UNIVERSIDADE FEDERAL DA PARAÍBA

CENTRO DE CIÊNCIAS E TECNOLOGIA

CURSO DE MESTRADO EM ENGENHARIA ELÉTRICA

LEVANTAMENTO ESTATÍSTICO DA LOCUÇÃO E CONVERSAÇÃO PARA SINAIS DE VOZ

Niomar Lins Pimenta

Campina Grande, Pb

UNIVERSIDADE FEDERAL DA PARAÍBA

CENTRO DE CIÊNCIAS E TECNOLOGIA

CURSO DE MESTRADO EM ENGENHARIA ELÉTRICA

LEVANTAMENTO ESTATÍSTICO DA LOCUÇÃO E CONVERSAÇÃO PARA SINAIS DE VOZ

Niomar Lins Pimenta

Campina Grande, Pb

LEVANTAMENTO ESTATÍSTICO DA LOCUÇÃO E CONVERSAÇÃO PARA SINAIS DE VOZ

Niomar Lins Pimenta

Dissertação apresentada ao Curso de Mestrado em Engenharia Elétrica da Universidade Federal da Paraíba, em cumprimento às exigências para obtenção do grau de Mestre.

Benedito Guimarães Aguiar Neto, Dr.-Ing.

Orientador

Marcos Antonio Gonçalves Brasileiro, D. Sc.

Co-Orientador

Campina Grande, Pb

Julho/1992

P644e Pimenta, Niomar Lins
Levantamento estatistico da locucao e conversacao para sinais de voz / Niomar Lins Pimenta. - Campina Grande, 1992.
180 f. : il.

Dissertacao (Mestrado em Engenharia Eletrica) - Universidade Federal da Paraiba, Centro de Ciencias e Tecnologia.

1. Sinais de Voz 2. Sinais de Voz (Locucao e Conversacao) 3. Engenharia Eletrica 4. Dissertacao I. Aguiar Neto, Benedito Guimaraes, Dr. II. Brasileiro, Marcos Antonio Goncalves, Dr. III. Universidade Federal da Paraiba - Campina Grande (PB) IV. Título

CDU 621.91(043)

LEVANTAMENTO ESTATÍSTICO DA OCLUSÃO E CONVERSAÇÃO
PARA SINAIS DE VOZ

NIOMAR LINS PIMENTA

DISSERTAÇÃO APROVADA EM 17.07.1992

BENEDITO GUIMARÃES AGUIAR NETO, Dr.-Ing., UFPB
Orientador

MARCOS ANTONIO GONÇALVES BRASILEIRO, D.Sc., UFPB
Co-Orientador

JOÃO MARQUES DE CARVALHO, Ph.D., UFPB
Componente da Banca

ABRAHAM ALCAIM, Ph.D., CETUC/PUC - IBM
Componente da Banca

CAMPINA GRANDE - PB
JULHO - 1992

À minha espôsa Soraya, aos meus filhos Caio e Naiara e à minha mãe Maria.

Ao encerrar os trabalhos desta dissertação, alguns agradecimentos especiais a todos que, de forma direta ou indireta, auxiliaram em sua execução, devem ser registrados.

À FUCAPI, muito particularmente ao seu Diretor Executivo Dr. Aluísio Brasil Barbosa e sua Diretora Técnica Drª Isa Assef dos Santos, pelo apoio e oportunidade.

À Universidade do Amazonas pela oportunidade concedida para a realização de meu curso de mestrado.

Aos professores Benedito Guimarães Aguiar Neto e Marcos Antonio Gonçalves Brasileiro, pela orientação, estímulo, apoio e confiança que me dedicaram ao longo de todo o mestrado.

Ao casal Sérgio e Mônica que, nos momentos mais difíceis, ofereceram expontânea mostra de amizade.

Ao amigo João Edgar Chaves e família, que muito me ajudaram no início de meu curso de mestrado.

Ao casal de amigos Antonio Luiz e Marisol, que ofereceram, a mim e minha família, desde nossa chegada, fraternos sentimentos de amizade.

Ao colega Carlos Akira, pela ajuda na cessão de diversos programas, utilizados como referência para a realização deste trabalho.

À professora Maria Auxiliadora Bezerra, pelo apoio fornecido na área de linguística, através do fornecimento de material bibliográfico, palestras e informações adicionais.

SUMARIO

Este trabalho trata do levantamento estatístico de eventos dos sinais de voz, tanto em locuções individuais, quanto em conversações telefônicas simuladas, utilizando parâmetros temporais.

