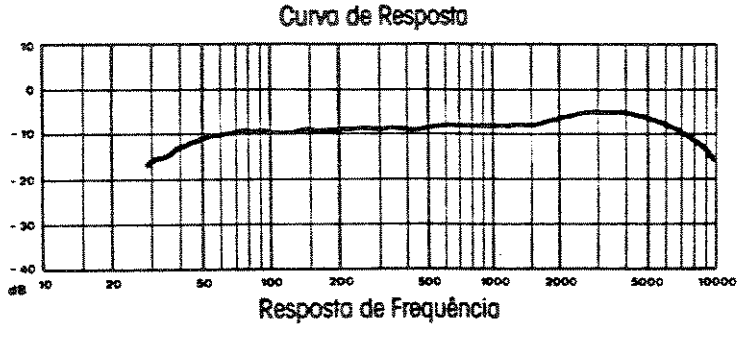


FIG. 4.3 - Resposta em freqüência do microfone de eletreto unidirecional utilizado no SAPDV-A. Ref.: catálogo do fabricante (Le-Son).

diretividade, contribuindo para a eliminação de ruídos estranhos à fala. Não foram constatados problemas devido a harmônicos da freqüência de 60 Hz.

4.2.2 SEGMENTAÇÃO PARA ANÁLISE A CURTO PRAZO

Os dados são lidos seqüencialmente da memória do microcomputador em quadros de 200 amostras (25 ms), com um passo de 40 amostras (5 ms) entre dois quadros, conforme a figura 4.4. Esta forma de segmentar o sinal de voz, conhecida por *segmentação não sincronizada com o pitch*, requer que comprimento do quadro seja uma solução de compromisso entre dois fatores fundamentais:

a) o comprimento deve ser pequeno, para evitar a possibilidade de ocorrer movimentos significativos dos órgãos articuladores durante o intervalo analisado. O movimento articulatório pode ser considerado estacionário em intervalos menores que cerca 30 ms (250 amostras);

b) para a estimação da freqüência fundamental de excitação, f0, o quadro deve incluir pelo menos dois

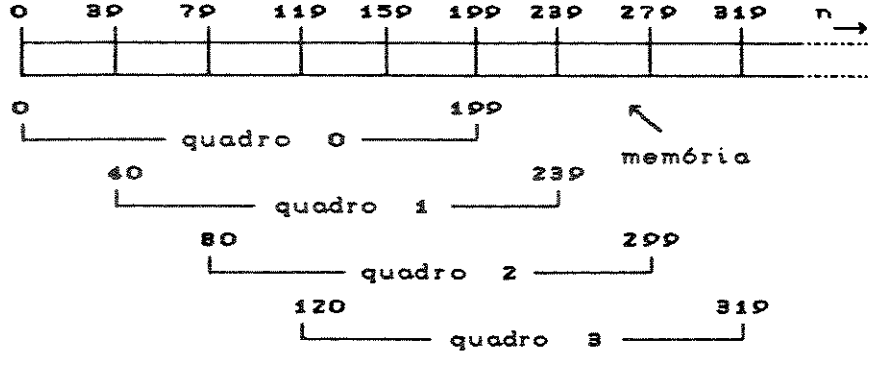


FIG. 4.4 - Divisão da seqüência de amostras do sinal de voz em quadros para a análise a curto prazo. Cada quadro possui N = 200 amostras (25 ms) e sofre um deslocamento de 40 amostras (5 ms) em relação ao quadro anterior.

períodos de pitch. A escolha de 25 ms para o comprimento do quadro possibilita, portanto, a estimação de valores de f0 tão baixos quanto 80 Hz.

O deslocamento dos quadros em intervalos de 5 ms permite a detecção de alguns movimentos mais rápidos, como a transição de formantes após a explosão de oclusivas, além de um bom acompanhamento da freqüência fundamental.

Vistos os aspectos da segmentação, nas próximas seções serão discutidos os três processamentos básicos *realizados sobre cada quadro*: estimação dos parâmetros temporais, espectrais e do período de pitch.

4.2.3 PARÂMETROS TEMPORAIS

Para a determinação do conjunto de parâmetros temporais foram utilizados, basicamente, os seguintes critérios:

a) opção por parâmetros de cálculo relativamente simples e que pudessem ser obtidos diretamente da forma de onda. Desta forma, foram excluídos os parâmetros de natureza temporal calculados com as técnicas de Processamento Homomórfico de sinais [2], [21];

b) utilização de parâmetros que acentuassem algumas características distintivas dos sinais, que podem ser perdidas ou mascaradas no espectro de potência, como, por exemplo, as assimetrias da forma de onda em relação ao eixo horizontal.

Após inúmeros testes, que avaliaram a eficácia de vários parâmetros descritos na literatura [2] e de outros, por nós escolhidos, chegou-se a um conjunto de quatro parâmetros temporais que atenderam aos critérios apresentados anteriormente. Estes parâmetros, calculados para cada quadro, estão indicados na figura 4.5. A energia total, Et, e o número de cruzamentos por zero, ZRX (ligeiramente modificado neste trabalho), são parâmetros tradicionais na análise de voz. Os dois outros parâmetros, número total de picos da forma de onda, NTP, e a

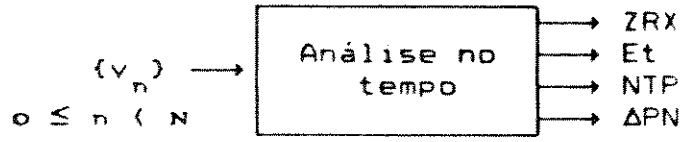


FIG.4.5 - Estimação de Parâmetros Temporais. (v_n) é a seqüência de amostras do quadro em análise.

diferença entre o número de picos do lado positivo e o número de picos do lado negativo da forma de onda, ΔPN, são aqui propostos para auxiliar a detecção de fricativos surdos e fricativos sonoros, respectivamente.

No restante desta seção são apresentados os detalhes do cálculo de cada um dos quatro parâmetros.

A) ENERGIA TOTAL (Et)

A energia total de cada quadro, em decibéis, é dada por:

$$Et = 10 \cdot \log(et1+et2) \qquad (4.1a)$$

onde

$$et1 = \sum_{n=0}^{N/2-1} (v_n)^2 \qquad (4.1b)$$

e

$$et2 = \sum_{n=N/2}^{N-1} (v_n)^2 \qquad (4.1c)$$

são a energia de cada metade do quadro, (v_n) é a seqüência de amostras do quadro de voz e N = 200 é o comprimento do quadro.

A Energia total é utilizada na determinação dos trechos vocálicos da elocução, como pode ser visto na figura 4.6. Esta figura apresenta a variação da energia total, Et, na pronúncia da palavra /ajuste/. No apêndice A são apresentados alguns histogramas, mostrando a distribuição da energia total para

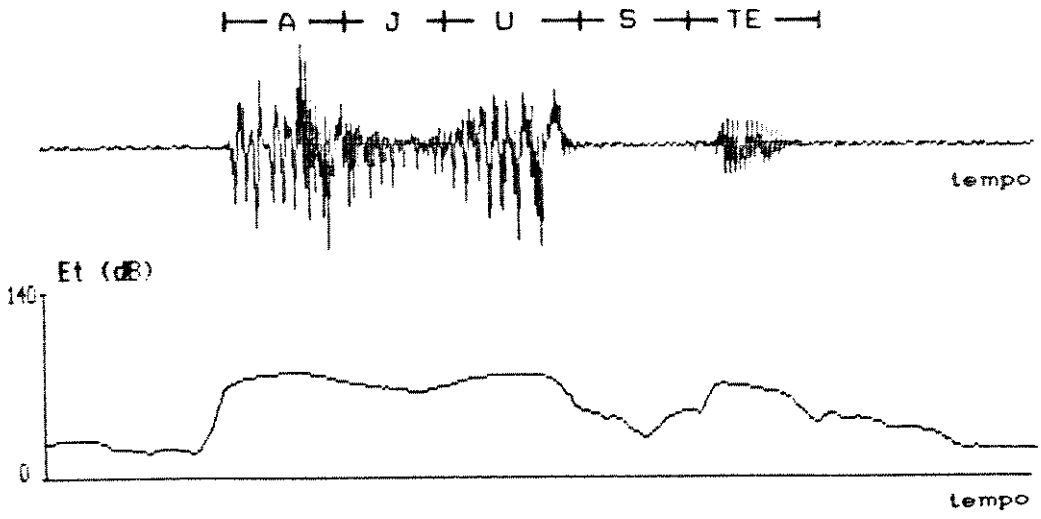


FIG.4.6 - Variação do parâmetro Et (energia total) ao longo da palavra /ajuste/.

vários tipos de sons. Os histogramas foram levantados para auxiliar a elaboração das regras para a classificação do quadro em uma das categorias fonéticas, como será visto posteriormente.

O cálculo de Et foi dividido em duas etapas para possibilitar a identificação da categoria "coarticulação", exemplificada na figura 4.7. Neste trabalho, a categoria "coarticulação" corresponde aos quadros situados na transição entre sons de pouca energia (como os fricativos surdos) e sons mais intensos (como os vocálicos). Para a detecção destas transições foi utilizada uma regra, que será discutida na

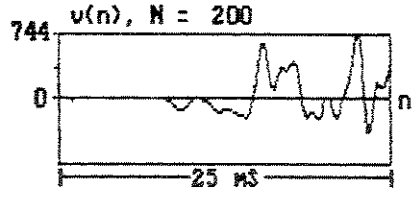


FIG. 4.7 - Exemplo de "coarticulação" na sílaba /pá/.

seção 4.2.8, que verifica se a relação et1/et2 (ou et2/et1), está acima de um determinado limiar. Deve ser ressaltado que, em lingüística, o termo coarticulação está mais ligado ao estudo da mudança do gesto articulatório de determinado fonema devido a um fonema vizinho. Como exemplo, na pronúncia do /s/, na sílaba /su/, ocorre um arredondamento dos lábios, já antecipando a pronúncia do /u/. É de se esperar, portanto, que as características acústicas do /s/, em /su/, sejam diferentes das características acústicas do /s/, em /si/.

B) CRUZAMENTOS POR ZERO (ZRX)

O número de cruzamentos por zero é um parâmetro de uso comum em Sistemas de Reconhecimento de Voz, sendo geralmente definido por [2]:

$$ZRX = \frac{1}{2} \sum_{n=1}^{N-1} \left| sinal(v_n) - sinal(v_{n-1}) \right| \qquad (4.2a)$$

onde

$$sinal(v_n) = \begin{cases} 1, & \text{se } v_n \geq 0 \\ -1, & \text{se } v_n < 0 \end{cases} \qquad (4.2b)$$

No nosso trabalho, foi observado que a substituição do "cruzamento por zero" pelo "cruzamento por um limiar" conduziu a melhores resultados. Isto pode ser visto no exemplo da figura 4.8a, onde é desejável que o parâmetro ZRX não seja influenciado pelas oscilações de pequena amplitude. Para contornar este problema, o cálculo do parâmetro ZRX foi realizado através do algoritmo 4.1, em substituição às equações 4.2a,b. Note-se que este algoritmo calcula, na realidade, o número de cruzamentos pelo limiar L da figura 4.8b.

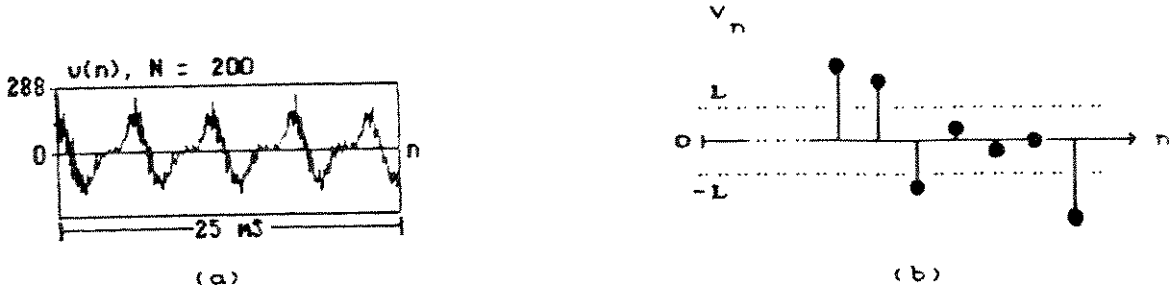


FIG. 4.8 - a) Sinal de amplitude elevada que apresenta oscilações "indesejáveis" (de baixa amplitude) em torno do zero; b) Introdução do limiar (L).

O aspecto a ressaltar no algoritmo 4.1 é o ajuste dinâmico do limiar, realizado da seguinte forma:

$$L = \begin{cases} 2, & |v_{max}| < 5 \\ 0, & 5 \leq |v_{max}| < 200 \\ |v_{max}|/200, & |v_{max}| \geq 200 \end{cases} \qquad (4.3)$$

onde v_{max} é a amostra de maior valor absoluto na seqüência $\{v_n\}$. Estas relações foram obtidas experimentalmente. O primeiro caso, $|v_{max}| < 5$, normalmente ocorre nos intervalos de silêncio entre palavras, resultando num valor nulo ou próximo de zero para ZRX; o segundo caso, $5 \leq |v_{max}| < 200$, é típico de sinais com pequena amplitude, como fricativos surdos; no último caso estão os vocálicos, que são os sinais de maior intensidade.

A figura 4.9 apresenta a variação do número de cruzamentos por zero, ZRX, ao longo da palavra /ajuste/. No Apêndice A são apresentados alguns histogramas do parâmetro ZRX; estes histogramas foram utilizados para a determinação dos

```
ZRX = 0;
para [n = 1; n < N; n = n + 1]
      se [ (v_n ≥ L  e  v_n-1 < -L) ou  (v_n ≤ -L  e  v_n-1 > L) ]
          ZRX = ZRX + 1;
```

ALG. 4.1 - Algoritmo para o cálculo do número de cruzamentos pelo zero, ZRX. O limiar, L, é determinado pela equação 4.3.

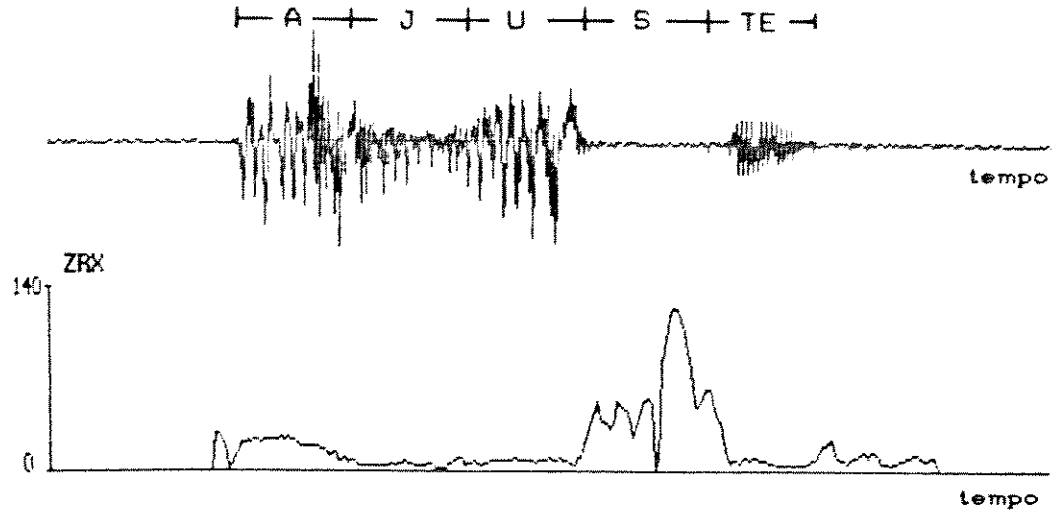


FIG.4.9 - Variação do parâmetro ZRX (cruzamentos por zero) ao longo da palavra /ajuste/.

limiares de decisão das regras do algoritmo de classificação, que será visto na seção 4.2.8.

Finalizando, pode ser observado pela figura 4.8a, que o número de cruzamentos por zero está bastante relacionado com a freqüência do primeiro formante, F1. Esta freqüência pode ser estimada grosseiramente por:

$$F1 \cong \frac{ZRX}{2 \cdot N \cdot T} \qquad (4.4a)$$

onde N é o número de amostras do bloco e T é o intervalo de amostragem; o produto N·T é a duração do quadro. Para os valores utilizados de N = 200 e T = 125 μs, ou seja, N·T = 25 ms, tem-se:

$$F1 \cong 20 \cdot ZRX \qquad (4.4b)$$

C) NÚMERO TOTAL DE PICOS (NTP)

O número total de picos, NTP, é um parâmetro que auxiliou a detecção de fricativos surdos de pequena intensidade, como o /f/ (figura 4.10). Ocorre que na pronúncia deste fonema, o

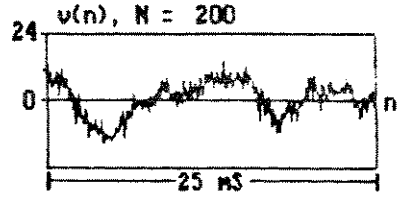


FIG.4.10 - Deslocamento do sinal de voz em relação ao eixo horizontal, em virtude do sopro, no som de /f/.

sopro (não audivel, de baixa freqüência e relativamente intenso) provoca um deslocamento do sinal em relação ao eixo n, diminuindo a contagem dos cruzamentos por zero. O efeito deste sopro pode ser reduzido colocando-se o microfone de um dos lados da boca, evitando a incidência direta do fluxo de ar sobre o transdutor.

O cálculo do parâmetro NTP foi realizado pelo algoritmo 4.2.

A figura 4.11 apresenta a variação do parâmetro NTP ao longo da palavra /ajuste/. Alguns histogramas, mostrando a distribuição do parâmetro NTP para vários sons, estão no Apêndice A.

```
PPOS = 0;
PNEG = 0;
para [i=1; i(N; i = i+1]
     se [(v_n ≥ 0) e (v_n ≥ v_n-1) e (v_n ) v_n+1)]
          PPOS = PPOS + 1;

     se [(v_n ( 0) e (v_n ≤ v_n-1) e (v_n ( v_n+1)]
          PNEG = PNEG + 1;

NTP = PPOS + PNEG;
```

ALG.4.2 - Algoritmo para o cálculo do número total de picos, NTP. As variáveis PPOS e PNEG correspondem ao número de picos da parte positiva e da parte negativa do sinal, respectivamente.

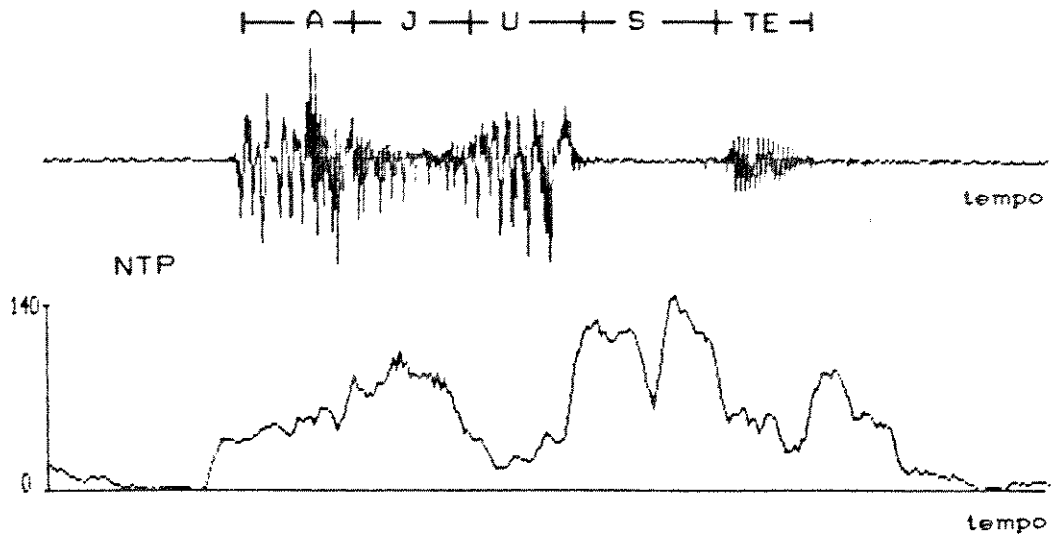


FIG.4.11 - Variação do parâmetro **NTP** ao longo da palavra /ajuste/.

D) DIFERENÇA ENTRE O NÚMERO DE PICOS POSITIVOS E NEGATIVOS (ΔPN)

A limitação da faixa do sinal digitalizado às freqüências abaixo de 3,4 kHz acarreta sérias dificuldades na análise de fricativos. Particularmente, os fricativos sonoros podem ser facilmente confundidos com vogais de pequena intensidade. Esta dificuldade pôde ser razoavelmente contornada através do parâmetro ΔPN.