No caso de locução, foi desenvolvido um detetor de voz com o objetivo de classificar e possibilitar um levantamento estatístico da ocorrência de sons sonoros, de sons surdos e de intervalos de silêncio.

Para os sinais de conversação, foi desenvolvido um outro detetor com a finalidade de determinar os intervalos com presença ou ausência de voz, de forma a possibilitar um levantamento estatístico de eventos de interesse em uma conversação telefônica, tais como duração média dos surtos de voz, duração média das pausas, atividade de voz e taxa de surtos de voz.

ABSTRACT

This work deals with the statistical measure of speech signals, obtained through individual expressions and simulated telephonic conversations, using temporal parameters.

In the event of individual expression, a speech detector was developed with the goal of classifying and obtaining statistical measures of voiced-unvoiced-silence ocurrences.

In the event of telephonic conversation signals, another speech detector was developed with the goal of determining the presence of speech. Statistical measures of events like mean talkspurt duration, mean pause duration, talkspurt rate and speech activity were obtained.

INDICE

INDICE

LISTA DE ILUSTRAÇÕES

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE TABELAS

Capítulo 4:

CAPITULO 1
INTRODUÇÃO

CAPÍTULO 1

INTRODUÇÃO

1.1 Comentários Iniciais

O processamento digital de sinais têm dois objetivos básicos [1]:

- estimar os parâmetros característicos do sinal;
- transformar o sinal em uma forma de representação mais adequada;

Os sinais podem ser definidos como funções ou seqüências numéricas de uma ou de várias variáveis independentes, que, tipicamente, transporta informação acêrca do estado ou do comportamento de um fenômeno ou sistema físico. Um sistema, em um contexto restrito, é definido como um algoritmo de processamento dos sinais [2].

Transformações matemáticas aplicadas a um sinal digital, têm efeito similar ao fluxo de um sinal analógico original, através de uma série de filtros. Em muitos aplicações, como, por exemplo, equipamentos criptográficos, análise de sinais em modems, síntese de voz

e sistemas de reconhecimento, é mais fácil e preciso processar os sinais com técnicas digitais, do que utilizando componentes analógicos. Nessas aplicações, uma das maiores vantagens está na utilização dos dispositivos de processamento digital de sinais com alta capacidade, de modo que os investimentos realizados em hardware são preservados, mesmo com o incremento de softwares mais sofisticados [3].

Os estudos na área de processamento digital de sinais de voz, objetivo específico deste trabalho, foram bastante intensificados a partir do momento em que se vislumbrou a enorme quantidade de aplicações possíveis, utilizando-se as técnicas desenvolvidas por diversos pesquisadores em todo o mundo e associando a esses estudos, áreas como a acústica, a lingústica e a microeletrônica, dentre outras [4].

A implementação dos conhecimentos teóricos elaborados ao longo dos muitos anos de pesquisa, de forma econômicamente viável, foi possível graças ao extraordinário avanço da microeletrônica, que possibilitou o desenvolvimento de processadores específicos, utilizados em aplicações em tempo real, capazes de operacionalizar os diversos algoritmos até então construídos e, por impossibilidades práticas, não utilizados [5].

Incontinenti, começaram a surgir aplicações práticas importantes, especialmente entre pessoas e entre estas e máquinas.

Uma das principais áreas de aplicação é a de transmissão digital de sinais de voz, onde procura-se desenvolver técnicas que aumentem a capacidade do canal por usuário, através de codificação da voz, buscando a redução da largura de faixa para a transmissão. A utilização de sinais de voz digitalizados, com taxa de bits mais baixa possível, é compatível com aplicações futuras que utilizarão terminais de baixo custo, acoplados a redes totalmente digitais [1].

A tecnologia TASI (time assignment speech interpolation) foi desenvolvida para utilizar os intervalos de silêncio em uma conversação entre dois assinantes, inserindo outras conversações e aumentando a capacidade dos canais telefônicos. Estatísticas dos surtos de energia dos sinais de voz presentes em um grupo de canais de comunicação, são fundamentais nesse processo. Em cabos transatlânticos esses grupos de canais chegam a 36 [4].

Outra área de aplicação para o processamento digital de sinais de voz, refere-se à comunicação homem/máquina. Três sub-áreas são, reconhecidamente, as mais importantes nessa atividade [1]:

- sistemas de reconhecimento de locutor;
- sistemas de reconhecimento de fala;
- sistemas de resposta vocal.

Os sistemas de reconhecimento de locutor obtêm a representação do sinal de voz, utilizando técnicas que preservam as características da fala, relevantes à identificação do locutor. A configuração resultante é comparada a outras previamente preparadas, utilizadas como referência e uma lógica de decisão escolhe a adequada, dentre as alternativas disponíveis.

Os sistemas de reconhecimento de fala têm por função básica, reconhecer, exatamente, a expressão falada, como na utilização de máquinas de datilografia operadas por voz, ou compreender a expressão falada, i.e., responder, de maneira correta, ao que foi falado.

Os sistemas de resposta da voz são projetados para responder a pedidos para informações, utilizando mensagens faladas. Ao contrário das duas sub-áreas anteriores, nos sistemas de resposta da voz a comunicação vocal é realizada na direção da máquina para o homem.

1.2 Caracterização do Trabalho

Este trabalho têm por objetivo, obter levantamentos estatísticos de eventos dos sinais de voz, tanto em locuções individuais quanto em conversações telefônicas simuladas, a partir de parâmetros temporais, utilizando para isso dois processos distintos:

- um conjunto de locutores emitindo uma frase, elaborada com base nos fonemas mais utilizados na língua portuguesa falada no Brasil;
- um conjunto de conversações telefônicas simuladas.

No processo de locução, é efetuada a determinação dos sons empregados na pronúncia de uma palavra ou frase, realizada por uma pessoa. Neste trabalho, a caracterização da voz utilizada para locução é realizada, classificando-se cada segmento do sinal em análise, como um som sonoro, um som surdo ou como ausência de voz ou silêncio. Essa classificação é muito importante em comunicações homem/máquina como, por exemplo, os sistemas que funcionam independentemente do locutor, onde o conhecimento da área de lingüística é necessário [1].

A caracterização dos sinais de voz originados de conversações, é efetuada através da geração de um sinal ON-OFF indicando a presença (estado ON) ou ausência (estado OFF) da voz no segmento em análise. O conhecimento dos intervalos de uma conversação telefônica que, efetivamente, contenham voz, é muito importante para melhorar a eficiência dos meios de comunicação utilizados, através da construção de interpoladores de voz, que associam vários assinantes em um mesmo canal, a partir do conhecimento das estatísticas desses intervalos. A

denominação "conversação telefônica simulada" foi empregada, por não terem sido utilizadas, nas conversações, as redes telefônicas das empresas concessionárias desses serviços.

Os sons sonoros, os sons surdos e os intervalos de ausência de voz, são denominados de eventos de locução e os intervalos ON (surtos de voz) e OFF (ausência de voz), são denominados de eventos de conversação.

Para que um segmento do sinal em análise seja classificado como um som surdo, um som sonoro ou silêncio (ausência da voz) no caso das locuções, e seja considerado como um estado ON ou OFF no caso das conversações, é utilizado um dispositivo denominado de detetor de voz.

Neste trabalho esse dispositivo foi construído através de parâmetros temporais do sinal de voz. É importante mencionar que cada um dos processos têm um detetor de voz específico, dadas às peculiaridades de cada aplicação.

A utilização de uma janela muito curta para a segmentação do sinal de voz, como a empregada neste trabalho, permite que os parâmetros temporais reflitam com maior precisão as variações do sinal de voz, mas não possibilitam a determinação dos parâmetros espectrais como a freqüência fundamental dos sinais quase-periódicos da voz humana (sons sonoros), e dos formantes. Daí a utilização apenas dos parâmetros temporais.

No caso das locuções o detetor foi desenvolvido

utilizando-se os parâmetros temporais energia, taxa de cruzamento por zero, variação de energia, número total de picos, coeficiente de autocorrelação normalizado e diferença de picos. Justifica-se o número elevado de parâmetros temporais para a construção do detetor, pela não utilização de parâmetros espectrais. Isso exige testes adicionais, visando a classificação mais precisa possível do tipo de som emitido.

Para as conversações telefônicas, não há a necessidade do rigor em se determinar o tipo de som emitido, mas sim de garantir, com precisão, se o intervalo em análise representa a presença ou ausência do sinal de voz. Em vista disso, o detetor foi desenvolvido a partir dos parâmetros temporais energia, taxa de cruzamento por zero e variação de energia.

A partir dos resultados obtidos nas saídas dos detetores, são levantados dados estatísticos dos eventos mais representativos tanto para locução quanto para conversação.

Os eventos cujos levantamentos estatísticos foram obtidos, no caso da locução, foram os sons sonoros, os sons surdos e o silêncio. Para as conversações telefônicas simuladas, além das estatísticas dos surtos de voz e dos intervalos de ausência da voz, foram obtidas ainda, medidas da duração média dos surtos de voz, da atividade de voz, da taxa de surtos de voz e da duração média das pausas.