No capítulo 3 foi visto que os fricativos sonoros apresentam uma assimetria muito peculiar: o lado positivo da forma de onda, correspondente à compressão do ar, apresenta uma característica de turbulência e o lado negativo, associado à rarefação do ar, é mais suave. Esta assimetria torna-se mais acentuada à medida que as palavras são pronunciadas mais cuidadosamente. A diferença entre o número de picos positivos e negativos, ΔPN, é, portanto, um parâmetro razoável para a caracterização desta categoria. Este parâmetro foi obtido por:

$$\Delta PN = PPOS - PNEG \qquad (4.5)$$

onde PPOS e PNEG, calculados pelo algoritmo 4.2, correspondem ao

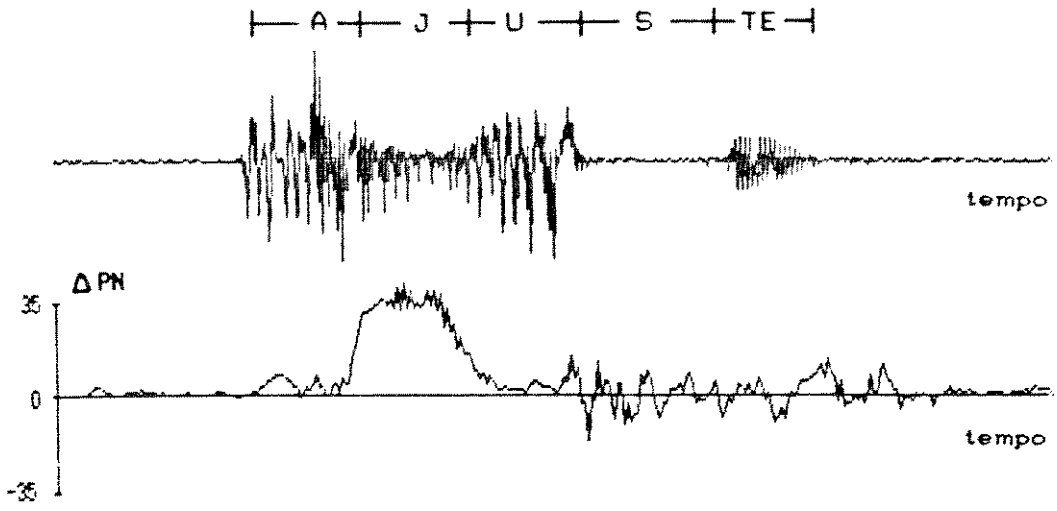


FIG.4.12 - Variação do parâmetro ΔPN ao longo da palavra /ajuste/.

número de picos positivos e negativos, respectivamente.

O comportamento de ΔPN, de forma semelhante aos três parâmetros discutidos anteriormente, é apresentado na figura 4.12 (variação temporal) e no Apêndice A (histogramas).

É importante ressaltar que no nosso sistema de aquisição de voz, SAPDV-A, a seqüência digitalizada sofre uma inversão de fase, compensando uma inversão provocada pelo microfone utilizado. Desta forma, o lado positivo da seqüência digitalizada corresponde à maior intensidade da onda de pressão acústica no microfone. Para a utilização deste parâmetro em outros sistemas (ou entrando-se com a voz através de um gravador ou outro microfone no SAPDV-A) deve ser verificado se o número total de inversões de fase é ímpar, quando então deve ser utilizado o parâmetro ΔNP (diferença entre o número de picos negativos e positivos do sinal) em lugar do parâmetro ΔPN. Note-se este tipo de parâmetro, que depende da fase do sinal, não deve ser utilizado quando o sinal de voz for transmitido através de uma linha telefônica, pois o atraso introduzido pelo canal é desconhecido.

4.2.4 ANÁLISE LPC

Além dos parâmetros temporais, a extração de parâmetros espectrais é de fundamental importância para a caracterização dos sinais de voz. A análise LPC, discutida nesta seção, foi a ferramenta básica utilizada para a estimativa da resposta em freqüência do trato vocal, donde são extraídos os parâmetros espectrais.

A simplicidade dos algoritmos, a precisão dos resultados e o aspecto suave dos espectros obtidos com as técnicas de Predição Linear [2], [3], [6], [20], foram os fatores que determinaram a escolha da análise LPC em lugar das técnicas de Processamento Homomórfico [2], [4], [21]. A análise cepstral realizada com o Processamento Homomórfico, além de ser computacionalmente muito dispendiosa, resulta em estimativas espectrais contendo vários picos, dificultando a localização dos formantes. Estes picos extras podem ser atribuídos a um mal posicionamento da janela cepstral (filtro passa-baixos tempos). Este posicionamento é bastante crítico e deve ser feito dinamicamente, em função do período de pitch.

Nesta seção serão apresentados os aspectos práticos e os resultados da análise LPC, representada esquematicamente na figura 4.13.

A) ORDEM DO PREDITOR

Como foi visto no capítulo anterior, a função de sistema do filtro do modelo digital simplificado da produção da fala pode ser dada por

(3.22)
$$H(z) = \frac{G}{1 - \sum_{i=1}^{M} a_i \cdot z^{-i}}$$

onde G é o fator de ganho, $\{a_i\}$ é o conjunto de coeficientes do polinômio característico do filtro e M a ordem do preditor, que

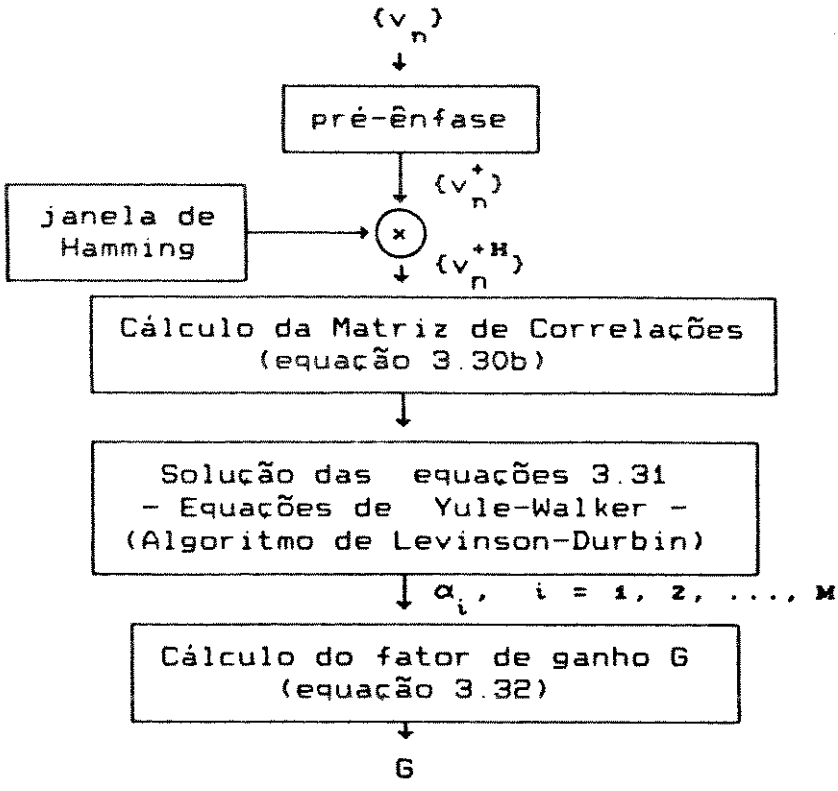


FIG 4.13 - Análise LPC

deve ser determinada. Recorrendo à equação 3.19b, o valor de M foi estabelecido inicialmente em M = 8, pois f_s = 8 kHz; em seguida, avaliando-se a qualidade das estimativas espectrais obtidas com M maior e menor que 8, observou-se que o valor mais adequado para este trabalho foi M = 12. Com este valor, as ressonâncias espectrais mostraram-se bem acentuadas, e não houve o aparecimento de picos espúrios, provocados pelos lóbulos do espectro do pulso glotal.

Estabelecida a ordem do preditor, serão apresentados os detalhes da estimativa da resposta em freqüência do trato vocal, acompanhando o esquema da figura 4.13.

B) PRÉ-ÊNFASE

A pré-ênfase é utilizada para compensar a queda espectral de 6 dB/oitava no espectro do sinal de voz, devido à

combinação da envoltória do filtro conformador, $G(z)$, e da irradiação, $T(z)$. A pré-ênfase sobre o sinal de voz, (v_n), foi implementada através da expressão usual [3].

$$v_n^+ = v_n - a \cdot v_{n-1} \qquad 0 \leq n < 200 \qquad (4.6)$$

onde (v_n^+) é o sinal pré-enfatizado e "a" é um número real positivo com valor próximo de 1 (entre 0,9 e 1,0). A figura 4.14 apresenta as curvas de pré-ênfase para valores típicos de "a". As curvas foram obtidas fazendo-se $z = e^{j\omega}$ em

$$Pr(z) = \frac{V^+(z)}{V(z)} = 1 - a \cdot z^{-1}, \qquad (4.7)$$

onde $V^+(z)$ e $V(z)$ representam as Transformadas $\mathcal{Z}$ de (v_n^+) e (v_n), respectivamente.

Uma análise da equação 3.22, reapresentada no início desta seção, nos permite escrevê-la da seguinte forma:

$$H(z) = \begin{cases} G(z) \cdot F(z) \cdot T(z), & \text{para sons sonoros} \\ F(z) \cdot T(z), & \text{para sons não sonoros (surdos)} \end{cases} \qquad (4.8)$$

onde $G(z)$, $F(z)$ e $T(z)$ são as Transformadas $\mathcal{Z}$ do filtro conformador do pulso glotal, do aparelho fonador e da impedância de irradiação, respectivamente. Sob o ponto de vista da produção da fala, como está expresso na equação 4.8, a pré-ênfase somente deveria ser aplicada aos intervalos sonoros, onde a combinação de

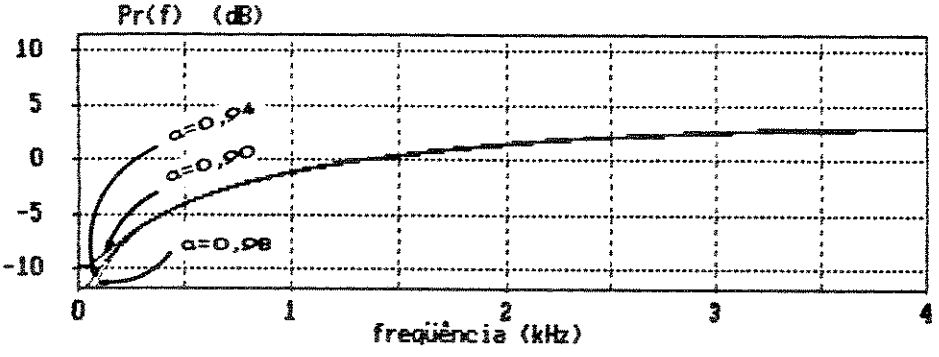


FIG. 4.14 - Curvas de pré-ênfase para valores típicos de "a"

$T(z)$ e da envoltória de $G(z)$ provocam uma queda de 6 dB/oitava no espectro de $F(z)$. Neste caso, o valor de "a" pode ser ajustado dinamicamente por

$$a = |R(1)/R(0)| \qquad (4.9)$$

onde R(1) e R(0) correspondem à função de autocorrelação a curto prazo da seqüência de voz, calculada com atrasos 1 e 0, respectivamente. Nos sons sonoros, em que há uma forte correlação entre as amostras, o valor de "a" será próximo de 1; para os sons não sonoros, "a" assumirá um valor pequeno. No nosso trabalho, contudo, utilizou-se uma pré-ênfase fixa de 90% (a = 0,9), para quaisquer sons. Foi verificado que a pré-ênfase também contribuiu para a identificação dos fricativos surdos. É curioso lembrar que a pré-ênfase aplicada a fricativos surdos equivale ao efeito da primeira ressonância do ouvido externo (seção 3.7.1). Adiantando o assunto do próximo item desta seção, a pré-ênfase (com sua característica passa-altas - figura 4.14) também é recomendada [5] para evitar o aparecimento do espectro da janela de Hamming na região de baixas freqüências da estimativa espectral do trato vocal, quando a seqüência $\{v_n\}$ tiver um valor médio não nulo.

C) JANELA DE HAMMING

Após a pré-ênfase, a seqüência $\{v_n^+\}$ é multiplicada pela janela de Hamming, resultando em

$$v_n^{+H} = v_n^+ \cdot H_n \qquad 0 \leq n < N \qquad (4.10a)$$

onde

$$H_n = \begin{cases} 0,54 - 0,46 \cdot \cos\left(\frac{2\pi n}{N-1}\right), & 0 \leq n < N \\ 0, & \text{caso contrário} \end{cases} \qquad (4.10b)$$

é a janela de Hamming.

Com este janelamento, as características espectrais do centro do quadro são mantidas e as transições abruptas das extremidades são eliminadas.

A janela de Hamming é uma função consagrada na análise de voz; não foram feitas experiências com outras janelas. A multiplicação da seqüência de voz pela janela de Hamming antes da pré-ênfase, segundo Markel & Gray [3], não produz mudanças significativas nos resultados.

A ponderação pela janela de Hamming (ou outra janela equivalente) também é indicada para reduzir o erros de predição do Método da Autocorrelação, que são maiores no início e fim do segmento, pois, para $0 \leq n < M$ (supondo um preditor LPC de ordem M), v_n é prevista a partir de amostras nulas; de forma semelhante, para $N \leq n < N+M$, v_n é nula e é estimada a partir de amostras não nulas.

D) ALGORITMO DE LEVINSON-DURBIN

Após o cálculo das autocorrelações a curto prazo (equação 3.30b), o sistema de equações 3.31 (Equações de Yule-Walker) deve ser resolvido para a determinação das estimativas dos parâmetros do trato vocal, conforme a figura 4.13.

Explorando a simetria da matriz de autocorrelações e o fato de os elementos da diagonal principal, assim como os elementos de qualquer diagonal paralela a ela, serem iguais entre si (matriz Toeplitz), Levinson desenvolveu um eficiente algoritmo recursivo para a solução do sistema, sem a necessidade de inverter a matriz de autocorrelações; neste algoritmo [5], os elementos do vetor Mx1 da parte direita da equação 3.31 são considerados elementos genéricos. Observando que os elementos deste vetor são os mesmos elementos de qualquer linha da matriz de autocorrelações, Durbin aprimorou o algoritmo de Levinson, tornando-o mais rápido e utilizando menos posições de memória [6].

O algoritmo de Levinson-Durbin, cuja demonstração pode ser encontrada em [22], é dado por

$$E^{(0)} = R(0) \qquad (4.11a)$$

para $1 \leq i \leq M$:

$$r_i = \frac{R(i) - \sum_{j=1}^{i-1} \alpha_j^{(i-1)} R(i-j)}{E^{(i-1)}} \qquad (4.11b)$$

$$\alpha_i^{(i)} = r_i \qquad (4.11c)$$

$$\alpha_j^{(i)} = \alpha_j^{(i-1)} - r_i \alpha_{i-j}^{(i-1)} \qquad 1 \leq j \leq i-1 \qquad (4.11d)$$

$$E^{(i)} = (1-r_i^2)E^{(i-1)} \qquad (4.11e)$$

onde $E^{(i)}$ é a energia do sinal de erro na i-ésima iteração, R(i) é a função de autocorrelação a curto prazo com atraso i, r_i é o i-ésimo coeficiente de reflexão e $\alpha_j^{(i)}$ é o valor do j-ésimo termo a ser determinado, na i-ésima iteração. As equações 4.11b-e são resolvidas recursivamente para i = 1, 2, ... M, resultando em

$$\alpha_j = \alpha_j^{(M)} \qquad 1 \leq j \leq M \qquad (4.11f)$$

Neste algoritmo, a determinação dos coeficientes de um preditor de ordem M implica no cálculo intermediáriao dos coeficientes de todos os preditores de ordem menor que M. No apêndice B é apresentada a implementação do algoritmo, em linguagem C.

Após o cálculo dos coeficientes α, o ganho é determinado pela equação 3.32. Em seguida, é calculada a resposta em freqüência do trato vocal, como será visto na próxima seção.

4.2.5 RESPOSTA EM FREQÜÊNCIA DO TRATO VOCAL

Os coeficientes "α", determinados conforme a figura 4.13, são utilizados como estimativas dos coeficientes "a" do modelo do trato vocal, uma vez que

a) a escolha apropriada da ordem do preditor impediu que os lóbulos do espectro do pulso glotal estivessem presentes na estimativa da resposta em freqüência do trato vocal;

b) a pré-ênfase compensou a queda de 6 dB/oitava na resposta em freqüência do trato vocal, devido ao efeito combinado da envoltória do espectro do pulso glotal e da irradiação.

Portando, a estimativa da resposta em freqüência do trato vocal, $\hat{F}(e^{j\omega})$, pode ser calculada fazendo-se $z = e^{j\omega}$ e $\{a_i\} = \{\alpha_i\}$ na equação 4.6, resultando em

$$\hat{F}(e^{j\omega}) = \frac{G}{1 - \sum_{i=1}^{M} \alpha_i \cdot e^{-j\omega i}} \qquad 0 \leq |\omega| \leq \pi \qquad (4.12)$$

Para o cálculo de $\hat{F}(e^{j\omega})$ em um número muito elevado de freqüências, a equação 4.12 não é adequada. Neste trabalho, optou-se por um método mais eficiente, proposto por Markel [5]. A idéia básica deste método consiste em calcular o denominador da equação 4.12 para um elevado número de freqüências, com o auxílio de um algoritmo FFT. Isto é realizado explorando o fato de a seqüência $\{1, -\alpha_1, -\alpha_2, \ldots -\alpha_M\}$, correspondente aos coeficientes do denominador da equação 4.12 (filtro inverso), ser uma seqüência de comprimento finito; desta forma, a resposta em freqüência discreta pode ser obtida eficientemente com o cálculo da Transformada Discreta de Fourier, em tantos pontos quanto a resolução em freqüência desejada, ou seja,

$$A_k \underset{0 \leq k < L}{=} DFT\left\{ 1, -\alpha_1, -\alpha_2, \ldots, -\alpha_M, \underbrace{0, 0, \ldots, 0}_{L-M-1 \text{ zeros}} \right\} \qquad (4.13)$$

onde $A_k = A(e^{j\omega})$ para $\omega = (2\pi/L)k$, $A(e^{j\omega})$ é a Transformada de Fourier do denominador da equação 4.12 e L é o número de pontos da DFT. No nosso trabalho, L foi fixado em 1024, proporcionando

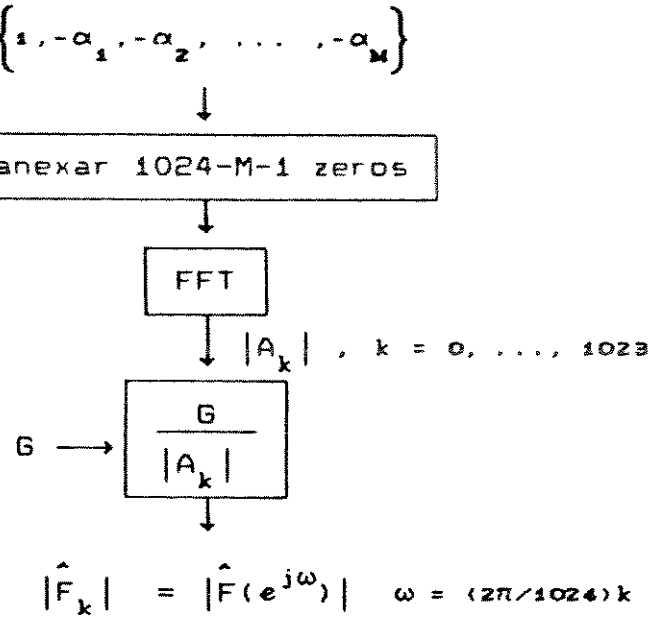


FIG.4.15 - Procedimento para o cálculo da resposta em freqüência discreta do trato vocal a partir dos coeficientes da Predição Linear.

uma resolução espectral de 8 kHz/1024 ≅ 7,8 Hz. Esta resolução permitiu uma estimativa simples e precisa dos formantes, a partir da localização dos picos da resposta em freqüência do trato vocal.

Devido à inclusão de um elevado número de zeros (1024 - 12 - 1 = 1011), o cálculo da DFT foi realizado com um algoritmo FFT em que as operações com zero são eliminadas ("prunned") [7], [8]. O programa da FFT, em linguagem C, é apresentado no Apêndice B.

Após o cálculo da DFT do filtro inverso, a estimativa da amplitude da resposta em freqüência discreta do trato vocal,

$$|\hat{F}_k| = |\hat{F}(e^{j\omega k})| \qquad 0 \leq k < 1024 \qquad (4.14)$$

é determinada por:

$$|\hat{F}_k| = \frac{G}{|A_k|} \quad , \; 0 \leq k < 1024 \qquad (4.15a)$$

onde

$$|A_k| = \left[Re^2(A_k) + Im^2(A_k)\right]^{1/2} \qquad (4\ 15b)$$

e Re() e Im() denotam a parte real e imaginária, respectivamente. Note-se que o módulo de $\hat{F}_k$ é suficiente para as análises pretendidas, uma vez que o ouvido humano é insensível às variações de fase, como foi discutido no capítulo 2.

O diagrama em blocos da figura 4.15 apresenta esquematicamente o procedimento descrito nos parágrafos anteriores. Esta figura é complementada com a figura 4.13.

Nas figura 4.16a-f são apresentadas diversas estimativas da resposta em freqüência do trato vocal. As estimativas estão sobrepostas à DFT da seqüência v_n^{+H} (sinal de voz com pré-ênfase e janelamento de Hamming) para que se possa fazer uma avaliação do método.