Em têrmos da contribuição do trabalho realizado, no processo de locução o detetor desenvolvido apenas com parâmetros temporais, utiliza um processamento original para classificar os sons da língua portuguesa falada no Brasil, apesar de basear-se em outros algoritmos existentes. Pretende-se mostrar que é possível obter um detetor robusto para a classificação sonoro-surdo-silêncio, utilizando-se parâmetros temporais, desde que seja utilizada uma janela de duração bastante reduzida. Neste trabalho a duração da janela é de 4 ms.

No caso do processo de conversação, a contribuição é mais acentuada devido ao pioneirismo do trabalho para a língua portuguesa falada no Brasil. O detetor elaborado utiliza métodos eficientes para detetar a transição entre os intervalos com presença de voz e ausência de voz, além de *corrigir* adequadamente o sinal ON-OFF, de possíveis ações do ruído ou de interrupções intrínsecas ao ato da fala, como pequenas hesitações e pausas intersilábicas. Os resultados obtidos podem ser utilizados para auxiliar na construção de interpoladores ou mesmo no modelamento de tráfego telefônico.

É importante mencionar que o ruído está presente nos dois casos, por ser um sinal inerente ao processo. Os ruídos ambientais, como os produzidos pelo homem, foram convenientemente reduzidos durante o processo de gravação das fontes de voz, através da utilização de uma sala fechada. Os ruídos dependentes do sinal, os ruídos naturais

internos como o ruído térmico e os causados por componentes ativos, os ruídos da fita cassete, os ruídos do microfone e o ruído de quantização da placa utilizada na digitalização do sinal de voz, foram minimizados pela utilização de dispositivos e aparelhos de boa qualidade.

1.3 Configuração da Dissertação

Após esta breve introdução, o Capítulo 2 deste trabalho faz considerações gerais a respeito do sinal de voz, apresentando suas características básicas e a forma como é produzida, definindo os intervalos básicos da fala, além do processo de deteção.

No Capítulo 3 são definidos os parâmetros temporais utilizados na construção dos detetores de voz e caracterizados os eventos de voz, tanto para a locução quanto para a conversação.

Já no Capítulo 4 são apresentadas as características da locução, a elaboração do detetor de voz e os resultados obtidos para os eventos definidos nesse processamento.

No Capítulo 5 a abordagem é semelhante ao capítulo anterior, mas refere-se às características de conversação.

Finalmente, no Capítulo 6 é feita uma avaliação geral do trabalho, com a consolidação dos resultados

obtidos e propostas de melhoria dos algoritmos aqui desenvolvidos.

1.4 Símbolos e Abreviaturas

A seguir são apresentados os símbolos e abreviaturas utilizados ao longo do trabalho:

Av	atividade de voz;
cor	coeficiente de autocorrelação do sinal;
DP	estado de deteção primária;
dpic	diferença de picos do sinal;
DPp	desvio padrão dos quadros por pausa;
DPv	desvio padrão dos quadros por surto de voz;
edf	saída obtida pela combinação dos 4 últimos estados de PDF;
En	energia do sinal;
EV	estado de existência de voz;
FDP	função distribuição de probabilidade;
fdp	função densidade de probabilidade;
HO	estado de hangover;
IND	nº de quadros indefinidos;
LAPS	Laboratório de Automação e Processamento de Sinais do Deptº de Engª Elétrica da UFPB;
Mnp	média do nº de quadros por pausa;

Mns	média do nº de quadros por surto de voz;
Mp	duração média das pausas;
Msd	duração média dos sons surdos;
Msil	duração média dos intervalos de silêncio;
Msn	duração média dos sons sonoros;
Msv	duração média dos surtos de voz;
npico	número total de picos do sinal;
NQ	número de quadros;
NSD	número total de surtos de sons surdos;
NSIL	número total de intervalos de silêncio;
NSN	número total de surtos de sons sonoros;
OFF	segmento indicando ausência de voz;
ON	segmento indicando presença de voz;
PDF	parâmetro sinalizador do estado de saída do detetor de voz utilizado para conversação;
PTS	quadros de voz segmentados em 40 amostras (4 ms a uma taxa de amostragem de 10 ms);
SD	nº total de quadros de sons surdos;
SI	estado de silêncio;
SIL	nº total de quadros de silêncio;
SN	nº total de quadros de sons sonoros;
Sp	soma total dos quadros de pausa;
Sv	soma total dos quadros de voz;
TASI	time assignement speech interpolation;
TOsd	taxa de ocorrência dos sons surdos;

TOsil	taxa de ocorrência de intervalos de silêncio;
TOsn	taxa de ocorrência dos sons sonoros;
Tp	duração total das pausas;
TS	taxa de surto de voz média;
Tsd	duração dos sons surdos;
Tsil	duração dos intervalos de silêncio;
TSl	taxa de surto de voz para locução;
Tsn	duração dos sons sonoros;
Tsv	duração total dos surtos de voz;
zcr	taxa de cruzamento por zero do sinal.