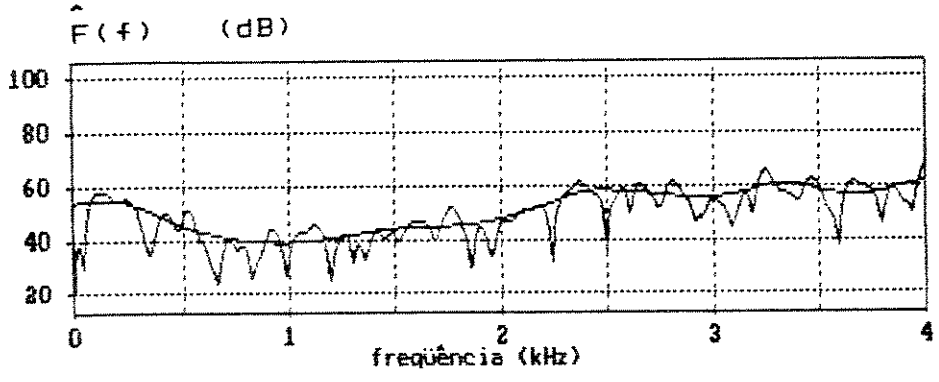


FIG. 4.16a - Fricativo Surdo /ch/ em /chá/

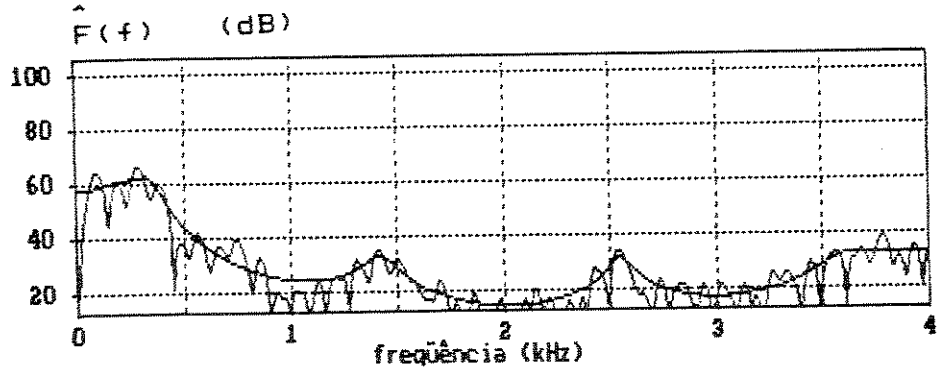


FIG. 4.16b - Fricativo Sonoro /z/ em /dezoito/

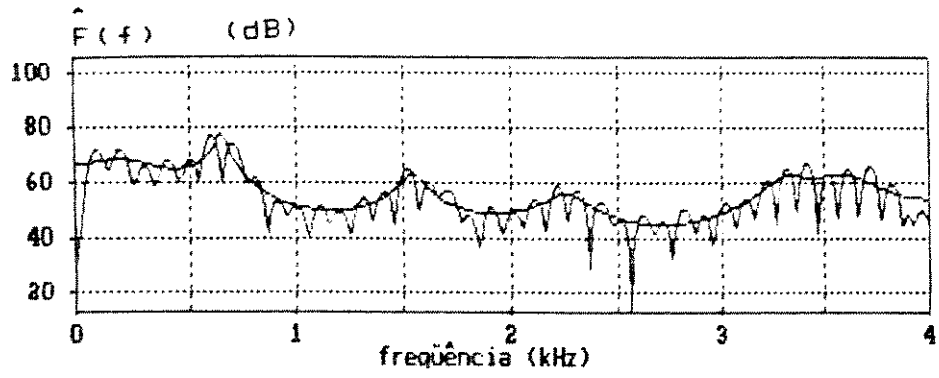


FIG. 4.16c - Vogal /a/ em /chá/.

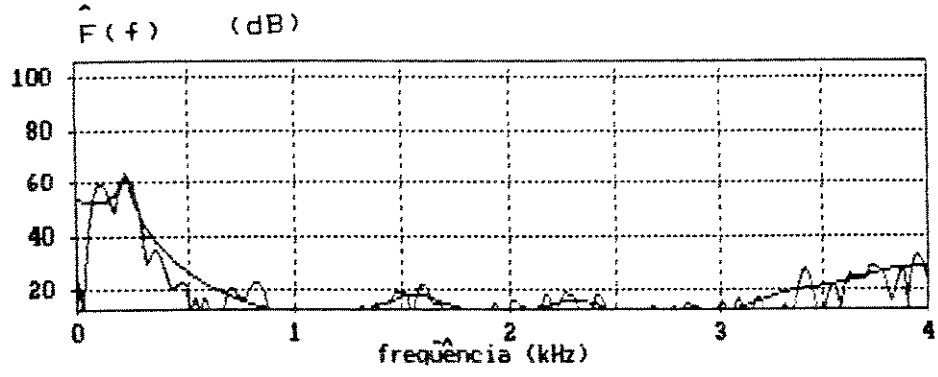


FIG. 4.16d - Oclusão Sonora do /b/ em /bá/.

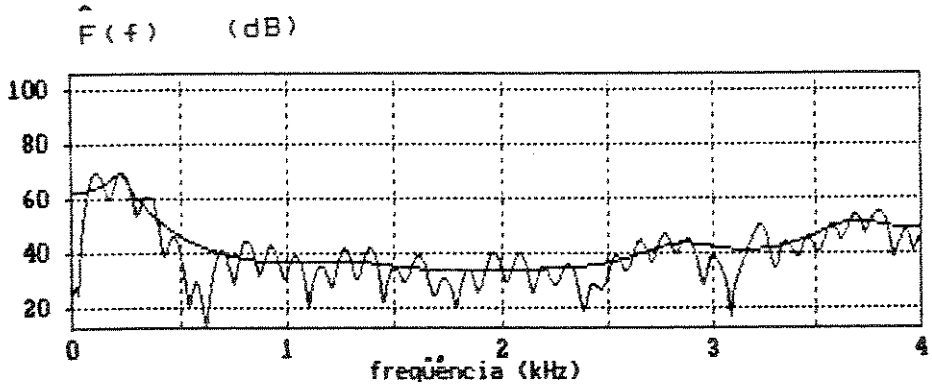


FIG. 4.16e - Vocálico /m/ em /má/.

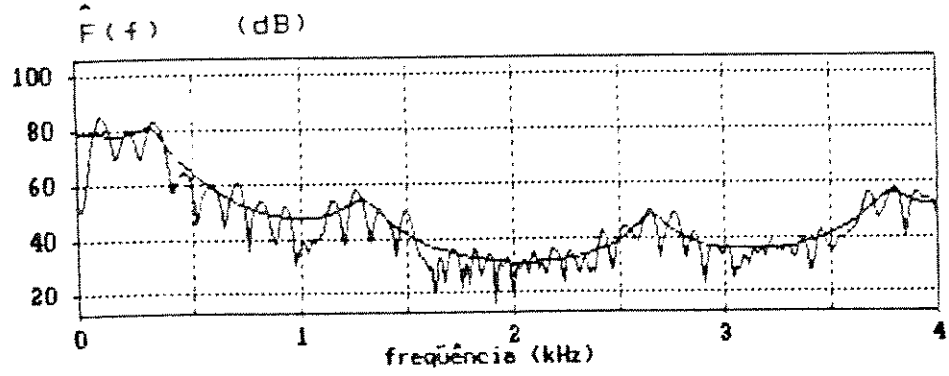


FIG. 4.16f - Vocálico /l/ em /lá/.

4.2.6 PARÂMETROS ESPECTRAIS

A) LOGARITMO DA ENERGIA EM FAIXAS SELECIONADAS

A escolha das faixas de freqüência a serem utilizadas foi baseada nas características espectrais dos diversos sons, discutidas no capítulo 3 e na literatura disponível [9-14], [23]. O fato de se trabalhar com um sinal limitado em 4 kHz foi a maior dificuldade encontrada neste trabalho. Foram avaliadas várias combinações de faixas de freqüência, testando-se tanto a quantidade de faixas quanto os limites de cada faixa. Após diversos testes, que envolveram a análise da voz de mulheres e homens adultos, buscando a determinação de características espectrais que fossem *invariantes com o locutor*, foi possível a implementação dos algoritmos de classificação utilizando medidas em apenas quatro faixas de freqüência, calibradas da seguinte forma: a faixa total (0-4000 Hz), e as faixas de 0-500 Hz , de 750-2000 Hz e de 1000-3000 Hz, associadas aos parâmetros LogEne(0), LogEne(1), LogEne(2) e LogEne(3), respectivamente. Estes parâmetros foram calculados por

$$\text{LogEne}(i) = 10 \cdot \log \left(\sum_{k=p}^{q} |\hat{F}_k|^2 \right)^{1/2}, \quad 0 \le i \le 3 \qquad (4.16)$$

onde $\hat{F}_k$ é a k-ésima amostra da estimativa espectral discreta do trato vocal (equação 4.15a) e os limites do somatório são dados por

LogEne(0): faixa(0-4000 Hz) → (p, q) = (0, 511)
LogEne(1): faixa(0-500 Hz) → (p, q) = (0, 64)
LogEne(2): faixa(750-2000 Hz) → (p, q) = (96, 256)
LogEne(3): faixa(1000-3000 Hz) → (p, q) = (128, 384)

A descrição do algoritmo que combina estes quatro parâmetros em freqüência e os outros quatro, de natureza temporal, para determinar a categoria fonética associada ao quadro em análise, será adiada para a seção 4.2.8. No restante desta seção e na seção 4.2.7, serão discutidos os métodos empregados para a estimativa das outras saídas do Módulo Frontal, isto é, os três primeiros formantes, F1, F2 e F3, e a freqüência fundamental, f0.

B) FORMANTES

Os formantes, ou ressonâncias do aparelho fonador, constituem um subconjunto dos picos da estimativa da amplitude da

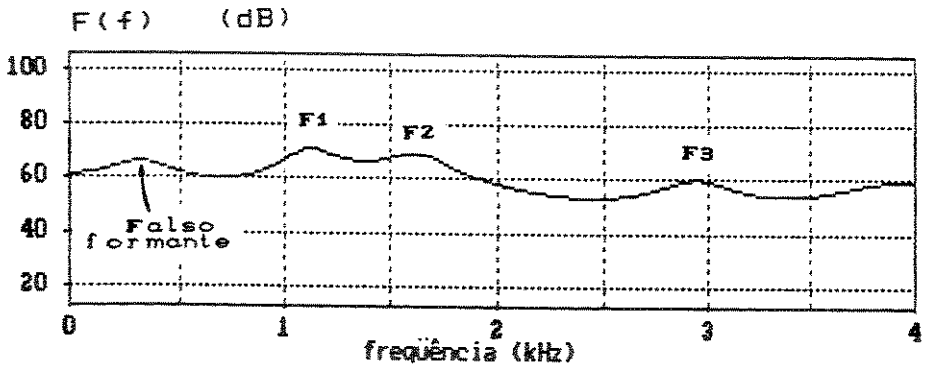


FIG.4.17a - Exemplo de um "falso formante" na estimativa da resposta em freqüência do trato vocal.

resposta em freqüência do trato vocal, $|\hat{F}_k|$, ou seja, nem todos os picos correspondem a formantes. Como exemplo, é comum o aparecimento de um pico com uma freqüência inferior a F1, (figura 4.17a). Foi observado que este "falso formante" geralmente acontece na pronúncia das vogais /a/ ou /o/, se:

a) o valor de F1 é relativamente alto (acima de 1 kHz);
b) não existem vales espectrais acentuados, isto é, a pronúncia não é nasalizada;
c) o quarto formante não está presente na estimativa espectral.

Na existência de todas estas condições, o preditor de ordem M = 12 está superdimensionado, possibilitando a modelagem da região abaixo de F1 (que é uma região de alta energia) por um pico espectral. A seqüência de figuras 4.17b-e mostra o surgimento do falso formante e de outros picos adicionais em $|\hat{F}_k|$, à medida que a ordem do preditor é aumentada.

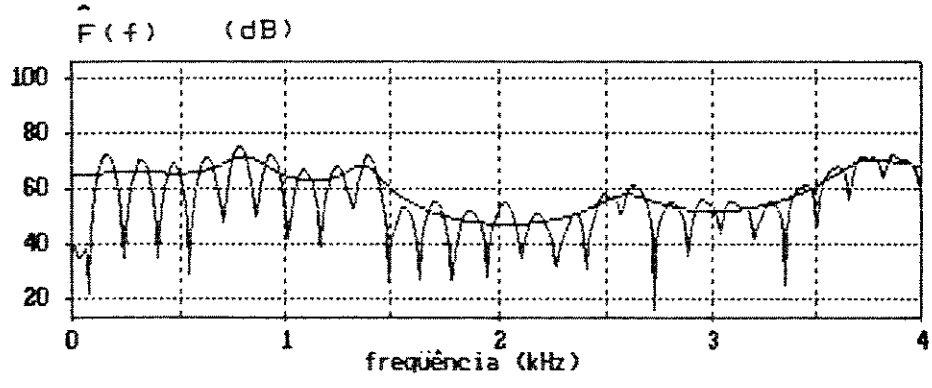


FIG. 4.17b - Estimativa espectral com preditor de ordem 12.

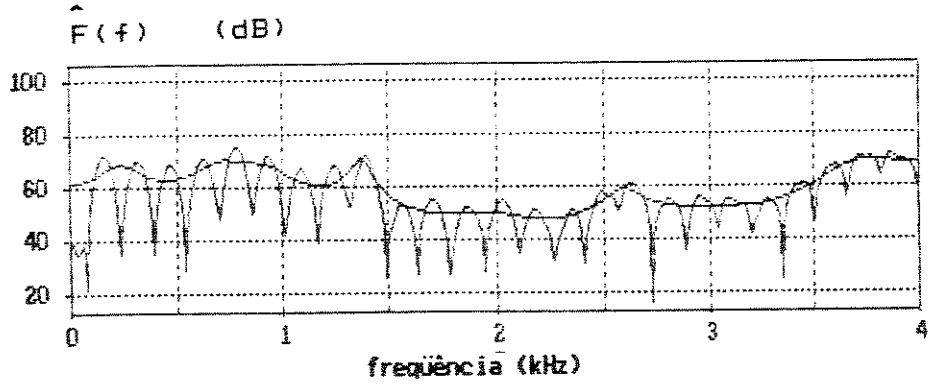


FIG. 4.17c - Estimativa espectral com preditor de ordem 18.

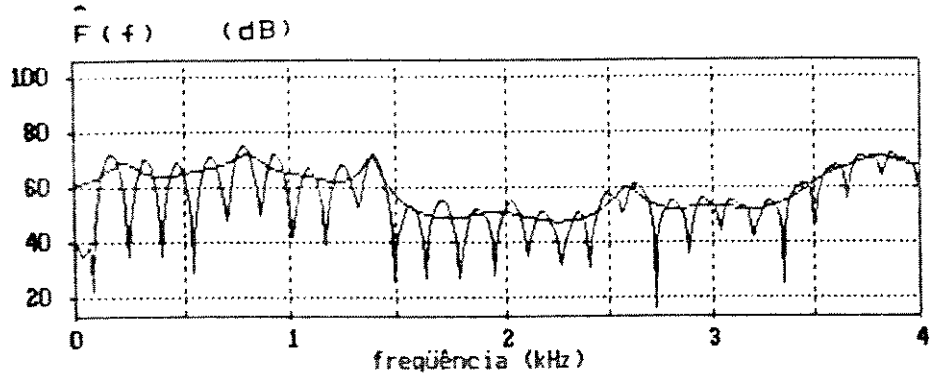


FIG. 4.17d - Estimativa espectral com preditor de ordem 24

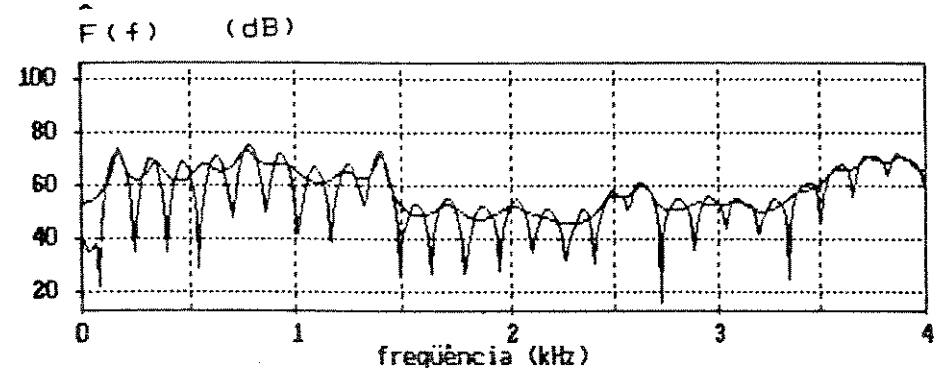


FIG. 4.17e - Estimativa espectral com preditor de ordem 40

A determinação dos picos espectrais correspondentes aos formantes requer o acopanhamento do movimento dos picos ao longo do tempo, assim como o conhecimento das regiões do espectro onde podem ocorrer cada formante. No nosso trabalho não foi implementado nenhum algoritmo para o acompanhamento ("tracking") de formantes. A estratégia utilizada para a estimação de F1, F2 e F3 para um determinado quadro é independente dos quadros adjacentes. Inicialmente, as freqüências dos possíveis formantes são estimadas a partir de uma interpolação parabólica para as curvas de ressonância de $\{|\hat{F}_k|\}$. A parábola que passa por um pico espectral e pelos dois pontos adjacentes está indicada na figura 4.18. Os valores (em Hz) da freqüência de pico, f_p, e da largura de faixa (-3 dB), bw, podem ser facilmente obtidos em função de

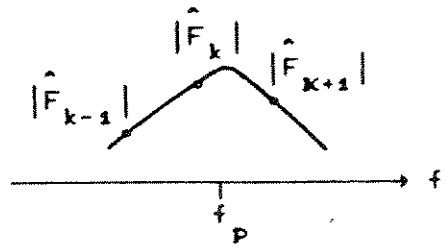


FIG.4.18 - Exemplo de um pico espectral. f_p é a freqüência do pico da parábola.

$|\hat{F}_{k-1}|$, $|\hat{F}_k|$ e $|\hat{F}_{k+1}|$, resultando em:

$$f_p = -b/2a \qquad (4.17a)$$

e

$$bw = \frac{-fs}{a \cdot L} \sqrt{\frac{b^2}{2} - 2ac}\ , \qquad (4.17b)$$

sendo

$$a = (|\hat{F}_{k+1}| + |\hat{F}_{k-1}|)/2 - |\hat{F}_k|, \qquad (4.17c)$$

$$b = (|\hat{F}_{k+1}| - |\hat{F}_{k-1}|)/2, \qquad (4.17d)$$

$$c = |\hat{F}_k|, \qquad (4.17e)$$

onde fs = 8 kHz é a freqüência de amostragem e L = 1024 é o número de pontos da DFT. O número de picos está limitado teoricamente em M/2 = 6, onde M = 12 é a ordem do preditor; na prática, verificou-se que, para os sons vocálicos, o número de picos detectados oscila entre 2 e 5, sendo detectados exatamente 3 picos em cerca de 90% dos casos. Finalmente, definindo-se

f_i = freqüência do i-ésimo pico,
a_i = amplitude do i-ésimo pico ,
bw_i = largura de faixa do i-ésimo pico,
i = 1, ..., tot,

onde tot é o total de picos detectados, procede-se à determinação dos picos correspondentes aos três primeiros formantes, F1, F2 e F3, através do algoritmo 4.3. O algoritmo verifica três

possibilidades

a) se o total de picos é menor que 3, os dados são considerados insuficientes para a análise;

b) se o número de picos é exatamente 3, a freqüência dos picos é associada diretamemente à freqüência dos formantes;

c) se o número de picos é maior que 3, é verificada a ocorrência do falso formante (figura 4.17); o primeiro pico será eliminado se sua amplitude for menor que a amplitude do segundo pico ou se seu fator de mérito (f_1/bw_1) for menor que 1,5; este limiar foi determinado experimentalmente. Desta forma, não havendo o falso formante, são aproveitados os três primeiros picos.

```
se [tot<3]
    retorne;
se [tot = 3]
    para [j=1; j≤3; j=j+1]
        F_j = f_j;
se [tot>3]
    se [a1<a2 ou f1/bw1<1,5]
        para [j=1; j<3; j=j+1]
            F_j = f_j+1;
    senão
        para [j=1; j<3; j=j+1]
            F_j = f_j;
```

ALG.4.3 - Determinação dos três primeiros formantes a partir dos picos da estimativa da resposta em freqüência do trato vocal.

Nas figuras 4.19a,b são apresentados alguns espectrogramas obtidos com a aplicação do algoritmo 4.3.

Como foi dito no início da seção, neste trabalho não foi realizado nenhum algoritmo para o acompanhamento dos formantes. Alguns exemplos de tais algoritmos podem ser encontrados em [4], [15] e [16].

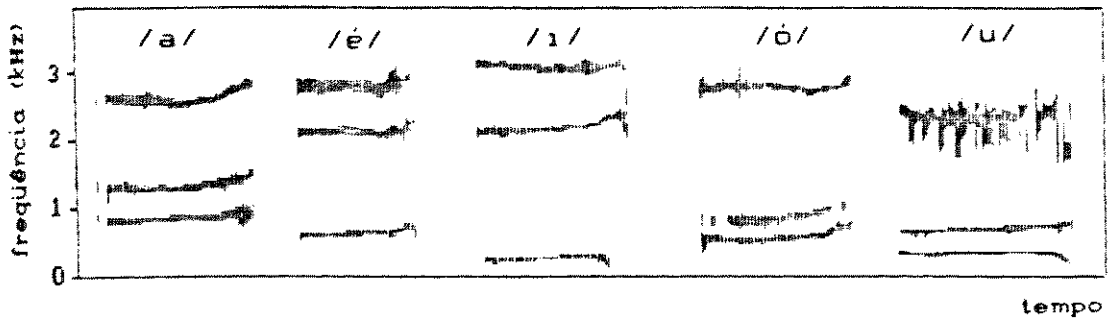


FIG. 4. 19a - Espectrograma das vogais /a/, /e/, /i/, /o/, /u/, de um mesmo locutor.

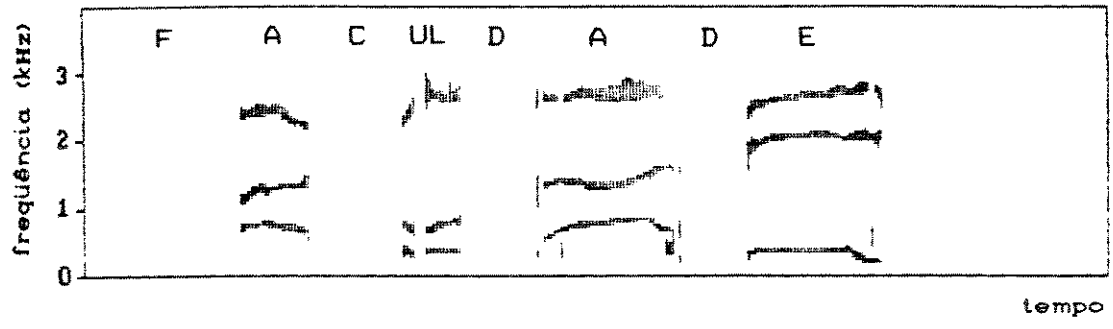


FIG.4.19b - Espectrograma dos trechos vocálicos da palavra /faculdade/.

4.2.7 ESTIMATIVA DE F0

A freqüência fundamental de excitação, f0, é estimada pela média dos pulsos pseudo-periódicos produzidos pelo movimento das cordas vocais, durante os 25 ms de análise. O uso do pitch[1]

[1] Em processamento de voz (e neste trabalho, inclusive), os termos pitch e freqüência fundamental são utilizados como sinônimos, embora o conceito de pitch seja mais abrangente. A rigor, o pitch de um determinado estímulo sonoro (não necessariamente um sinal de voz), corresponde à freqüência, em Hz, de um tom senoidal que está "afinado" com o estímulo, segundo a percepção auditiva de um determinado indivíduo. Como, na percepção de voz, o pitch dos sons sonoros geralmente corresponde ao valor da freqüência fundamental, para as pessoas com audição normal, os dois termos passaram a ser empregados indistintamente. Detalhes sobre o assunto podem ser encontrados em [24] e [25].

no reconhecimento de voz é importante por três fatores fundamentais:

a) a hipótese de independência entre o aparelho fonador e o sistema glotal não é totalmente verdadeira, de forma que cada vogal interfere de modo diferente na excitação. Como exemplo, na pronúncia do ditongo /ai/, em uma fala natural, a vogal /a/ terá um valor de f0 ligeiramente menor que o da vogal /i/ [17];

b) através das curvas de variação de f0 é possível, por exemplo, determinar o caráter afirmativo ou interrogativo de uma frase;

c) as variações de f0, juntamente com a duração e a amplitude das vogais, permitem a determinação do acento ortográfico das palavras.

Após a elaboração de um algoritmo de análise em freqüência, explorando a periodicidade das ondulações do espectro a curto prazo (figura 4.20), com resultados não satisfatórios, optou-se pelo desenvolvimento de um algoritmo de análise no tempo, utilizando a AMDF (Average Magnitude Difference Function). A técnica da AMDF é um bom compromisso entre complexidade computacional e precisão dos resultados [19]. O algoritmo de análise em freqüência foi abandonado devido às dificuldades encontradas para determinar, automaticamente, que regiões do espectro de cada quadro apresentam ondulações regulares.

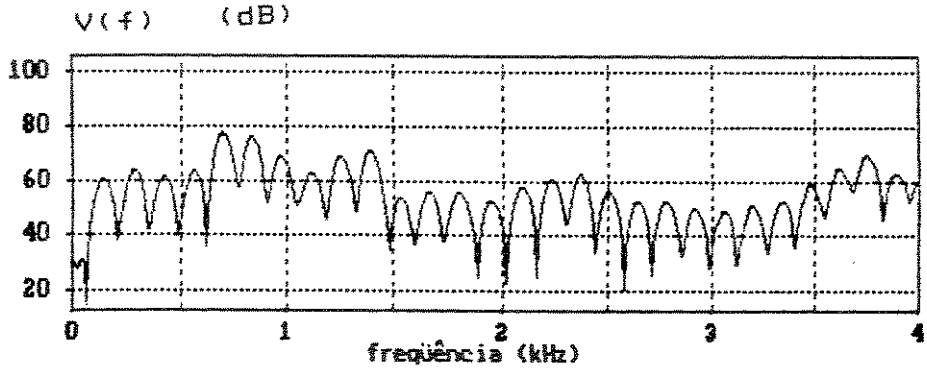


FIG.4.20a - vogal /a/. Note-se que a regularidade das ondulações (associadas à freqüência da excitação impulsiva) mantém-se por toda a faixa.

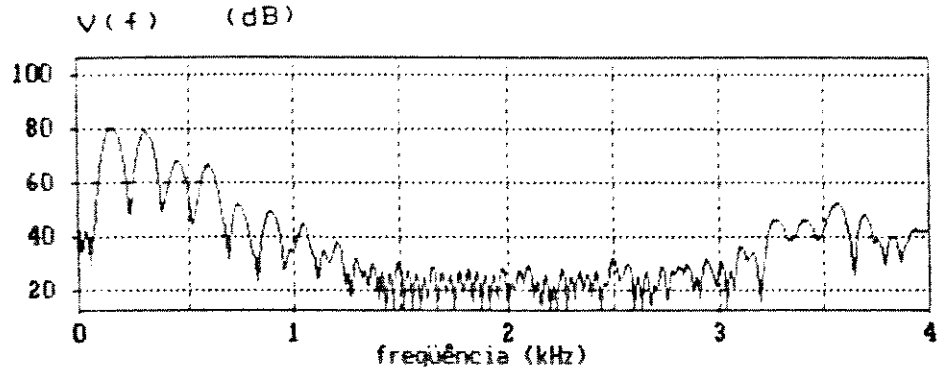


FIG.4.20b - vogal /u/. Note-se a quebra de regularidade na faixa de 1,4-3,0 kHz (no caso particular deste espectro), devido à constricção que é formada pelo arredondamento dos lábios.

A) AMDF

A AMDF (Average Magnitude Difference Function) é uma variação da função de autocorrelação onde o produto das amostras é substituído pelo módulo da diferença, isto é:

$$AMDF_k = \frac{1}{L} \sum_{n=0}^{L-1} |v_n - v_{n+k}| \qquad k = 0, 1, \ldots, k_{máx} \qquad (4.18a)$$

onde $AMDF_k$ é o valor da AMDF para um atraso k, L é escolhido apropriadamente e $\{v_n\}$ é a seqüência de voz. No nosso caso, utilizou-se $L = k_{máx} = N/2 = 100$, onde N é o comprimento do quadro, e eliminou-se a divisão por L, por ser desnecessária. Desta forma, a equação 4.18a pode ser reescrita como

$$AMDF_k = \sum_{n=0}^{99} |v_n - v_{n+k}| \qquad k = 0, 1, \ldots, 100 \qquad (4.18b)$$

Para os sons sonoros, a AMDF apresenta vales acentudados nos atrasos correspondentes ao período de pitch; estes vales não são observados para os sons surdos. A figura 4.21 mostra dois exemplos típicos.

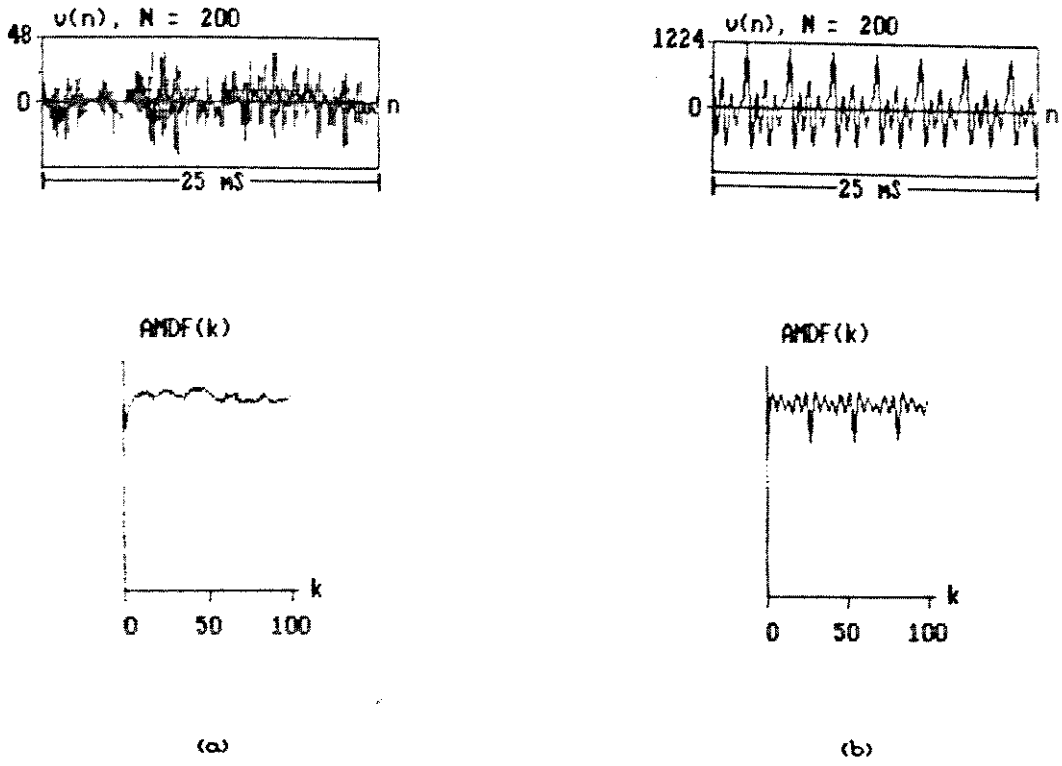


FIG.4.21 - Exemplos típicos da **AMDF**: a) **AMDF** para um quadro do fricativo surdo /ch/; b) idem, para a vogal /a/.

B) DETECTOR DE PITCH

O algoritmo implementado para a detecção de pitch, mostrado esquematicamente na figura 4.22, é uma modificação do algoritmo proposto em [18]. O programa, em Linguagem C, está apresentado no Anexo B. As simplificações aqui introduzidas reduziram (sem, contudo, eliminar) a possiblidade de propagação de erros, verificada no algoritmo original, quando a primeira decisão for incorreta. Uma das modificações introduzidas consistiu em apenas realizar o algoritmo para a extração de pitch quando o resultado da expressão lógica

$$et1/et2)50 \text{ ou } et2/et1)50 \text{ ou } Et(40 \text{ ou } ZRX)70 \text{ ou } (NTP)90 \text{ e } Et(50) \qquad (4.19)$$

for falso. Esta expressão, cujos limiares foram ajustados a

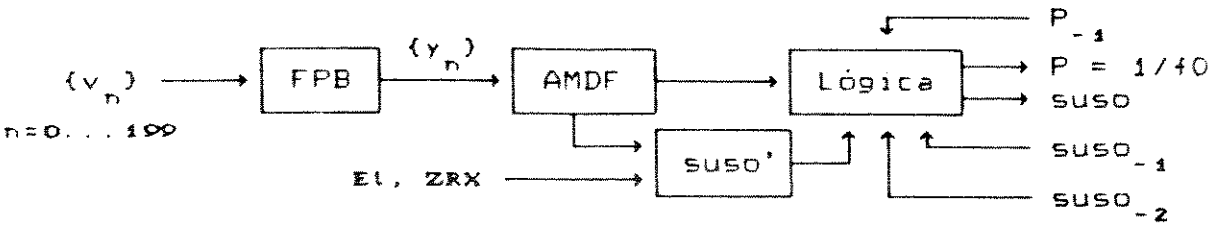


FIG. 4.22 - Detector de pitch. FPB é um filtro passa baixas, a AMDF é calculada pela equação 4.18; suso', suso, suso$_{-1}$ e suso$_{-2}$ são decisões surdo/sonoro; P e P_{-1} representam o pitch do quadro atual e do quadro anterior e a "Lógica" é realizada pelo algoritmo 4.4.

partir dos histogramas apresentados no Apêndice A, detecta com segurança três categorias sonoras que não necessitam ser processadas pelo Detector de Pitch:

a) Coarticulações, que são identificadas por uma alta relação et1/et2 ou et2/et1, como foi discutido na seção 4.2.3A.

b) Silêncio, correspondente aos quadros cuja energia total, Et, é menor que o limiar de 40;

c) Fricativos Surdos, divididos em dois grupos: o primeiro grupo, dos fricativos surdos fortes, como o /ch/, que apresenta um grande número de cruzamentos por zero (ZRX) 70); o segundo grupo, dos fricativos surdos fracos, como o /f/, que apresenta um elevado valor de NTP (número total de picos) e uma energia total abaixo do limiar de 50.

Caso a expressão 4.19 seja falsa, ou seja, é possível que o quadro seja sonoro, procede-se à execução do Algoritmo de Extração de Pitch, conforme a figura 4.22.

Inicialmente, a decisão suso' é determinada por

$$\text{suso'} = \begin{cases} 1, & \underline{\text{se}}\ (\text{ZRX})5\ \underline{\text{e}}\ \text{ZRX}(40)\ \underline{\text{ou}}\ \text{Et})40 \\ 0, & \text{caso contrário} \end{cases} \qquad (4.20)$$

onde ZRX é o número de cruzamentos por zero e Et a energia total. A relação 4.20 atribuiu o valor "1" à variável suso' nos quadros supostamente sonoros, isto é, que apresentam um número moderado

de cruzamentos por zero ou uma alta energia. Em seguida, a seqüência de voz, (v_n), é filtrada por um filtro passa-baixas não recursivo, com um corte em torno de 1 kHz, gerando a seqüência (y_n) dada por:

$$y_n = \begin{cases} \frac{1}{5} \sum_{i=0}^{4} v_{n+i} & 0 \leq n \leq 195 \\ x_n & 195 < n < 200 \end{cases} \qquad (4.21)$$

Esta filtragem reduz a influência das componentes de alta freqüência, dando um aspecto mais suave à AMDF da seqüência (y_n), que é calculada pela equação 4.18b. Após o cálculo da AMDF são determinados quatro parâmetros:

a) max = amplitude máxima da AMDF,

b) min = amplitude mínima da AMDF,

c) minp = posição do mínimo da AMDF, isto é, o provável período de pitch e

d) nrat=max/min.

Estes parâmetros serão utilizados no algoritmo 4.4, responsável pela lógica de detecção de pitch; outros detalhes da implementação são apresentados no Apêndice B.

O algoritmo executa um entre quatro caminhos possíveis, a partir do valor da variável "l", dada por:

$$l = suso' + 2 \cdot suso_{-1} + 4 \cdot suso_{-2} \qquad (4.22)$$

onde suso' é a decisão surdo/sonoro inicial para o quadro em análise, que pode ser alterada posteriormente, e $suso_{-1}$ e $suso_{-2}$ correspondem à decisão surdo/sonoro do último e do penúltimo quadros, respectivamente. O procedimento realizado em cada caminho é descrito a seguir:

a) <u>caminho 1</u> (l = 0, 2 ou 4)

Se o vale da AMDF for pouco profundo, a decisão suso será igual a 0, confirmando suso'; caso contrário, suso será igual a 1, e o período de pitch será dado por minp;

(Observação: as condições iniciais, antes do processamento do primeiro quadro, são: $P_{-1} = suso_{-1} = suso_{-2} = 0$)

```
l = suso' + 2•suso_-1 + 4•suso_-2;
caso [l=0 ou l=2 ou l=4]
      se [nrat<limiar1 ou (nrat≥limiar1 e max<limiar2]
           P = 0;
           suso = 0;
      senão
           P = minp;
           suso = 1;
caso [l=1]
      se [max≤limiar2 ou (max>limiar2 e nrat≤limiar3)]
           P = 0;
           suso = 0;
      senão
           P = minp;
           suso = 1;
caso [l=3 ou l=5 ou l=7]
      P = minp;
      suso = 1;
caso [l=6]
      P = P_-1;
      suso = 0;
se [P > 1,8•P_-1 e P < 2,2•P_-1]
    P = minp/2;
senão se [P > 0,45•P_-1 e P < 0,55•P_-1]
          P = 2•minp;
senão se [P > 2,7•P_-1 e P < 3,3•P_-1]
          P = minp/3;
suso_-1 = suso;   suso_-2 = suso_-1;
```

ALG. 4.4 - Lógica para a determinação do período de pitch, P. Inicialmente, $suso_{-1}$, $suso_{-2}$ e P_{-1} são iguais a zero. Os limiares 1, 2 e 3 foram ajustados em 5, 1200 e 2, respectivamente.

b) caminho 2 (l = 1)

É verificado se o vale da AMDF é profundo o bastante para confirmar a decisão suso', que é igual a 1; não confirmando, o algoritmo faz suso = 0 e P = 0;

c) caminho 3 (l = 3, 5 ou 7)

O algoritmo faz simplesmente pitch = minp e suso = 1;

d) caminho 4 (l = 6)

Neste caso, a decisão suso' é igual a zero, mas o algoritmo estende o período de pitch do quadro anterior para o quadro atual e faz suso = suso' = 0. Esta estratégia procura evitar a ocorrência de um pitch nulo no meio de uma longa seqüência de intervalos sonoros, embora ela possa gerar um erro (aceitável) na transição dos trechos sonoros para os trechos surdos.

Após a determinação do período de pitch, é verificado, finalmente, se ele não corresponde à metade, dobro ou triplo do período de pitch anterior, fazendo-se as devidas correções, quando necessário.

As figuras 4.23a,b apresentam algumas curvas de pitch extraídas com este algoritmo.

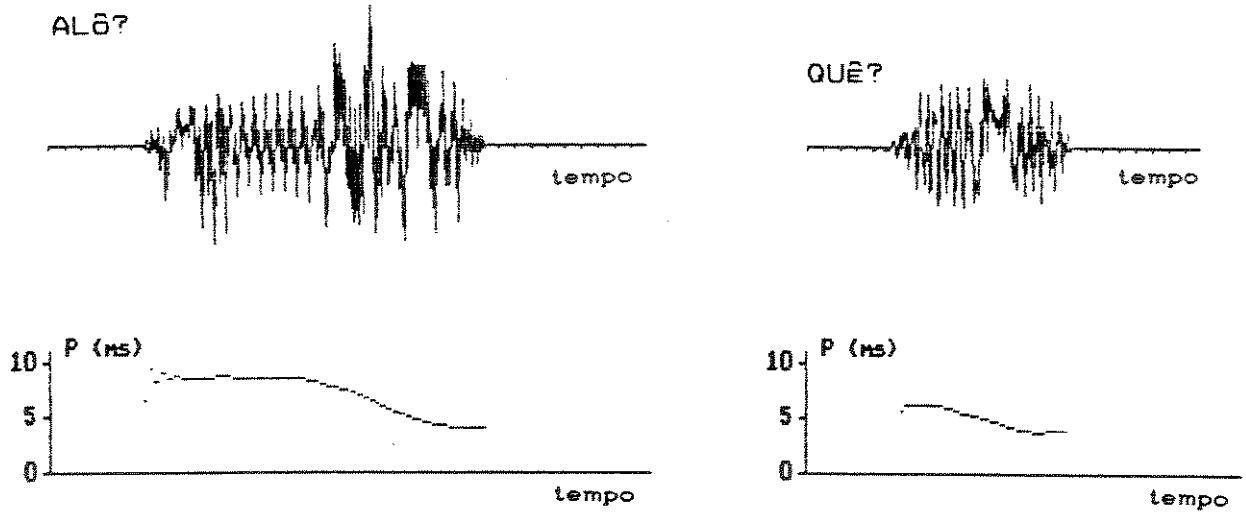


FIG.4.23a - Variação do período de pitch na pronúncia das palavras /alô?/ e /quê?/.

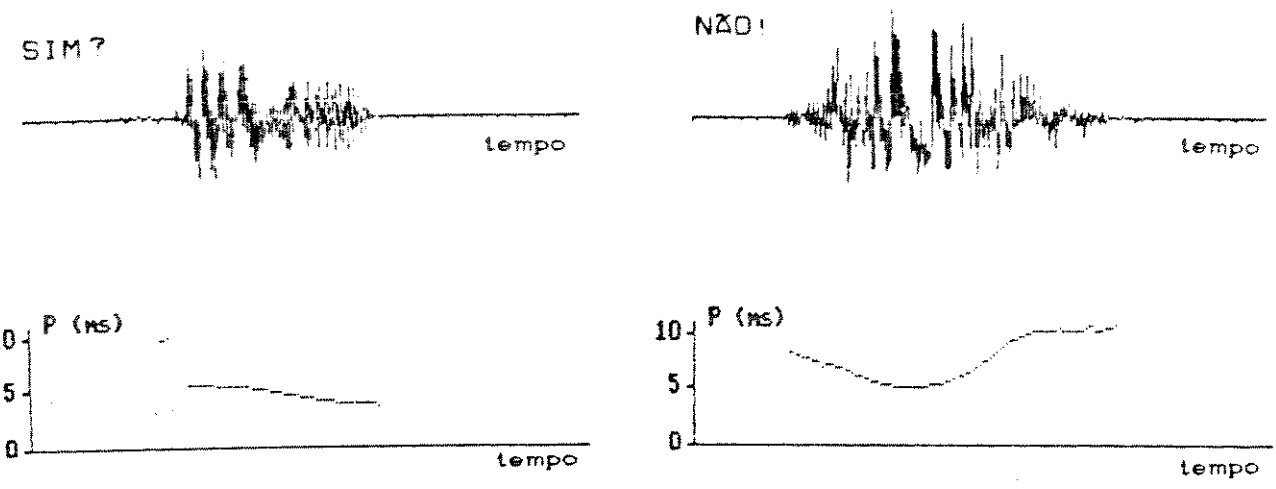


FIG.4.23b - Variação do período de pitch na pronúncia das palavras /sim?/ e /não!/.

4.2.8 CLASSIFICAÇÃO

A classificação do segmento de voz em uma das categorias fonéticas (figura 4.2) é feita pela combinação dos parâmetros temporais e espectrais discutidos nas seções anteriores, por meio de um sistema de regras. As regras foram criadas *heuristicamente* e os limiares de decisão foram ajustados, para um locutor geral, a partir da observação das médias estatísticas dos diversos parâmetros, com o auxílio de histogramas semelhantes aos apresentados no Anexo A. A elaboração das regras, num total de seis (onde cada regra está associada a uma categoria fonética), foi considerada a parte mais delicada do nosso trabalho. A idéia básica foi descrever cada categoria em termos dos diversos parâmetros, como será visto no algoritmo de classificação.

O Algoritmo de Classificação (ALG. 4.5) verifica o resultado da aplicação das 6 regras hierarquicamente, isto é, a regra 1 tem prioridade sobre a regra 2, a regra 2 sobre a regra 3, e assim sucessivamente.