CAPITULO 2
CONSIDERAÇÕES SOBRE O SINAL DE VOZ

CAPÍTULO 2

CONSIDERAÇÕES SOBRE O SINAL DE VOZ

2.1 Características Básicas do Sinal de Voz

Os sons da voz são de natureza complexa. O objetivo básico da voz é transmitir idéias, pensamentos, opiniões, sentimentos, de uma pessoa para outra. Desse modo, em sistemas de comunicação é necessário considerar tanto o aspecto da produção da voz por parte do locutor, quanto sua percepção pelo ouvinte [6].

O sinal de voz ocupa a faixa entre 80 Hz e 12 kHz do espectro de freqüências. A freqüência fundamental da voz humana está situada entre 80 Hz e 350 Hz, estando o valor típico para os sons produzidos pelos homens em torno de 120 Hz e para os sons produzidos pelas mulheres em torno de 240 Hz. A faixa dinâmica da voz (diferença entre seus valores máximo e mínimo) varia entre 30 dB e 50 dB [6,7,8].

A figura 2.1 apresenta a forma de onda de um sinal de voz, na qual é possível perceber uma combinação de características inerentes ao processo da fala. Em alguns intervalos, o sinal apresenta níveis elevados de energia

além de uma certa periodicidade, e em outros tem a aparência de um sinal aleatório com níveis de amplitude bastante reduzidos.

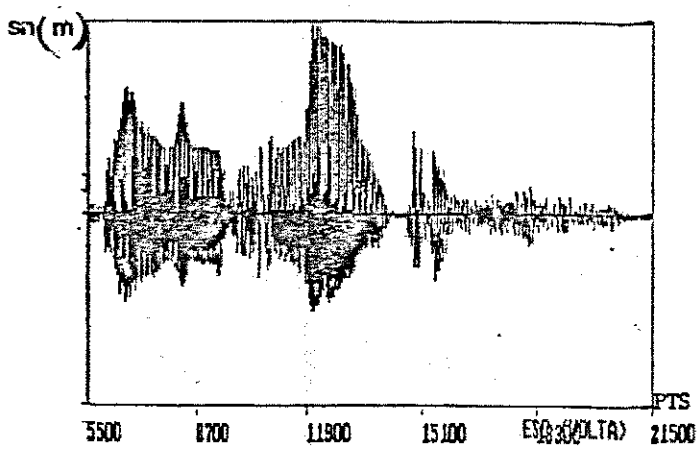


FIGURA 2.1 – Ilustração da forma de onda do sinal de voz, representando a frase "Dias Lentos".

A energia do sinal de voz está concentrada na região de freqüências mais baixas do espectro. A faixa de 500 Hz a 800 Hz centraliza esta concentração. No entanto, mesmo contendo baixos valores de energia, as componentes de freqüência mais altas são importantes pois determinam, em grande parte, a inteligibilidade da voz [9,10].

As freqüências abaixo de 500 Hz contribuem muito pouco para a compreensão da fala, mas tem um efeito importante na naturalidade da voz reproduzida. Em sistemas telefônicos a compreensão é fundamental e a naturalidade da voz é secundária, o que justifica o uso de larguras de banda mais estreitas que em sistemas de radiodifusão, por exemplo, onde a naturalidade da voz é prioritária [9].

2.2 Fisiologia da Produção da Fala

O aparelho fonador é constituído por três elementos principais: os pulmões, a laringe e a região vocal ou trato vocal. Físicamente, é composto por um tubo acústico não uniforme que se estende desde a glote, que é a abertura existente entre as cordas vocais, até os lábios. A extensão desse tubo está em torno de 17 cm a 20 cm em um adulto masculino e a área de sua secção transversal varia, de acordo com a posição dos articuladores (lábios, língua, úvula, maxilar, garganta, nariz, dentes), entre 0 cm^2 e 20 cm^2 [8].

Os pulmões podem ser considerados como a fonte de energia, pressionando o ar, anteriormente inspirado, para a laringe e a região vocal [8].

A laringe consiste de três cartilagens suportando as cordas vocais que, em combinação com a glote, constituem uma abertura de tamanho variável, assemelhando-se a dois lábios que abrem e fecham, através da qual o ar vindo dos pulmões, flui, produzindo ondas de pressão ou uma turbulência de ar, que excitarão o trato vocal [8].

O trato vocal consiste da faringe, do trato oral e do aparelho nasal. O trato oral, por sua vez, é formado pelos seguintes órgãos articulatórios: o véu palatino e a úvula, o palato duro, a língua, a arcada dentária e os lábios. A secção transversal do tubo formado pelo trato

oral e pela faringe, varia com a posição desses órgãos articulatórios. O aparelho nasal, no entanto, possui secção transversal fixa, situando-se em paralelo com o trato oral, sendo ativado ou desativado pelo deslocamento da úvula [8,11].

A figura 2.2 apresenta um diagrama esquemàtico do mecanismo vocal humano [11].

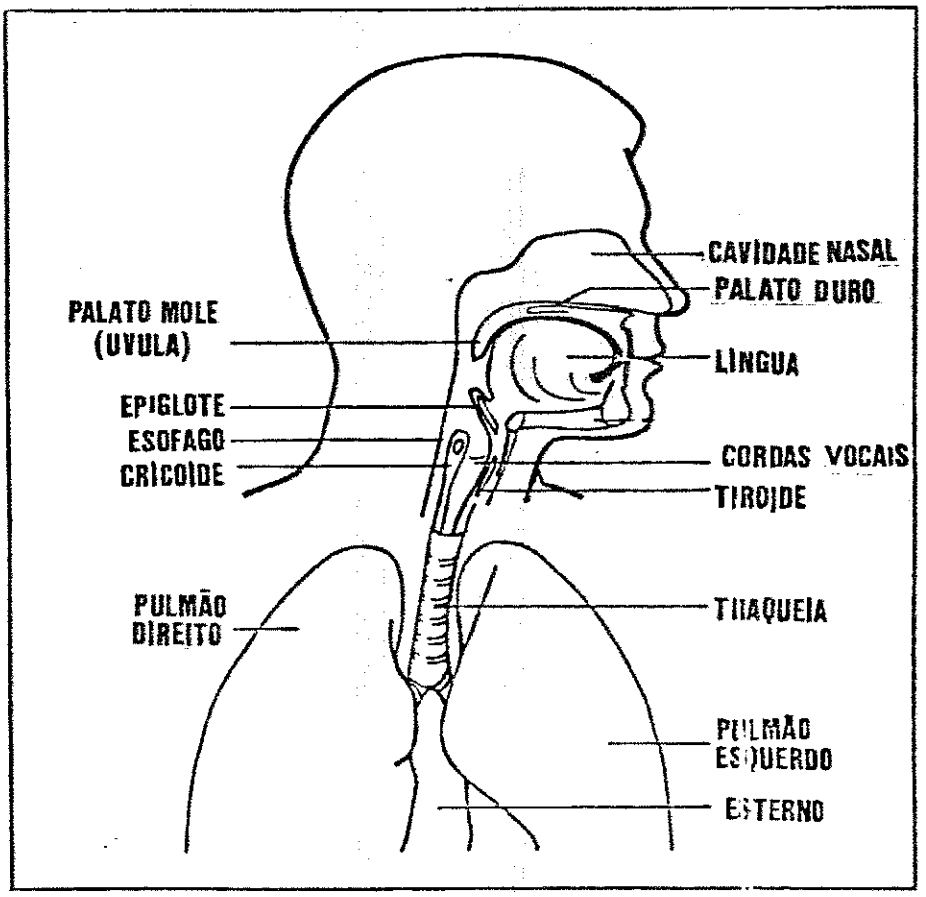


FIGURA 2.2 – Diagrama Esquemático do Aparelho Fonador.

A figura 2.3 representa o modelo acústico do mecanismo de produção da voz humana [12].