O próximo item da seção é dedicado à descrição de cada uma das regras.

```
regra 1 é aplicável?
 sim: categoria = silêncio;
 não: regra 2 é aplicável?
       sim: categoria = coarticulação;
       não: regra 3 é aplicável?
             sim: categoria = fricativo surdo;
             não: regra 4 é aplicável?
                   sim: categoria = oclusão sonora;
                   não: regra 5 é aplicável?
                         sim: categoria = fricativo sonoro;
                         não: regra 6 é aplicável?
                               sim: categoria = vocálico;
                               não: categoria = indefinido;
                                    fim de análise;
```

ALG. 4.5 - Classificação do segmento em uma das categorias fonéticas. As expressões lógicas que definem as regras 1 a 6 são apresentadas no texto.

A) REGRAS

Regra 1 (silêncio)

se { [LogEne(0)(29]
ou
[12(NTP(50 e 13(ZRX(40 e ΔPN(25 e Et(45] } ⟶ silêncio

O limiar de 29 dB na primeira parte da regra está cerca de 3 dB acima do ruído ambiente. A segunda parte da regra é mais complexa e procura detectar o silêncio no interior das palavras (como na pronúncia das oclusivas surdas /p/ e /t/, em /apostar/), assim como o silêncio entre a pronúncia de duas palavras.

Regra 2 (coarticulação)

se { (et1/et2)50) ou (et2/et1)50) } ⟶ coarticulação

Esta regra verifica simplesmente se a relação entre a

energia total de cada meio-quadro (12,5 ms) é maior que um limiar. É utilizada para a localização de transições entre fonemas, principalmente após a explosão de oclusivas (/p/, /b/, /t/, /d/, /g/, /k/)

Regra 3 (fricativo surdo)

se { [LogEne(1)(35 e (ZRX)25 ou NTP)40) e Et(55 e f0=0]
ou
[Et(60 e (ZRX)55 ou NTP)75) e |ΔPN|(20 e f0=0] } ⟶ fricativo surdo

Esta regra baseia-se fundamentalmente na decisão surdo/sonoro igual a zero (f0 = 0). Outros aspectos importantes são o baixo valor de Et, e uma elevada contagem de cruzamentos por zero (ZRX) ou do número total de picos (NTP). A regra foi dividia em duas partes. A primeira parte foi ajustada para a detecção de fricativos surdos fracos (/f/, principalmente), enquanto a segunda parte reconhece os fricativos surdos com maior energia (como o /ch/).

Regra 4 (oclusão sonora)

se { LogEne(0))30 e LogEne(0)(40 e LogEne(3)(21 e
(bw1(300 ou f0≠0) e (3(ZRX(12) e NTP(35 } ⟶ oclusão sonora

Esta regra detecta os instantes de oclusão das explosivas sonoras (/b/, /b/, /g/), onde há vibração laríngea. As consoantes /m/ e /n/, no final de palavras, também são incluídas nesta categoria. O parâmetro LogEne(3), associado à faixa de 1000-3000 Hz, é fundamental na distinção entre esta categoria e a categoria dos fricativos sonoros, uma vez que as oclusões sonoras têm pouca energia nesta faixa, enquanto os fricativos sonoros, em virtude da excitação turbulenta, têm uma energia mais acentuada. O parâmetro bw1 (largura de faixa do primeiro formante) também é utilizado para evitar confusões com os fricativos sonoros, que também apresentam uma ressonância nas baixas freqüências devido à

excitação glotal. Neste caso, as oclusões sonoras são identificadas por apresentar uma ressonância bastante aguda

Regra 5 (fricativo sonoro)

$$\underline{se}\ \left\{ Et<66\ \underline{e}\ ZRX<15\ \underline{e}\ (\Delta PN>ZRX\ \underline{ou}\ DPN>25)\ \underline{e}\ NTP>15\ \underline{e}\ f0\neq 0\ \underline{e}\ LogEne(2)<32 \right\} \rightarrowtail \text{fricativo sonoro}$$

A detecção de fricativos sonoros (/v/, /j/, /z/) está baseada principalmente num alto valor do parâmetro DPN. Isto permite uma boa distinção entre fricativos sonoros, oclusões sonoras e vocálicos. Note-se também a dependência com a decisão surdo/sonoro e com um pequeno número de cruzamentos por zero. A imposição LogEne(2) < 32, onde LogEne(2) é a energia de 750-2000 Hz, evita que os fricativos sonoros sejam confundidos com vogais /i/ de pequena intensidade, pois as vogais apresentam uma energia maior nesta faixa.

Regra 6 (vocálico)

$$\underline{se}\ \left\{ f0\neq 0\ \underline{e}\ Et>60 \right\} \rightarrowtail \text{vocálico}$$

A categoria detectada por esta regra inclui as vogais, as consoantes /m/ e /n/, quando próximas a vogais fortes, e as laterais, como o /l/. A regra verifica simplesmente se o quadro é sonoro e sua energia está acima de um limiar. Os vocálicos são relativamente simples de serem identificados, uma vez que toda a informação necessária está contida na faixa de 0-4 kHz.

B) EXEMPLO

Na figura 4.24 é apresentado um exemplo de classificação para a expressão /digitalização de voz/, pronunciada por um homem adulto. Lembre-se que é realizada uma classificação a cada 5 ms e que esta classificação é o resultado da análise de um quadro de 25 ms.

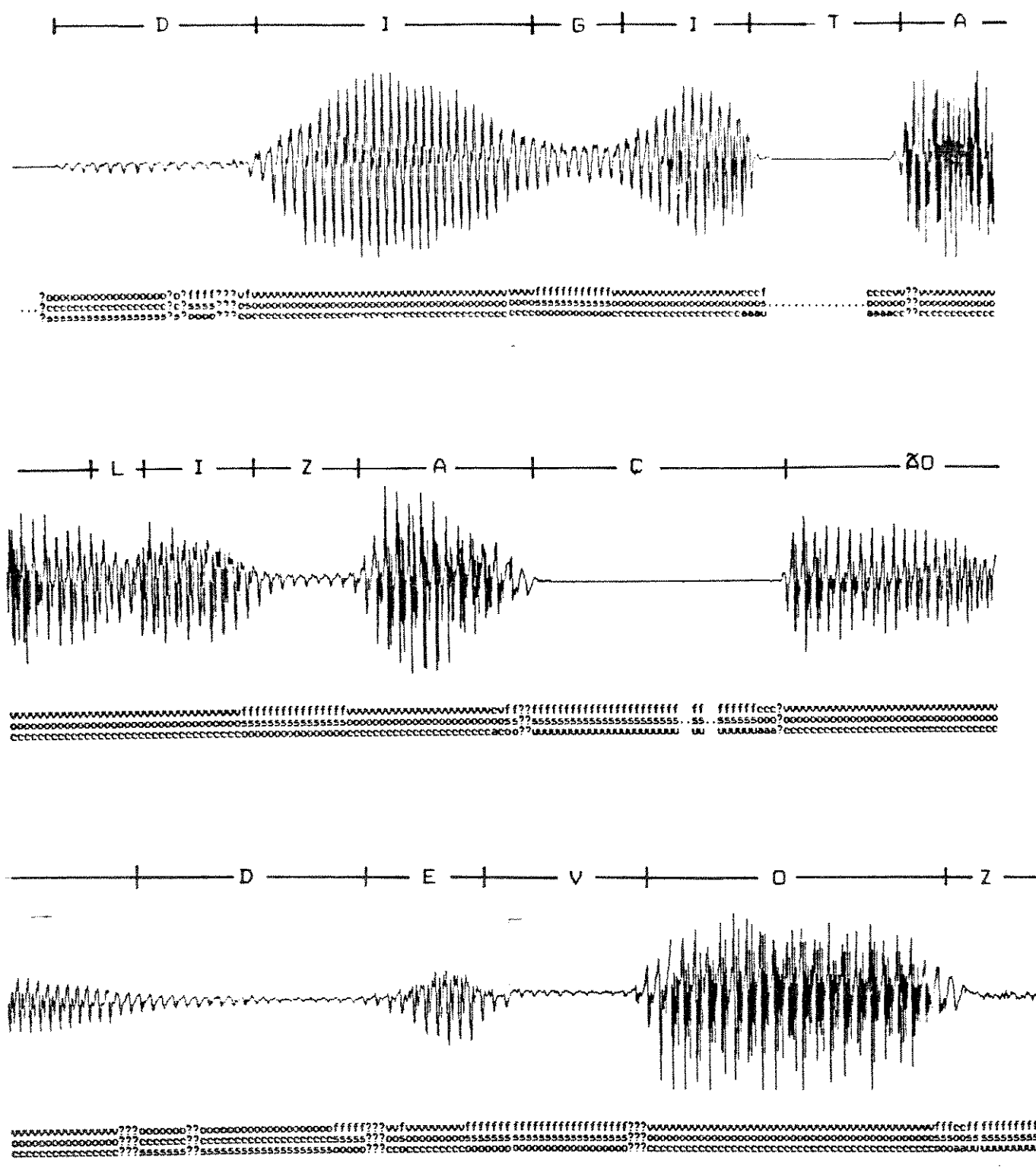


FIG.4.24 - Exemplo de classificação automática, onde o ponto (.) indica silêncio, coa = coarticulação, fsu = fricativo surdo, ocs = oclusão sonora, fso = fricativo sonoro, voc = vocálico e ??? = indefinido.

C) AVALIAÇÃO ESTATÍSTICA DE DESEMPENHO

A realização de uma estatistica não tendenciosa é uma tarefa bastante difícil na avaliação de um Sistema de Reconhecimento de Voz. Os índices de desempenho variam com o vocabulário utilizado na avaliação, o tipo de pronúncia (palavras isoladas, conectadas ou fala natural), e o ruído, dentre outros fatores.

O Quadro 4.1 representa o comportamento estatistico do Algoritmo 4.5 (algoritmo de classificação) na pronúncia das palavras /dezessete/ e /dezoito/ faladas por um primeiro locutor, e das expressões /digitalização de voz/ e /voz digital/ faladas por um segundo locutor. Ambos os locutores são adultos do sexo masculino. Para a realização deste teste, inicialmente foram obtidas as classificações automáticas de cada uma das elocuções,

(entradas) \ (classificações)	SILÊNCIO	FRICATIVO SURDO	OCLUSÃO SONORA	FRICATIVO SONORO	VOCÁLICO	COARTICULAÇÃO	INDEFINIDO
SILÊNCIO	**91,07** (204)	**5,36** (12)	**0,89** (2)				**2,68** (6)
FRICATIVO SURDO	**9,16** (12)	**90,84** (119)					
OCLUSÃO SONORA			**88,49** (123)	**5,04** (7)	**2,16** (3)		**4,32** (6)
FRICATIVO SONORO			**2,44** (4)	**90,85** (149)	**1,83** (3)		**4,88** (8)
VOCÁLICO			**0,75** (6)	**5,49** (44)	**91,39** (732)	**0,25** (2)	**2,12** (17)
COARTICULAÇÃO				**4,65** (2)		**81,40** (35)	**13,95** (6)

QUADRO 4.1 - Avaliação de desempenho do classificador. Os números em negrito são valores percentuais e os números entre parênteses são as contagens. Foram considerados apenas os intervalos de silêncio ocorridos dentro das palavras ou próximos ao início ou fim da elocução.

conforme o exemplo da figura 4.24. Como as elocuções eram conhecidas, sabia-se, a priori, a classificação correta de cada uma das entradas. Desta forma, os resultados obtidos com o algoritmo de classificação foram comparados com os resultados corretos ou desejados, montando-se o Quadro 4.1. Resultados semelhantes foram alcançados com vozes femininas. Não foram realizados testes com crianças. Pelo menos entre vozes de pessoas adultas, os resultados mostraram-se independentes do locutor.

No cálculo das estatísticas da categoria "silêncio", foram considerados apenas os quadros localizados no interior e próximos ao início e fim das elocuções. Note-se que o pior índice corresponde à categoria das coarticulações. Outros comentários sobre o Quadro 4.1, assim como a discussão dos demais resultados e as conclusões deste trabalho ficam reservadas para o capítulo final.

4.3 REFERÊNCIAS

[1] F. Violaro, *"Nova Versão do Sistema de Análise e Processamento Digital de Voz: SAPDV-A"*. Anais do 7º Simpósio Brasileiro de Telecomunicações, Florianópolis - S.C. (1989);

[2] L. R. Rabiner & R. W. Schafer, *"Digital Processing of Speech Signals"*, Prentice Hall, Inc., NJ (1978);

[3] J. D. Markel & A. H. Gray, Jr., *"Linear Prediction of Speech"*, Springer-Verlag, NY (1976);

[4] R. W. Schafer & L. R. Rabiner, *"System for Automatic Formant Analysis of Voiced Speech"*, The Journal of the Acoustical Society of America, vol. 47, pp 634-648 (1970);

[5] J .D. Markel, *"Digital Inverse Filtering - A New Tool for Formant Trajectory Estimation"*, IEEE Transactions on Audio and Electroacoustics, vol 20, pp 129-137 (1972);

[6] J. Makhoul, *"Linear Prediction: A Tutorial Review"*, Proceedings of the IEEE, vol 63, pp 561-580 (1975);

[7] J. D. Markel, *"FFT Pruning"*, IEEE Transactions on Audio and Electroacoustics, vol 19 (4), pp 305-311 (1971);

[8] D P. Skinner, "Pruning the Decimation In-Time FFT Algorithm", IEEE Transactions on Acoustic, Speech and Signal Processing, vol 24 (2), pp 193-194 (1976),

[9] K. N. Stevens, "Acoustic Correlates of Some Phonetic Categories", The Journal of The Acoustical Society of America, vol 68 (3), pp 836-842 (1980),

[10] Kung-Pu Li et al, "Segment Classification in Continuous Speech", IEEE Transactions on Audio and Electroacoustics, vol 21 (1), pp 50-57 (1973);

[11] C. J. Weinstein et al, "A System for Acoustic-Phonetic Analysis of Continuous Speech", IEEE Transactions on Acoustics, Speech and Signal Processing, vol 23, pp 54-67 (1975);

[12] R. W. Becher & F. Poza, "Acoustic Phonetic Research in Speech Understanding", IEEE Transactions on Acoustics, Speech and Signal Processing, vol 23 (5) pp 416-426 (1975);

[13] P. Regel, "A Module for Acoustic-Phonetic Transcription of Fluently Spoken German Speech", IEEE Transactions on Acoustics, Speech and Signal Processing, vol 30 (3), pp 440-450 (1982);

[14] J. M. Heinz & K. N. Stevens, "On The Properties of Voiceless Fricative Consonants", The Journal of The Acoustical Society of America, vol 33, pp 589-596 (1961);

[15] S. S. McCandless, "An Algorithm for Automatic Formant Extraction Using Linear Prediction Spectra", IEEE Transactions on Acoustics, Speech and Signal Processing, vol 22, pp 135-141 (1974);

[16] S. S. McCandless, "Modifications to Formant Tracking Algorithm of April 1974", IEEE Transactions on Acoustics, Speech and Signal Processing, vol 24, pp 192-193 (1976);

[17] J. Gandour & B. Weinberg, "On the Relationship Between Vowel Height and Fundamental Frequency: Evidence from Esophageal Speech", Phonetica, vol 37 (5-6) pp 344-354 (1980);

[18] M. J. Ross et al, "*Average Magnitude Difference Function Pitch Extractor*", IEEE Transactions on Acoustics, Speech and Signal Processing, vol 22, pp 129-137 (1972);

[19] L. R. Rabiner et al, "*A Comparative Performance Study of Several Pitch Detection Algorithms*", IEEE Transactions on Acoustics, Speech and Signal Processing, vol 24, pp 399-417 (1976);

[20] R. B. Monsen & A. M. Engebretson, "*The Accuracy of Formant Frequency Measurements: A Comparison of Spectrographic Analysis and Linear Prediction*", Journal of Speech and Hearing Research, vol 26, pp 89-97 (1983);

[21] A. V. Oppenheim & R. W. Schafer, "*Digital Signal Processing*", Prentice Hall, N. J. (1975);

[22] S. Haykin, "*Adaptive Filter Theory*", Prentice-Hall, N. J. (1986);

[23] J. E. Shoup & L. L. Pfeifer, "*Acoustic Caracteristics of Speech Sound*", capitulo 4 de "*Contemporary Issues in Experimental Phonetics*", N. J. Lass, ed., Academic Press, NY (1976);

[24] J. F. Brandt, "*Perceptual Psychophysics: Speech and Hearing*", capitulo 13 de "*Contemporary Issues in Experimental Phonetics*", N. J. Lass, ed., Academic Press, NY (1976);

[25] J. L. Flanagan, "*Signal Analysis in the Auditory System*", capitulo 6 de "*Human Communication: A Unified View*", E. E. David, Jr., & P. B. Denes, McGraw-Hill (1972);

[26] M. Shigenaga et al, "*A Speech Recognition System for Continuously Spoken Japanes Sentences - SPEECH YAMANASHI*", The Transactions of The IECE of Japan, vol 69 (5), pp 675-683 (1986);

[27] F. R. Chen, "*Lexical Access and Verification in a Broad Phonetic Approach to Continuous Digit Recognition*", Proceedings of the International Conference on Acoustics, Speech and Signal Processing, pp 1089-1092, Tóquio, Japão (1986);

[28] D. P. Huttenlocher, "*A Broad Phonetic Classifier*", Proceedings of the International Conference on Acoustics, Speech and Signal Processing, pp 2259-2262, Tóquio, Japão (1986);

[29] L. Fissore et al, "*A Word Hypothesizer for a Large Vocabulary Continuous Speech Understanding System*", Proceedings of the International Conference on Acoustics, Speech and Signal Processing, pp 453-456 (1989);

[30] V. Zue et al, "*Acoustic Segmentation and Phonetic Classification in the Summit System*", Proceedings of the International Conference on Acoustic, Speech and Signal Processing, pp 389-392 (1989).

CAPÍTULO 5

DISCUSSÃO FINAL

5.1 REVISÃO

O primeiro capítulo desta dissertação tratou, superficialmente, das várias possibilidades de uso prático das informações transportadas pelo sinal de voz.

No capítulo 2, referente aos Sistemas de Reconhecimento Automático de Voz (SRAV), mostrou-se que o bom desempenho dos SRAV depende, basicamente, de dois fatores:

a) de um bom modelamento da produção do sinal;

b) da aplicação de diversas fontes de conhecimento lingüístico, como a sintaxe, a fonética ou a prosódia, na tentativa de explorar as particularidades de cada idioma.

A caracterização dos sons da fala e o modelamento da produção do sinal de voz foram feitos no Capítulo 3. Neste capítulo também foi destacada a importância da realização de transformações sobre o sinal, com o intuito de simular os processamentos auditivos e perceptuais que ocorrem na decodificação da fala. Neste sentido, foi introduzido o conceito de Banda Crítica; o uso de bandas críticas, ou de outras transformações equivalentes, é um dos assuntos de maior interesse nas atuais pesquisas em reconhecimento de voz e pode ser colocado como um terceiro fator para a melhoria no desempenho dos SRAV. Este tópico será discutido novamente na seção 5.3.2.

No 4º capítulo foram apresentados os aspectos práticos do desenvolvimento do software para o Módulo Frontal de um Sistema de Reconhecimento Automático de Voz

Neste 5º e último capítulo serão avaliados os principais métodos e resultados do capítulo 4, assim como as perspectivas de utilização do nosso Módulo Frontal para o reconhecimento de grandes vocabulários, de forma independente do locutor.

5.2 AVALIAÇÃO DOS RESULTADOS EXPERIMENTAIS

5.2.1 ESTIMATIVA ESPECTRAL

A escolha da técnica a ser empregada na estimativa da resposta em freqüência discreta do aparelho fonador, $\tilde{F}(e^{j\omega k})$, resultou da comparação entre as duas formas básicas de análise espectral descritas na literatura: a técnica de Deconvolução Homomórfica e a técnica LPC. Optou-se pela Predição Linear, tendo em vista a simplicidade computacional e a suavidade dos espectros resultantes, como já foi dito no capítulo 4. A preocupação em obter-se um espectro suave é justificada pela possibilidade de se estimar a freqüência dos formantes a partir da simples localização dos picos de $\tilde{F}(e^{j\omega k})$. A título de ilustração, a figura 5.1 apresenta uma comparação entre a estimativa da resposta em freqüência do trato vocal obtida com Predição Linear e com Deconvolução Homomórfica. O processamento Homomórfico foi

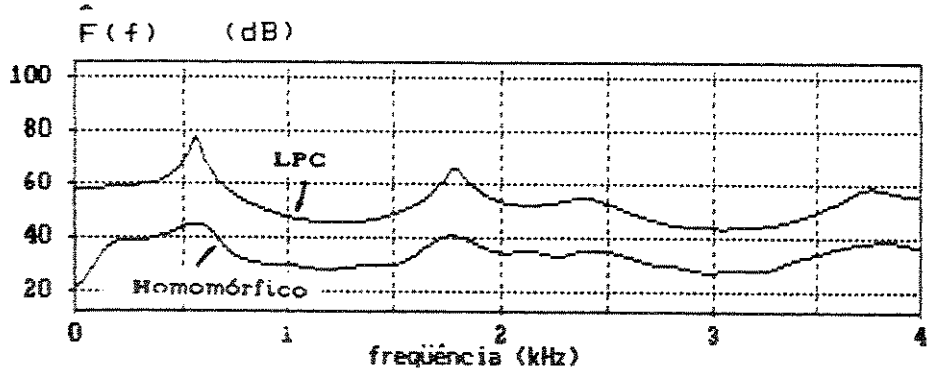


FIG.5.1 - Comparação entre a análise LPC e a Deconvolução Homomórfica.