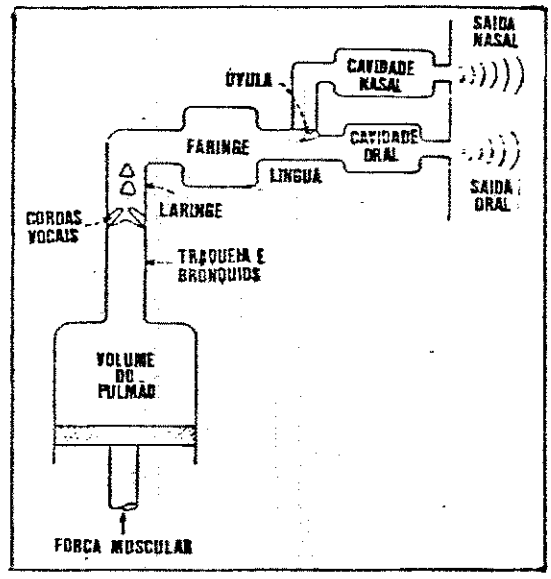


FIGURA 2.3 - Modelo Acústico do Mecanismo de Produção da Voz Humana.

O fluxo de ar dos pulmões através do aparelho vocal segue um caminho que depende da posição dos vários órgãos articulatórios, responsáveis pela considerável alteração nas cavidades do aparelho fonador. Basicamente, o ar vindo dos pulmões flui pela faringe, pela língua e, dependendo da posição da úvula, é irradiado, sob a forma de sinal acústico, através da cavidade oral, da cavidade nasal ou de ambas. A posição assumida pelos diferentes órgãos articulatórios determina a diferenciação dos fonemas da fala [8].

A voz é, portanto, o produto de uma excitação original, modificada, a seguir, pelo trato vocal [8].

2.3 Classificação dos Sons da Voz

A classificação dos sons da voz humana é um trabalho complexo pois depende da língua falada em cada país. Os sons da fala dependem do modo de excitação do trato vocal humano e, de acordo com as características físicas desse aparelho, os diferentes fonemas produzidos classificam-se de uma forma abrangente e rica quanto aos modos e aos pontos de articulação. Neste trabalho é estabelecida uma classificação mais geral e, conseqüentemente, mais simples para os sons, uma vez que o objetivo almejado é bastante específico. Desse modo, três grupos de sons básicos são apresentados a seguir:

- sons **sonoros**;
- sons **surdos**;
- sons **oclusivos**.

Os sons **sonoros** são componentes da voz com grande quantidade de energia. São produzidos através da combinação correta que a pressão do ar e a tensão muscular causam às cordas vocais, fazendo-as vibrar como um oscilador de relaxação, modulando o ar em pulsos discretos e quase-periódicos. A quase-periodicidade desses pulsos define uma *frequência fundamental fo*, cujo inverso (*1/fo*) é denominado *período de pitch* [13]. No caso dos sons

sonoros, são produzidos formantes ou freqüências naturais que são as freqüências de ressonância do tubo do trato vocal. Os formantes dependem do formato e das dimensões do aparelho vocal, sendo que cada formato apresenta 4 formantes que caracterizam o som emitido [8,10,11,12]. Como exemplo de sons sonoros temos os fonemas vocálicos /a/, /e/, /i/, /o/, /u/ [1,14]. A figura 2.4 mostra um exemplo de um sinal sonoro, correspondente ao fonema /a/ na palavra "AJUSTE".