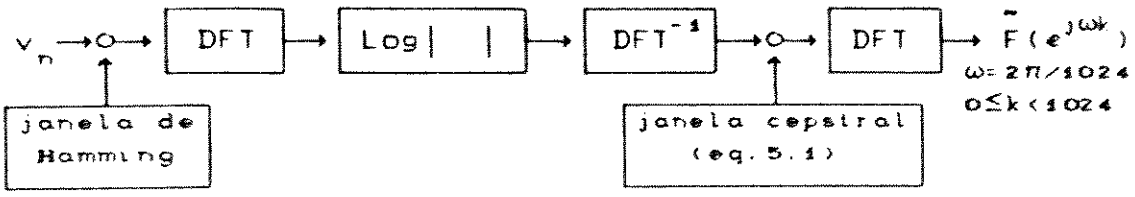


FIG. 5.2 - Sistema de Deconvolução Homomórfica utilizado para o exemplo da figura 5.1. DFT^{-1} é a Transformada de Fourier Discreta Inversa.

realizado conforme a figura 5.2, calculando-se as DFT's e a DFT^{-1} com 1024 pontos e utilizando-se uma janela cepstral (filtro passa baixos tempos) dada por

$$\ell(n) = \begin{cases} 1, & |n| \leq \tau \\ 0, & \text{caso contrário} \end{cases} \qquad (5.1)$$

O valor de τ foi ajustado em 35 amostras (4,4 ms).

A dificuldade do modelamento preciso dos vales espectrais com a análise LPC pode ser visualizada na figura 5.3. Note-se que a envoltória não acompanha com fidelidade os vales localizados em torno de 2,4 kHz e de 3,6 kHz. Deve ser ressaltado, entretanto, que esta deficiência da análise LPC não influenciou os nossos resultados, uma vez que não foi realizada a detecção do traço de nasalidade.

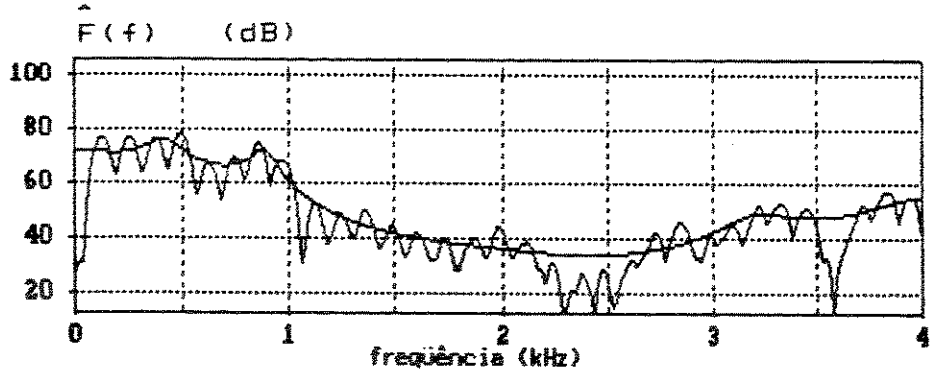


FIG. 5.3 - Dificuldade de modelamento de zeros espectrais

5.2.2 EXTRAÇÃO DE FORMANTES

A estimativa da freqüência dos formantes a partir da aproximação parabólica para os picos da resposta em freqüência do trato vocal mostrou-se eficiente em cerca de 90% dos casos. A detecção de picos foi ineficaz em situações como a indicada na figura 5.4. Neste caso, há uma ressonância em torno de 1100 Hz, que não será detectada. A detecção de ressonâncias deste tipo pode ser efetuada, por exemplo [1], pela localização dos picos da derivada segunda do espectro de amplitude ou pela solução da equação característica do preditor,

$$\sum_{i=0}^{12} \alpha_i \cdot z^{-i} = 0, \qquad \alpha_o = 1 \tag{5.2}$$

As raízes desta equação correspondem aos pólos de $\tilde{F}(e^{j\omega})$. Uma análise destas raízes permite separar os pólos da excitação dos pólos do trato vocal. Segundo [2], os pólos associados à excitação ou pertencem ao eixo real (não se tratando, portanto, de uma freqüência de ressonância), ou produzem um pico de pequena amplitude. Neste último caso, pode-se calcular a amplitude dos picos associados a cada par de pólos complexos conjugados e desprezar-se os picos com amplitude abaixo de um certo limiar.

No nosso trabalho não foi verificado o problema da fusão de dois picos espectrais muito próximos em um único pico de faixa

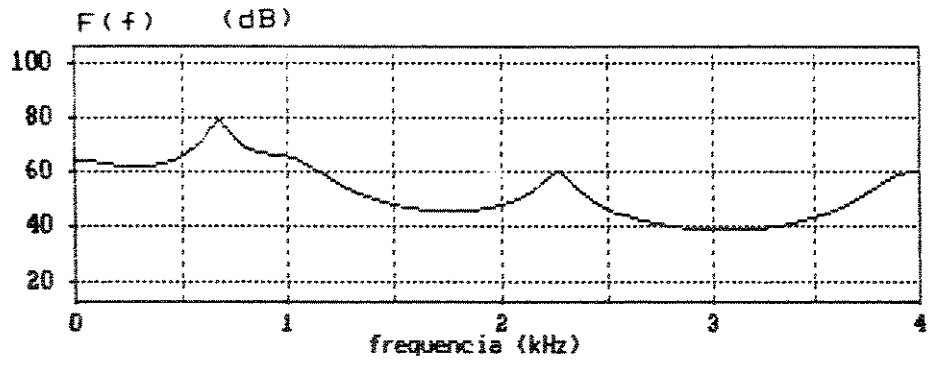


FIG.5.4 - Erro na localização de picos. Há um formante em torno de 1100 Hz, cujo pico não aparece no espectro.

larga. Este problema não aconteceu tendo em vista a escolha adequada da ordem do preditor (M = 12) e a alta resolução espectral utilizada no cálculo da resposta em frequência do filtro inverso (7,8 Hz).

Não foi realizado nenhum estudo sobre a precisão das estimativas da largura de faixa dos formantes. Estudos deste tipo podem ser feitos analisando-se formas de ondas geradas artificialmente, como no exemplo da figura 3.14.

5.2.3 DETECÇÃO DE PITCH

A elaboração do Detector de Pitch foi considerada uma das etapas mais críticas do trabalho. Note-se que este é o único algoritmo a utilizar informações de quadros anteriores para tomar uma decisão relativa a um determinado quadro. Embora ainda exista a possibilidade de propagação de erros, este tipo de problema foi significativamente reduzido em relação à versão original do detector (referência [18] do Capítulo 4). O algoritmo utilizado para a detecção de pitch, descrito no capítulo anterior e apresentado no apêndice B, mostrou-se satisfatório tanto em vozes masculinas como femininas. No primeiro caso, onde normalmente ocorre apenas um vale na AMDF, o algoritmo não encontrou dificuldades em detectar o vale e determinar o período de pitch correspondente. No segundo caso (em vozes femininas), geralmente existem dois ou três vales na AMDF e nem sempre o primeiro vale (correspondente ao período de pitch), é o mais profundo. O algoritmo possui uma estratégia para identificar os casos em que o vale mais profundo corresponde à metade, dobro ou triplo do valor correto[1].

Um outro aspecto a destacar no Detector de Pitch é o teste inicial (equação 4.19), que identifica os fricativos surdos, intervalos de silêncio ou coarticulações, não submetendo-os às demais análises do algoritmo. Esta estratégia, e

---

[1] O algoritmo não prevê o caso de o vale da AMDF corresponder a 1/3 do pitch correto porque esta condição não foi verificada em nossos testes.

em particular a detecção de coarticulações, também pode ser empregada em "vocoders". Temos observado, em vários testes de avaliação subjetiva, que é comum a degradação da qualidade da fala sintetizada, na pronúncia de sílabas do tipo /consoante oclusiva + vogal/, como /pa/, /bé/, ou /gá/. Esta degradação, semelhante a um soluço, certamente é causada por uma falha no Detector de Pitch após a explosão da consoante. Ocorre que no início da pronúncia da vogal (nos instantes seguintes à explosão), o movimento de vibração glotal é ainda bastante irregular, induzindo erros grosseiros nos detectores de pitch.

5.2.4 ALGORITMO DE CLASSIFICAÇÃO

A grande dificuldade encontrada no desenvolvimento do Algoritmo de Classificação esteve no problema da *representação do conhecimento*. Constatávamos que a "simples" análise visual das formas de onda e dos espectros era suficiente para a correta classificação dos quadros na quase totalidade dos casos. Entretanto, as primeiras tentativas de realização de uma classificação automática revelavam a complexidade do problema e a necessidade de uma análise conjunta de vários aspectos do sinal de voz. Esta análise foi realizada no sistema de regras, cuja versão final constitui o Algoritmo de Classificação, apresentado no capítulo anterior.

A avaliação estatística do algoritmo, resumida no Quadro 4.1, revela bons percentuais de acerto em todas as categorias. As principais fontes de erro estão localizadas na transição entre categorias, destacando-se a confusão entre os fricativos surdos fracos (como o /s/ no final de palavras) e o silêncio, e entre os fricativos sonoros e as vogais com pequena amplitude; os fricativos sonoros também foram confundidos com as oclusões sonoras. Deve ser ressaltado, entretanto, que não foi realizada nenhuma estratégia para a correção dos erros do Algoritmo de Classificação. Um processamento desta natureza reduziria significativamente o percentual de erros.

5.3 TRABALHOS POSTERIORES

Como encerramento desta dissertação, serão sugeridos diversos trabalhos de aprimoramento e aplicação dos resultados aqui obtidos. Uma característica indispensável nos próximos trabalhos é a utilização da dinâmica da fala, não restringindo as análises a apenas um quadro, como foi realizado no nosso trabalho. Também devem ser empregadas técnicas de Inteligência Artificial e modelos probabilísticos usando Cadeias Ocultas de Markov tanto num melhoramento do Módulo Frontal quanto no desenvolvimento de um Módulo Lingüítico.

5.3.1 SUBCLASSIFICAÇÃO DAS CATEGORIAS

A tentativa de extrair informações fonéticas detalhadas, como a classificação a nível de vogais, deve ser realizada com bastante cuidado, tendo em vista dois fatores principais:

a) a *necessidade* de uma subclassificação fonética minuciosa;

b) os *riscos* desta subclassificação detalhada.

Como exemplo de subclassificação, observando-se esses dois fatores, poder-se-ia realizar uma divisão da categoria dos *vocálicos* conforme o esquema abaixo:

- vocálicos
  - consoantes nasais
  - consoantes laterais
  - vogais e ditongos orais
    - vogais altas (/i/, /ê/)
    - vogais médias (/a/, /ó/)
    - vogais baixas (/ô/, /u/)
  - vogais nasalizadas

No nosso trabalho chegamos a realizar um pequeno estudo em reconhecimento de vogais orais, implementando um algoritmo para a classificação dos quadros correspondentes às vogais em sílabas do tipo /fricativo surdo + vogal/, como /fá/, /ché/, /si/, etc. Verificou-se que o reconhecimento de vogais, independente do locutor, é bastante complexo, uma vez que a posição dos formantes se altera bastante entre diversos

locutores. Além disso, na pronúncia natural das palavras, a configuração dos formantes é influenciada pelos efeitos da coarticulação entre fonemas e dificilmente atinge as mesmas posições que são verificadas em vogais pronunciadas sustentadamente.

Uma proposta para o reconhecimento de vogais, dentro da idéia do emprego de categorias fonéticas amplas, seria restringir a classificação ao nível de vogais altas, médias e baixas, num sistema de regras baseado, por exemplo, nos parâmetros

$$\Delta_1 = F1 - f0$$
$$\Delta_2 = F2 - F1$$
$$\Delta_3 = F3 - F2$$

onde f0 é a freqüência fundamental (em barks) e F1, F2 e F3 são os três primeiros formantes, (em barks). Tem sido demonstrado [5] que medidas em barks, semelhantes aos parâmetros Δ_1, Δ_2 e Δ_3 mantêm-se dentro de faixas constantes, de forma independente do locutor. Deve ser ressaltado, contudo, que essas diferenças se alteram entre os idiomas devendo ser estabelecidas as relações válidas para o idioma Português.

5.3.2 ACOMPANHAMENTO DE FORMANTES

Um algoritmo eficiente, que resolva os problemas de descontinuidade nas estimativas de formantes é uma ferramenta indispensável à análise de voz. Como exemplo, o movimento dos formantes traz consigo informações importantíssimas a respeito das consoantes oclusivas que antecedem uma vogal. Isto pode ser visto na figura 5.5, para o caso da vogal /a/. Note-se a trajetória do segundo formante possibilita a separação entre as consoantes oclusivas bilabiais (/p/ e /b/) e as demais consoantes. De forma semelhante, com a trajetória do terceiro formante, pode-se separar as oclusivas velares (/k/ e /g/) das alveolares (/t/ e /d/). Portanto, dispondo-se de um bom algoritmo para o acompanhamento de formantes também serão possíveis trabalhos em reconhecimento de consoantes oclusivas [6].

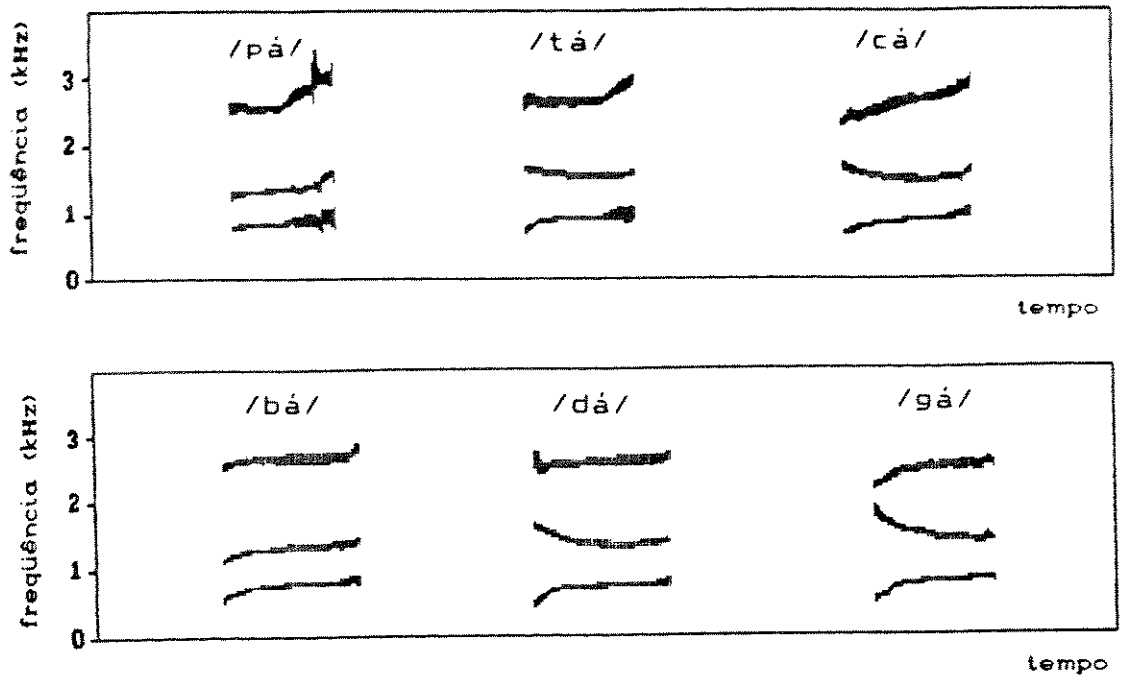


FIG.5.5 - Trajetórias de formantes nas sílabas /pá/, /tá/, /cá/, /bá/, /dá/, /gá/.

5.3.3 RECONHECIMENTO DE GRANDES VOCABULÁRIOS, COM PRONÚNCIA ISOLADA

Inicialmente, as classificações fonéticas realizadas pelo Módulo Frontal deverão ser processadas por um algoritmo que agrupe os quadros adjacentes similares em um único segmento. A representação da palavra /seis/, por exemplo, deverá ser dada pela seqüência

[fricativo surdo] [vocálico] [fricativo surdo].
/s/ /ei/ /s/

Note-se que as palavras /fás/, /fis/, /fês/, /seus/, e /suas/, terão a mesma representação.

Um *grande* vocabulário pode ser estruturado em *pequenos* subconjuntos de palavras que possuem a mesma representação. Estudos em vários idiomas [7], [8], [9], mostram que o tamanho médio de cada um desses subconjuntos é de 15 a 35 palavras, para vocabulários da ordem de 10.000 a 20.000 palavras. Desta forma, com uma análise fonética grosseira - mas confiável - é possível

uma redução significativa no número de palavras a serem verificadas. Observe-se que a análise fonética detalhada somente será realizada para a dicriminação entre palavras de um mesmo subconjunto. A classificação em categorias fonéticas também pode ser combinada com a classificação segundo a acentuação ortográfica (dividindo-se as palavras em oxítonas, paroxítonas ou proparoxítonas) com o objetivo de reduzir o tamanho dos subconjuntos.

5.4 REFERÊNCIAS

[1] L. G. Meloni, "*Algoritmos Rápidos de Extração de Formantes a Partir do Modelo Auto-Regressivo*", Anais do 7º Simpósio Brasileiro de Telecomunicações, pp 36-43, Florianópolis, S.C. (1989);

[2] G. L. Campos, "*Síntese de Voz para o Idioma Português*", Tese de Doutorado - USP (1980);

[3] A. V. Oppeheim & R. W. Schafer, "*Digital Signal Processing*", Prentice Hall (1975);

[4] L. R. Rabiner & R. W. Shafer, "*Digital Processing of Speech Signals*", Prentice Hall, NJ (1978);

[5] A. K. Syrdal & H. S. Gopal, "*A Perceptual Model of Vowel Recognition Based on The Auditory Representation of American English Vowels*", The Journal of The Acoustical Society of America, vol 79 (4), pp 1086-1100 (1986);

[6] P. Demichelis et al, "*Computer Recognition of Plosive Sounds Using Contextual Information*", IEEE Transactions on Acoustics, Speech and Signal Processing, vol 31 (2), pp 359-377 (1983);

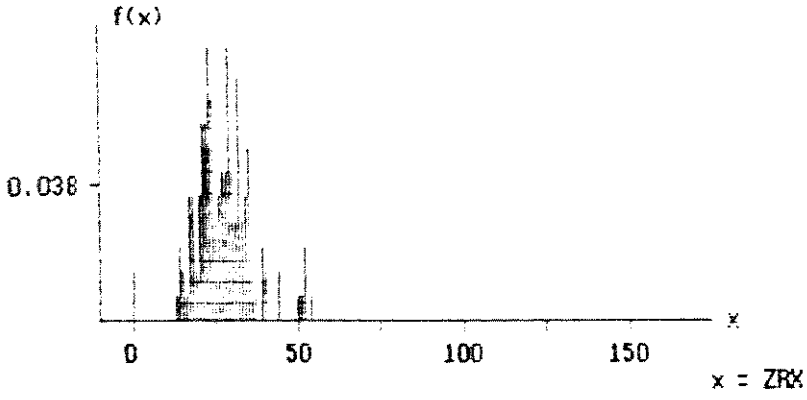
[7] V. Zue, "*The Use of Speech Knowlegde in Speech Recognition*", Proceedings of the IEEE, vol 73 (11), pp 1602-1615 (1985).

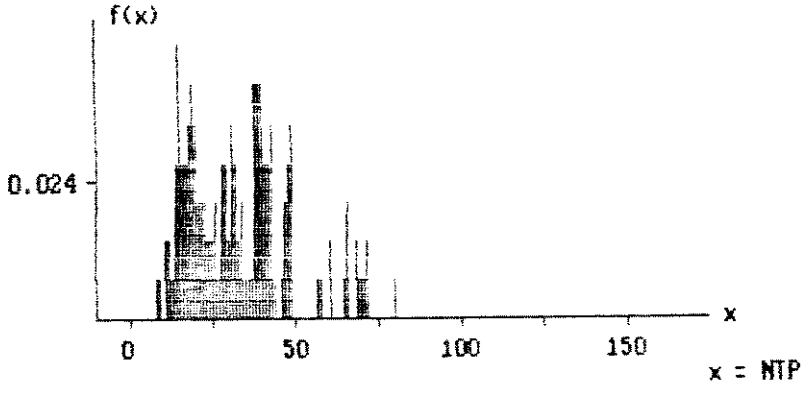
[8] J. N. Larar, "*Lexical Access Using Broad Acoustic-Phonetic Classifications*", Computer, Speech and Language, vol 1 (1), pp 47-59 (1986);

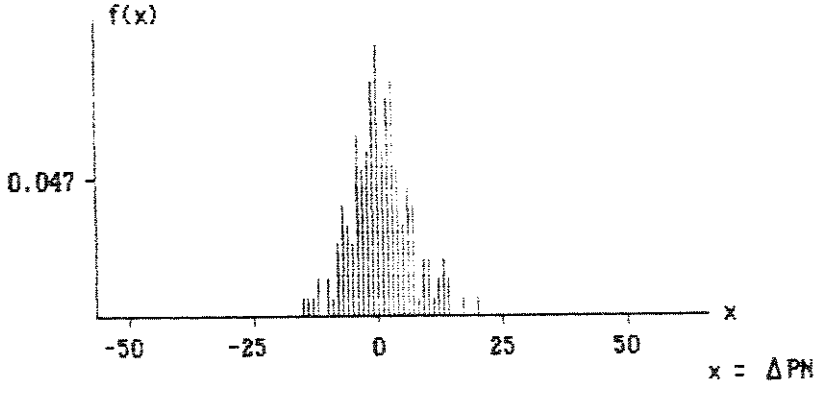
[9] G J Vernooij et al, "A Simulation Study on The Usefulness of Broad Phonetic Classification in Automatic Speech Recogniton", Proceedings of The International Conference on Acoustics, Speech and Signal Processing, pp 85-89 (1989).

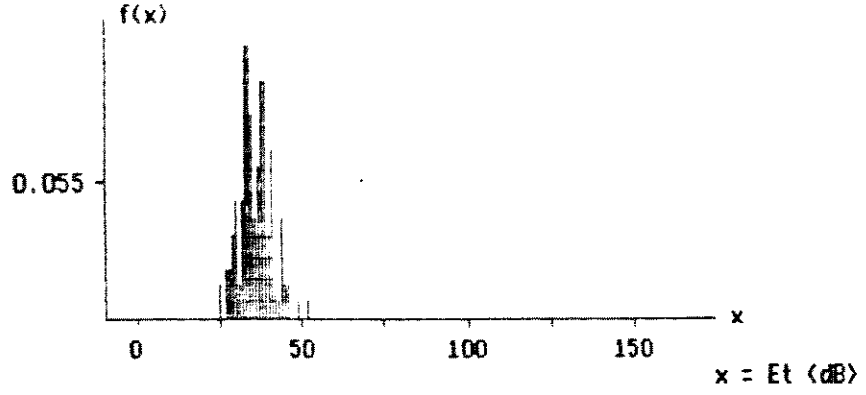
APÊNDICES

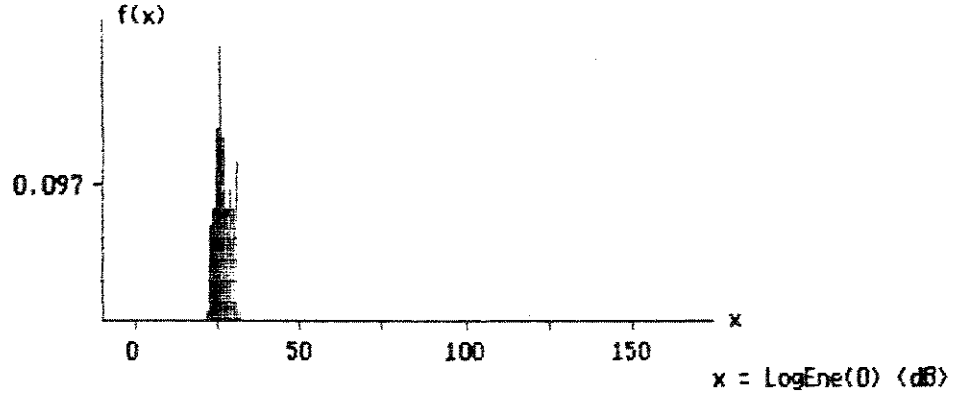
SILÊNCIO


f(x)
0.038
x
0
50
100
150
x = ZRX

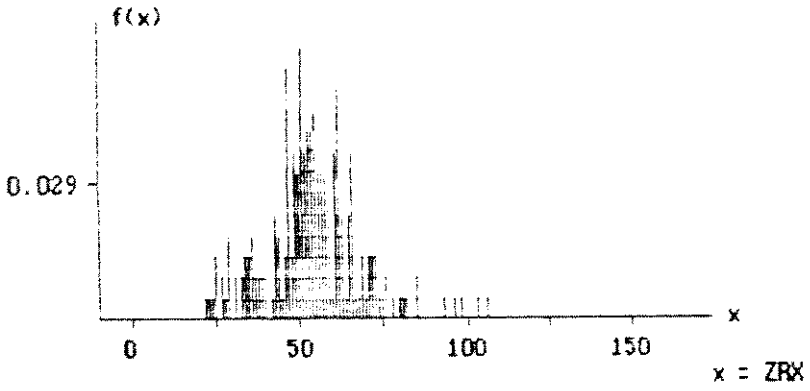

f(x)
0.024
x
0
50
100
150
x = NTP

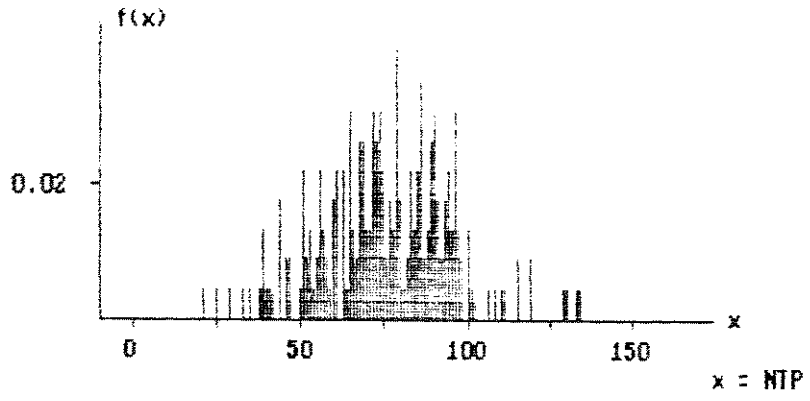

f(x)
0.047
x
-50
-25
0
25
50
x = ΔPN

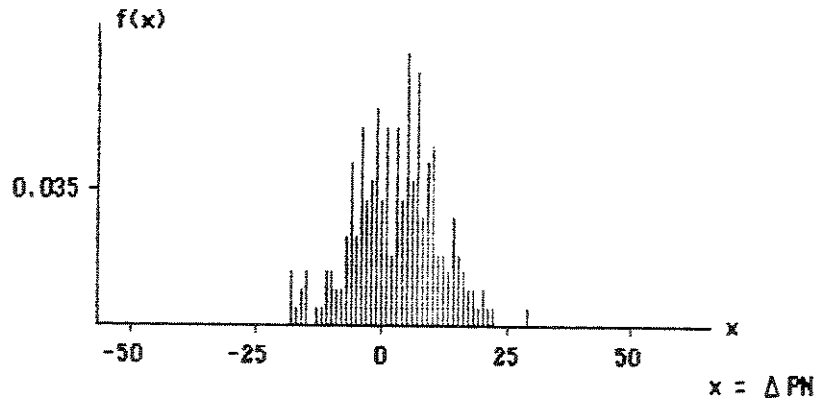

f(x)
0.055
x
0
50
100
150
x = Et (dB)

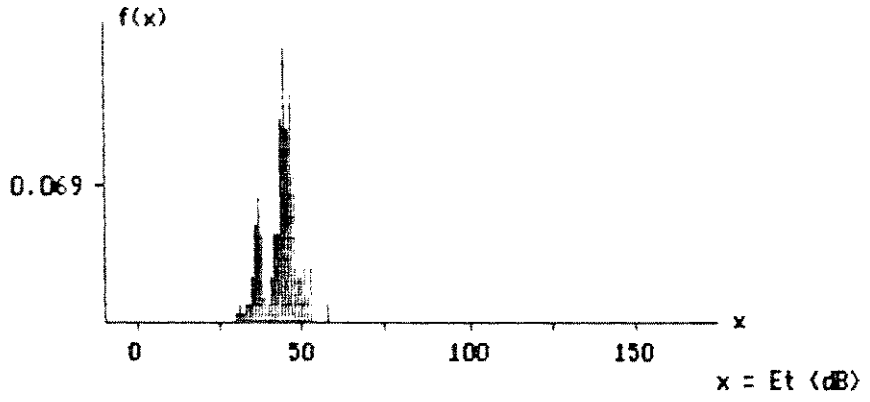

f(x)
0.097
x
0
50
100
150
x = LogEne(0) (dB)

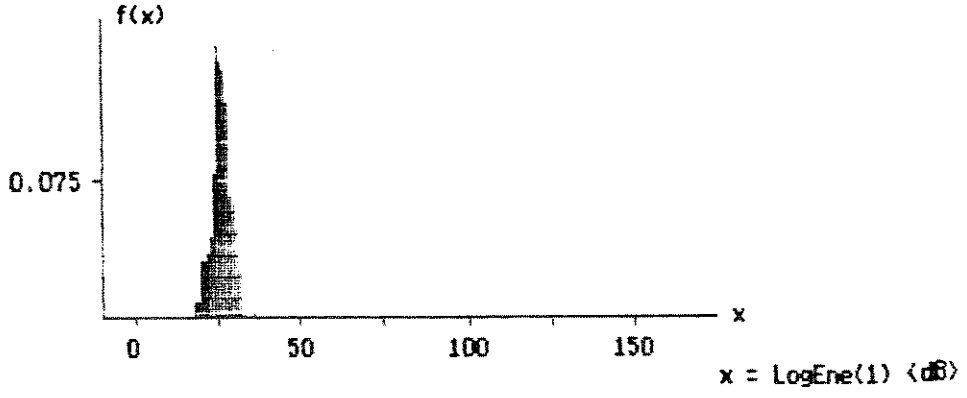
FRICATIVO SURDO (FRACO)



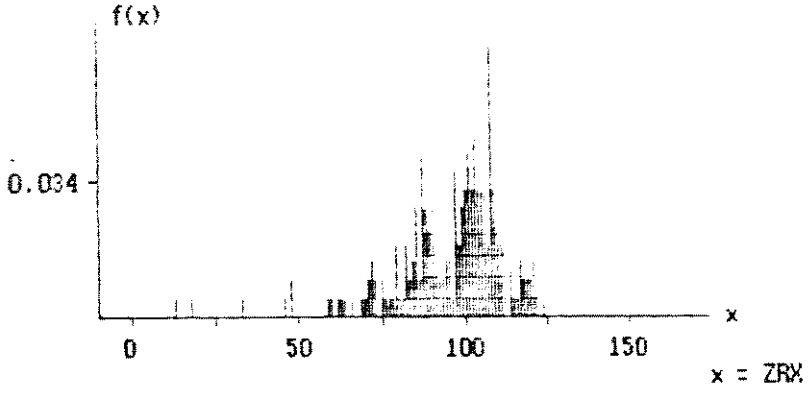


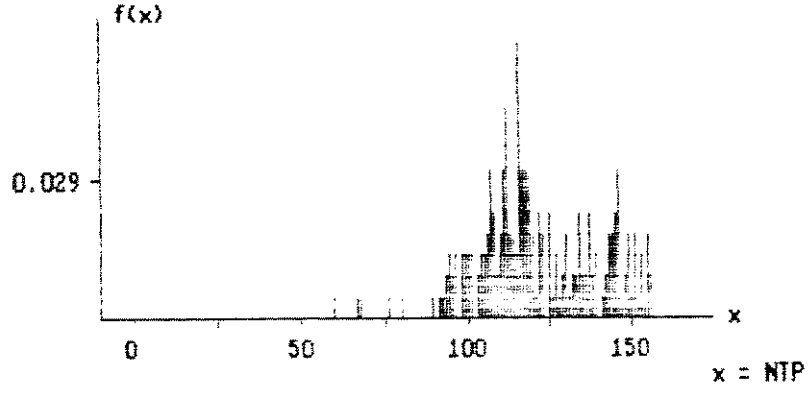


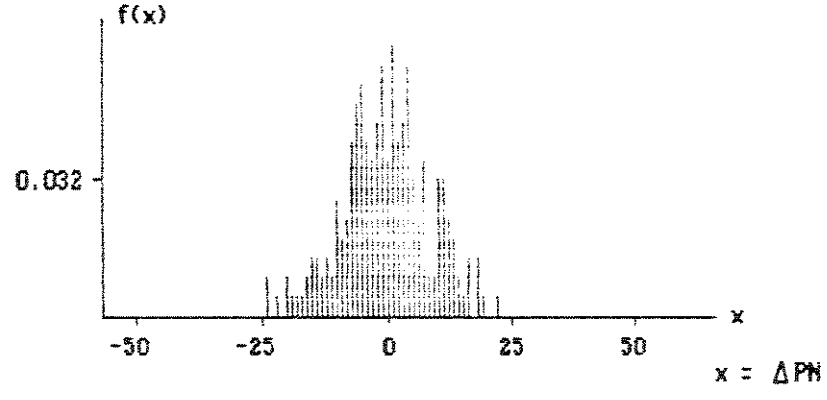

f(x)
0.069
x
0
50
100
150
x = Et (dB)

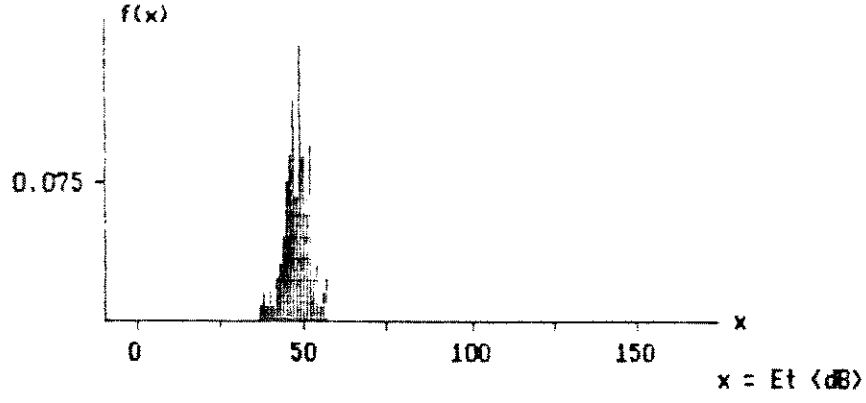

f(x)
0.075
x
0
50
100
150
x = LogEne(1) (dB)

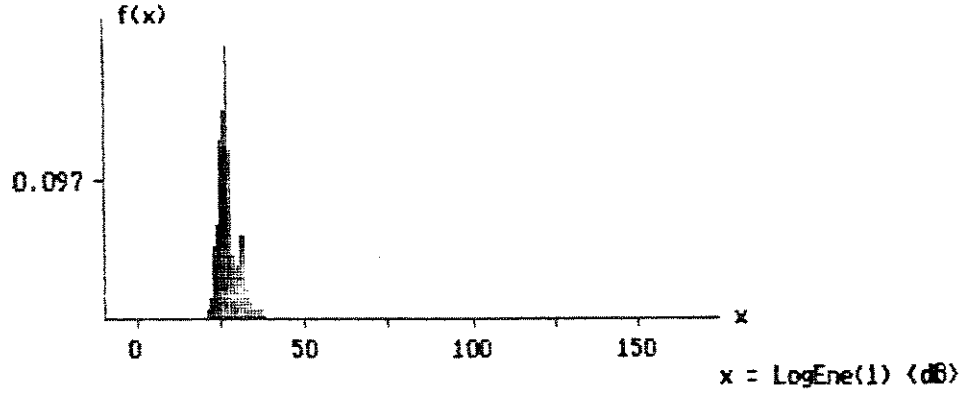
FRICATIVO SURDO (FORTE)



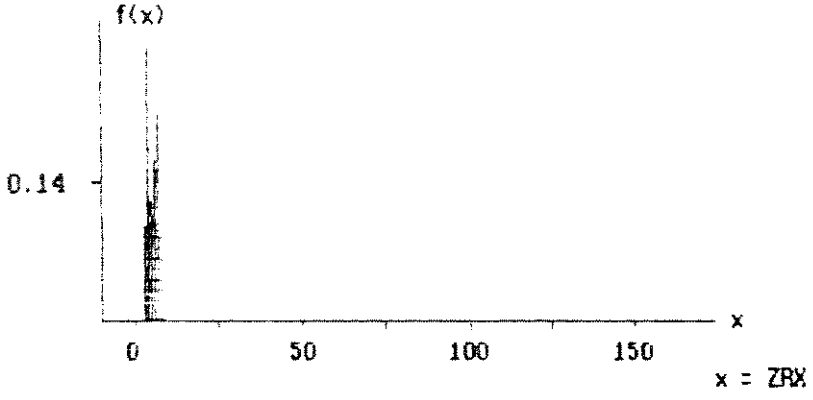


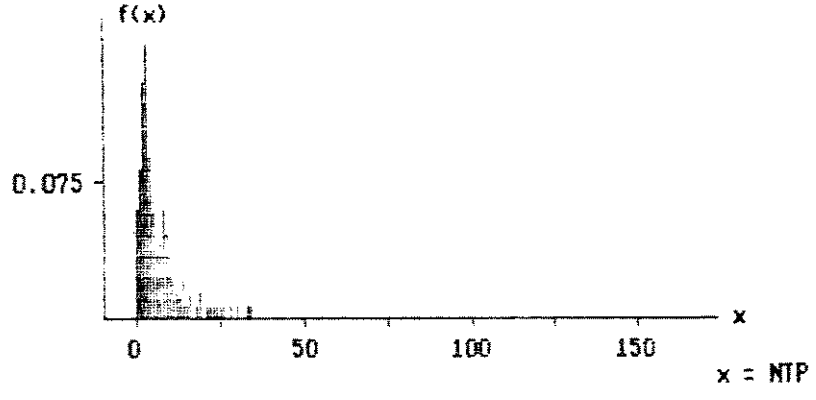


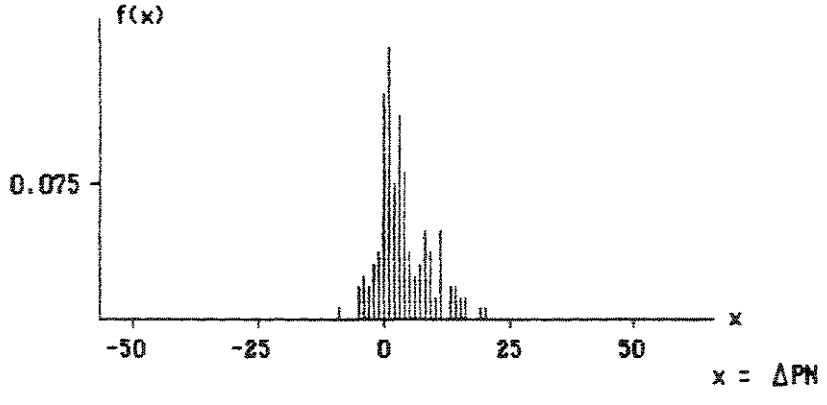

f(x)
0.075
x
0
50
100
150
x = Et (dB)

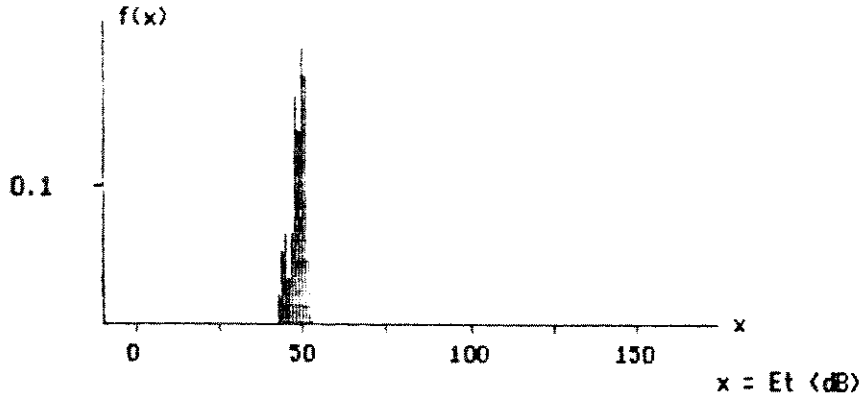

f(x)
0.097
x
0
50
100
150
x = LogEne(1) (dB)

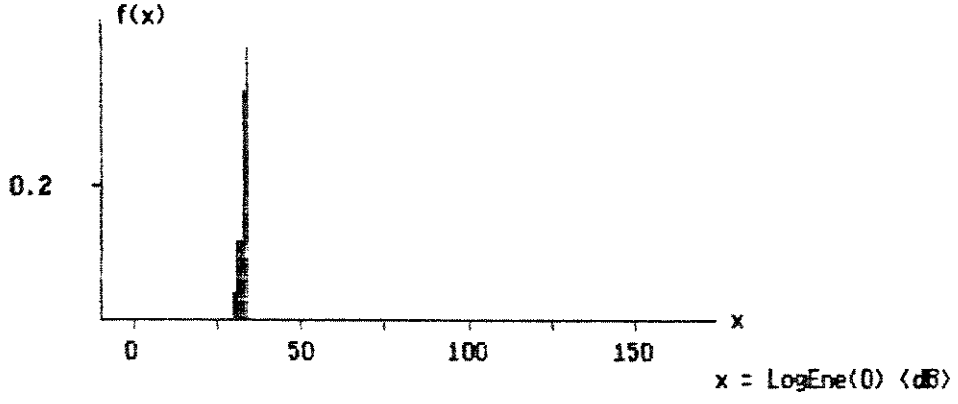
OCLUSÃO SONORA

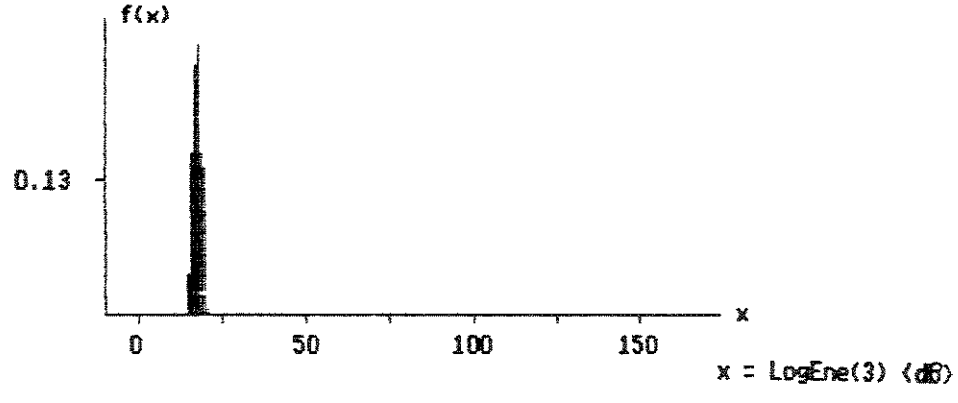




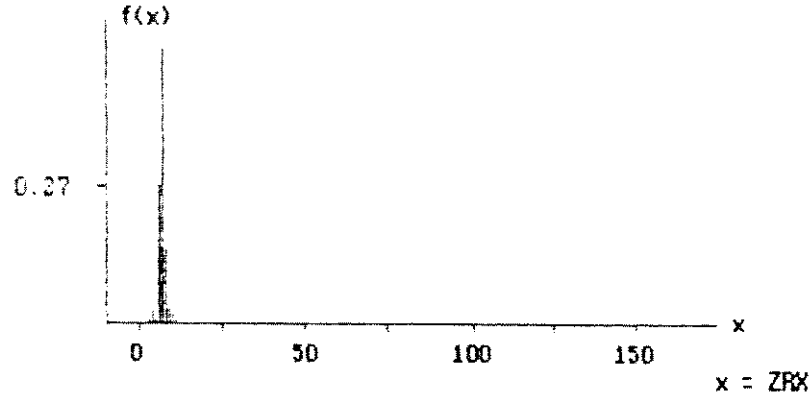


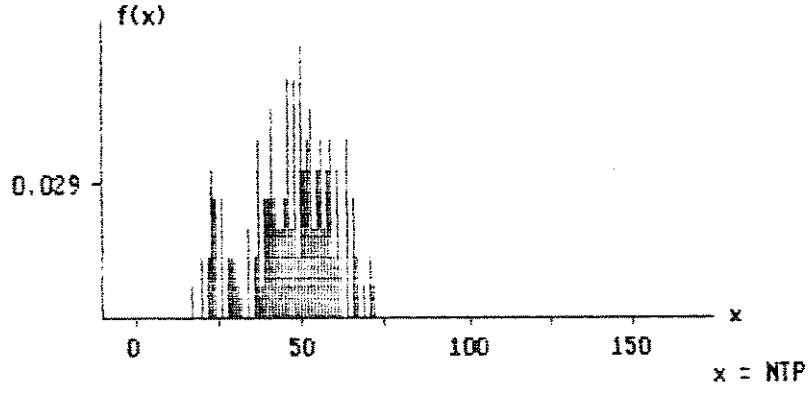

f(x)
0.1
x
0
50
100
150
x = Et (dB)

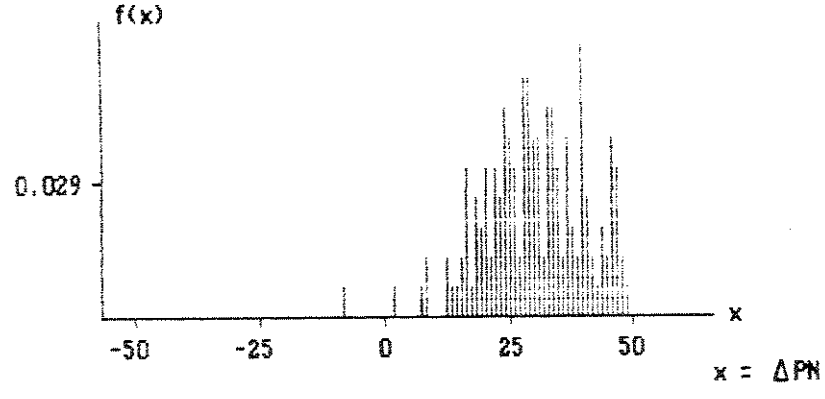

f(x)
0.2
x
0
50
100
150
x = LogEne(0) (dB)

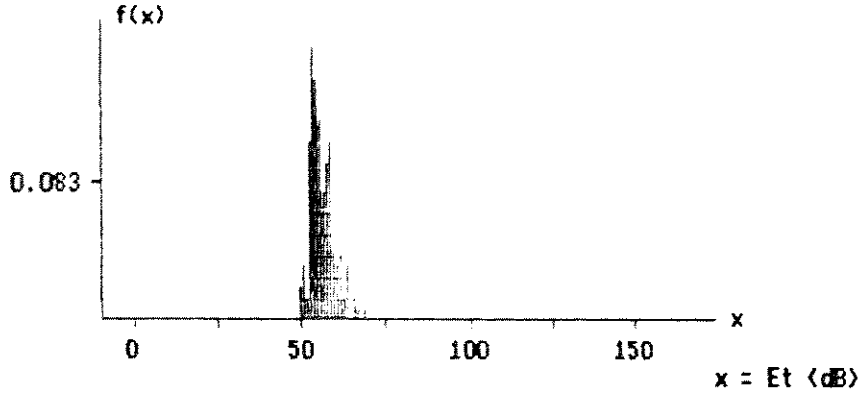

f(x)
0.13
x
0
50
100
150
x = LogEne(3) (dB)

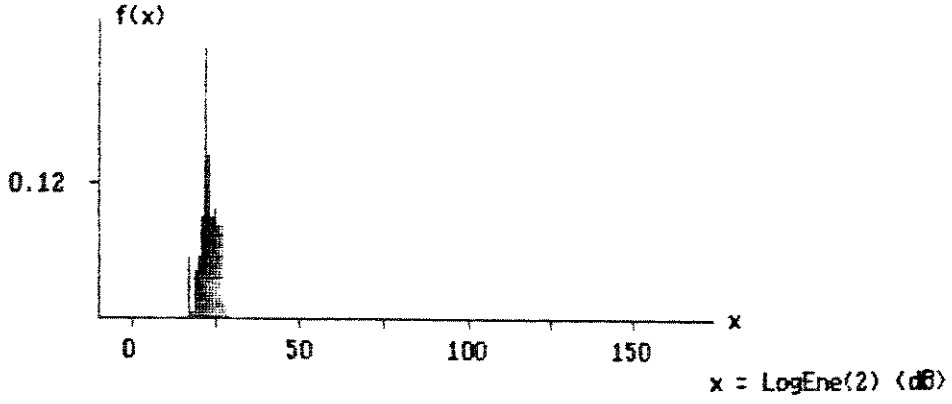
FRICATIVO SONORO



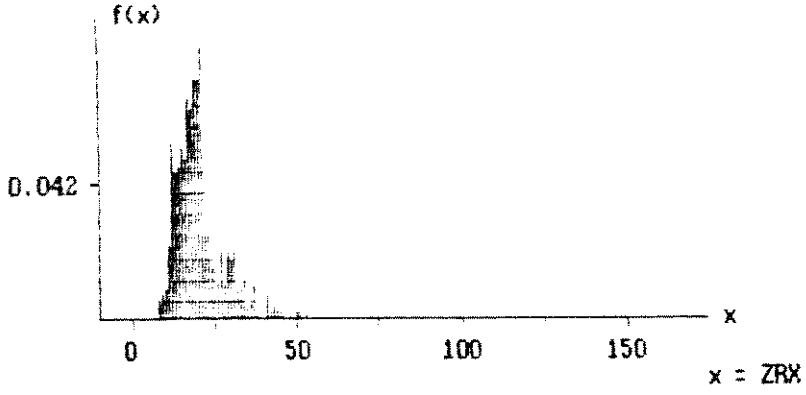


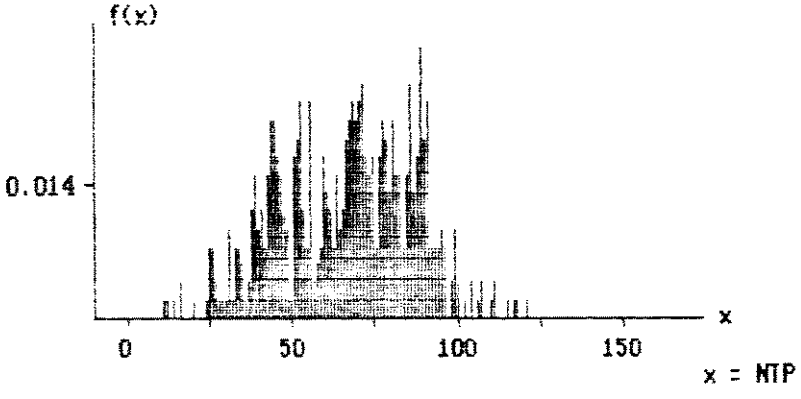


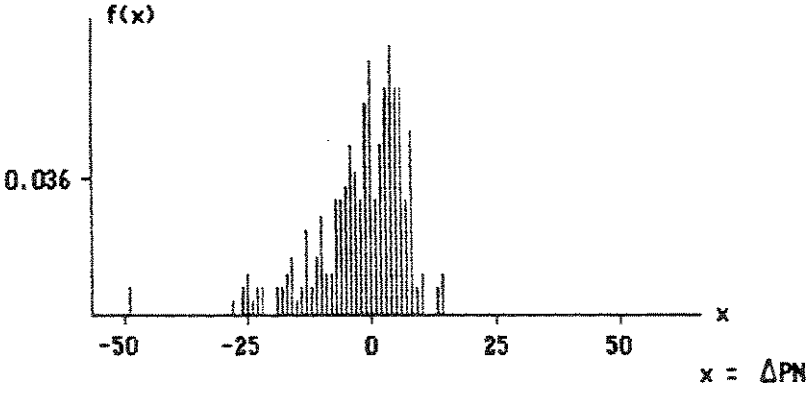

f(x)
0.083
x
0
50
100
150
x = Et (dB)

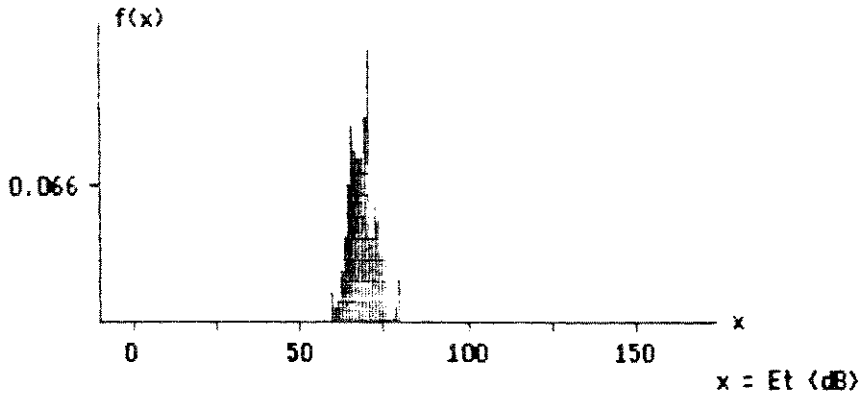

f(x)
0.12
x
0
50
100
150
x = LogEne(2) (dB)

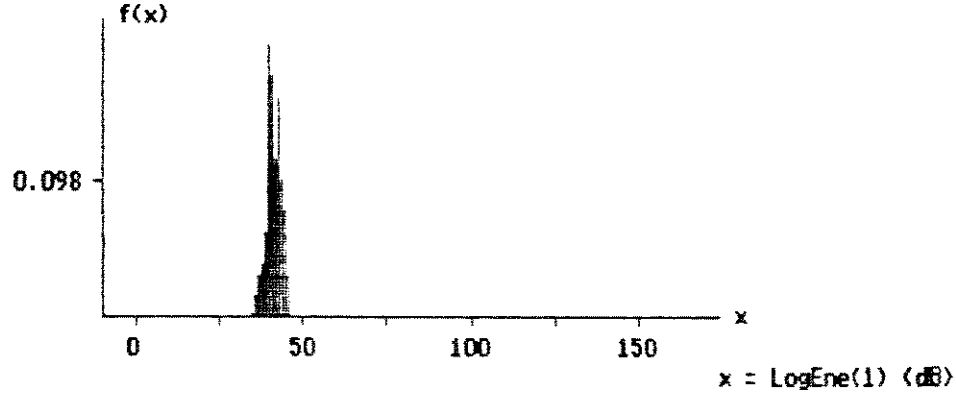
VOCÁLICO

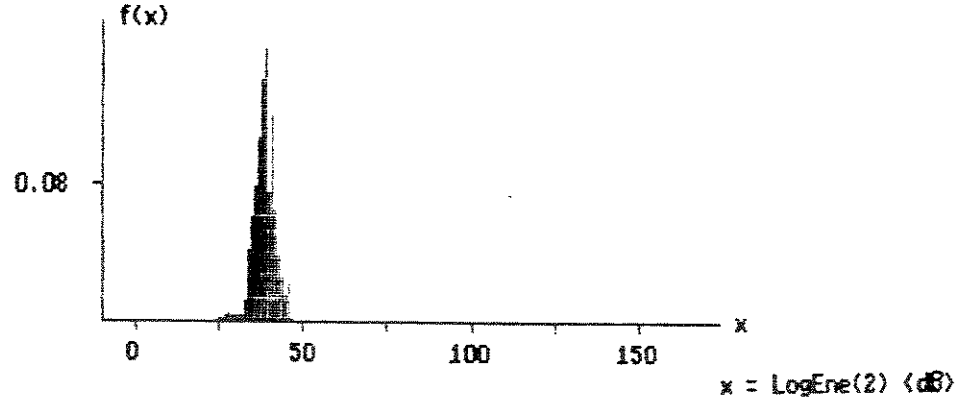

f(x)
0.042
x
0
50
100
150
x = ZRX


f(x)
0.014
x
0
50
100
150
x = NTP


f(x)
0.036
x
-50
-25
0
25
50
x = ΔPN


f(x)
0.066
x
0
50
100
150
x = Et (dB)


f(x)
0.098
x
0
50
100
150
x = LogEne(1) (dB)


f(x)
0.08
x
0
50
100
150
x = LogEne(2) (dB)

ALGORITMO DE LEVINSON/DURBIN

```
#include <math.h>
#include <conio.h>
#include "const.h"

durbin()
{
    register int i, j;
    double R[M+1], k[M+1], E, aux, a_anterior[M+1];

    for(i=0; i<=M; i++)
        {
           R[i]  = 0.0;
           k[i] = 0.0;
           a_anterior[i] = 0.0;
           a[i] = 0.0;

           for(j=0; j<TAMX-i; j++)
               {
                  R[i] += x[j]*x[j+i];
               }
        }


    E = R[0];

    for(i=1; i<=M; i++)
        {
           aux = 0.0;
           for(j=i; j<i; j++)
               {
                  aux += a_anterior[j]*R[i-j];
               }

           k[i] = (R[i]-aux)/E;

           a[i] = k[i];

           for(j=i; j<i; j++)
               {
                  a[j] = a_anterior[j] - k[i]*a_anterior[i-j];
               }

          E *= (1 - k[i]*k[i]);

           for(j=1; j<=i; j++)
              {
                a_anterior[j] = a[j];
              }

        }
```

```
    a[0] = 1.0;

    G = R[0];
    for(j=1; j<=M; j++)
        {
            G -= a[j]*R[j];
        }

}
```

```
/* ARQUIVO const.h */


#define M  12                       /* ordem da lpc */
#define NEST 10                     /* num. de estagios da fft */

#define TAMX 200                    /* tamanho do quadro */
#define TAMY 512                    /* tam. do vetor espectral */
#define TAMA M%2 == 0 ? M+2 : M+1     /* tam. do vetor de coef.
                                                    lpc */


#define PI 3.14159265358979

extern double x_orig[TAMX];         /* sequencia de voz */
extern double x[TAMX];              /* quadro de voz c/ jan. Hamming
                                                  e pre-enfase */
extern double y[TAMY];              /* sequencia espectral */
extern double a[TAMA];              /* coeficientes do LPC */
extern double G;                    /* fator de ganho do lpc */

extern int ZRX, NTP, pitch[2], suso[3];
extern double et1, et2, Et, f0;
```

OBSERVAÇÃO: para facilitar a utilização das funções da biblioteca de matemática, foi utilizado o comprimento "double" para todas as variáveis reais. Numa próxima etapa, pretende-se estudar a viabilidade de uso de variáveis do tipo "float" ou mesmo "int" em substituição às variáveis "double".

ALGORITMO FFT

```
#include (math.h)
#include "const.h"

trfp(vetor , tam_vetor , m, l)
double vetor[];
int   tam_vetor , m, l;

/*   m = numero de estagios da FFT - i       */
/*   l = numero de estagios nao nulos - i    */

{
   register int i, j;
   int k, rep, disp, j2;
   double arg, twf, c, s;
   double V_re_par, V_re_impar, V_im_par, V_im_impar;

   int n  = (int) pow(2, m);
   int nl = (int) pow(2, l);
   int np = n/nl;

   struct complexo
        {
          double re;
          double im;
        } t, u, v[TAMY+1];

   /* "ENTRELACA" A SEQUENCIA A SER TRANSFORMADA */
   for(i=0; i(tam_vetor/2; i++)
       {
         v[i+1].re = (double) vetor[2*i];
         v[i+1].im = (double) vetor[2*i+1];
       }

   /* ZERA POSICOES FINAIS DE v(i) */
   if(n)nl)
     {
       for(i = tam_vetor/2 + 1; i(=n; i++)
           {
             v[i].re = 0.0;
             v[i].im = 0.0;
           }
     }

   /* BIT REVERSO (COOLEY & TUKEY) */
   j=1;
   for(i=1; i(nl; i++)
        {
          if(i(j)
            {
              t.re = v[j].re;
              t.im = v[j].im;
```

```
            v[j].re = v[i].re;
            v[j].im = v[i].im;

            v[i].re = t.re;
            v[i].im = t.im;
          }
        k = nl/2;
        while(k<j)
              {
                j-=k;
                k/=2;
              }
        j+=k;
      }

/* TRANSFORMADA RAPIDA "PODADA" (DAVID P. SKINNER) */
            /* ASSP abril 1976, pag. 193 */
for(i=1; i<=nl; i++)
    {
      for(j=1; j<=np; j++)
          {
            k = nl-i;
            v[k*np+j].re = v[k+1].re;
            v[k*np+j].im = v[k+1].im;
          }
    }

for(i = m-l+1; i<=m; i++)
    {
      rep = (int) pow(2, i);
      disp = rep/2;
      arg = 2*Pi/rep;

      for(j=1; j<=disp; j++)
          {
            twf = (j-1)*arg;
            c = cos(twf);
            s = sin(twf);

            for(k=j; k<=n; k+=rep)
                {
                  j2 = k+disp;

                  t.re =  c*v[j2].re + s*v[j2].im;
                  t.im = -s*v[j2].re + c*v[j2].im;

                  v[j2].re = v[k].re - t.re;
                  v[j2].im = v[k].im - t.im;

                  v[k].re += t.re;
                  v[k].im += t.im;
                }
          }
    }
```

```
/* "DESEMBARALHA" (VER OPPENHEIM & SCHAFER, EXERCICIO 6.10) */
/* pontos particulares */
y[0] = pow( (v[0].re + v[0].im), 2 );
y[n/2] = pow( v[n/2].re, 2) + pow( v[n/2].im, 2);

/* demais pontos */
j = n+2;
for(k=1; k<n; k++)
    {
      i = k+1;    arg = k*PI/n;

      V_re_par    =  v[i].re + v[j-i].re;
      V_re_impar  =  v[i].re - v[j-i].re;
      V_im_par    =  v[i].im + v[j-i].im;
      V_im_impar  =  v[i].im - v[j-i].im;

      u.im = V_ im_par * cos(arg) - V_re_impar * sin(arg);
      u.re = -  V_re_impar * cos(arg) - V_im_par * sin(arg);

      t.re = V__re_par + u.im;
      t.im = V__im_impar + u.re;
      y[k] = (  pow(t.re, 2) + pow(t.im, 2) )/4;

      t.re = V__re_par - u.im;
      t.im = -  V_im_impar + u.re;
      y[n-k] =  ( pow(t.re, 2) + pow(t.im, 2) )/4;
    }

}
```

DETECTOR DE PITCH

```
#include (math.h)
#include "const.h"

pitch_amdf()
{
   double w[TAMX];
   double  amdf[TAMX/2], limiar1, limiar 2, limiar3,
           nrat, min, max, aux, auxi;
   int n, j, k, logica, minp;

   limiar1=5.0; limiar2=1200.0; limiar3=2.0;

   if( et1/et2)50 || et2/et1)50 || ZRX)70 || Et(30 )
     {
       pitch[i] = 0;
       suso[2] = suso[1]; suso[0] = 0;
       f0 = 0;
       return;
     }

   /* faz filtragem */
   for(n=0; n(TAMX-4; n++)
      {
         aux=0.0;
         for(j=0; j(5; j++)  aux += x_orig[n+j];

         w[n] = (double) aux/5.0;
      }

   for(n=TAMX-4; n(TAMX; n++) w[n] = x_orig[n];

   /* calcula AMDF(k) */
   for(k=0; k(TAMX/2; k++)
      {
         amdf[k] = 0.0;

         for(j=0; j(TAMX/2 ; j++)
            {
               amdf[k] += fabs(w[j]-w[j+k]);
            }
      }


   if(ZRX)5 && ZRX(40 && Et)40 ) suso[0] = 1;
   else suso[1] = 0;
```

```
 * localiza maximo, minimo e posicao do minimo da AMDF(k) */
max=amdf[24]; min = 10000000.0; minp = 1;

for(j=25; j<TAMX-1; j++)
    {
       if(amdf[j]>max)  max = amdf[j];

       aux = (double) amdf[j]/min;
       if( amdf[j-1]>=amdf[j] && amdf[j]<=amdf[j+1] && amdf[j]<=min)
         {
           min = amdf[j];
           minp = j;
         }
     }

nrat = max/min;
logica = suso[0] + 2*suso[1] + 4*suso[2];

switch(logica)
       {
          case 0:
          case 2:
          case 4:
                  if( nrat<limiar1 || (nrat>=limiar1 && max<limiar2) )
                    {
                       pitch[0] = 0;
                    }

                  else
                      {
                         pitch[0] = minp;
                         suso[0] = 1;
                      }

                  break;

          case 1:
                  if( max<=limiar2 || (max>limiar2 && nrat<=limiar3) )
                    {
                       pitch[0] = 0;
                       suso[0] = 0;
                    }

                  else pitch[0] = minp;

                  break;

          case 3:
          case 5:
          case 7:
                  pitch[0] = minp;
                  break;
```

```
            case 6:
                    pitch[0] = pitch[1];
                    break;

            default : break;

        }

  if(minp==1) pitch[0] = 0;

  aux = (double) pitch[0];
  if(pitch[1]!=0)
    {
      if(aux)1.8*pitch[1] && aux(2.2*pitch[1]) pitch[0] = minp/2;
      else if(aux)2.7*pitch[1] && aux(3.3*pitch[1]) pitch[0] = minp/3;
      else if(aux)0.45*pitch[1] && aux(0.55*pitch[1]) pitch[0] = 2*minp;
    }

  f0 = (pitch[0] == 0) ? 0.0 : 8000.0/pitch[0];

  suso[2] = suso[1]; suso[1] = suso[0];

  pitch[1] = pitch[0];

}
```