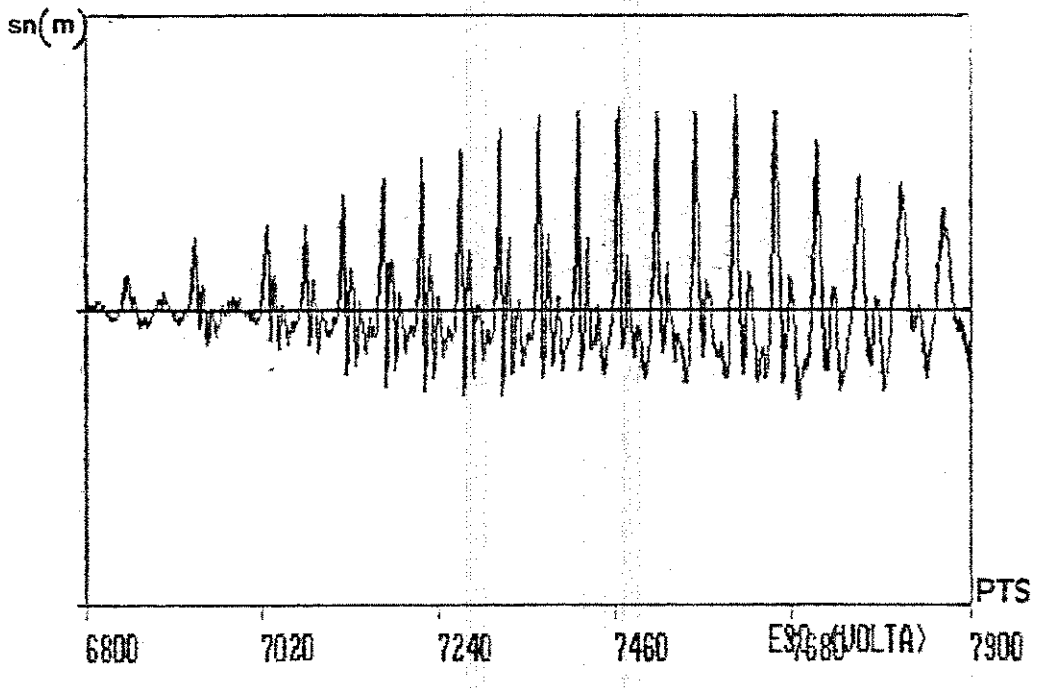


FIGURA 2.4 - Representação de um som sonoro da voz constituído pelo fonema /a/.

Outras componentes possuem menos energia que os sons sonoros, sendo denominadas de sons surdos. Esses sons são produzidos mantendo-se as cordas vocais abertas, não havendo qualquer vibração. É estabelecida uma compressão

efetuada pelos articuladores, em algum ponto do trato vocal, normalmente próximo ao final da boca, forçando o ar, a uma alta velocidade, de modo a produzir turbulência. Esse processo cria uma fonte de ruído de banda larga, que excita o trato vocal. Os sons surdos possuem as características de um sinal de ruído. Como exemplo de sons surdos pode-se citar os fonemas consonantais fricativos como o /f/ e o /s/ [1,14,15]. A figura 2.5 apresenta um segmento de sinal de voz constituído por um som surdo, correspondente ao fonema /s/ da palavra "AJUSTE". Nota-se que a forma de onda apresenta características de um sinal aperiódico.

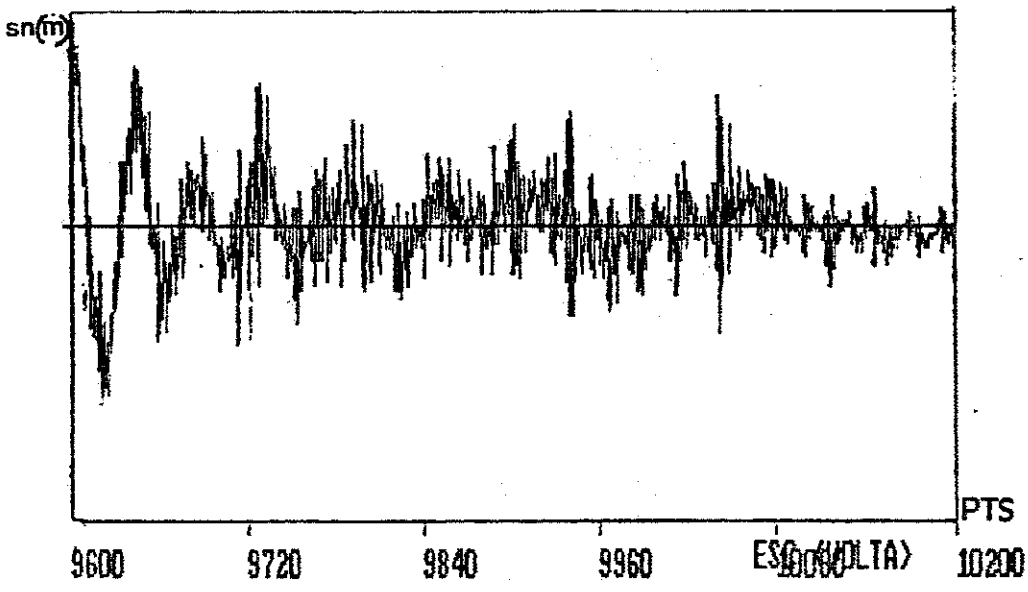


FIGURA 2.5 - Representação de um som surdo da voz constituído pelo fonema /s/.

O terceiro tipo encontrado na classificação apresentada é representado pelos sons oclusivos, obtidos a partir de um completo fechamento do trato vocal, usualmente próximo à sua parte frontal, estabelecendo uma pressão

por trás desse fechamento, liberado completamente logo a seguir. O período de fechamento do aparelho vocal gera um intervalo com níveis de energia bastante reduzidos. Alguns exemplos de fonemas consonantais oclusivos são o /p/, o /t/, o /b/, o /d/ e o /g/. A figura 2.5 apresenta um segmento de voz contendo um som oclusivo, correspondente à sílaba "te" na palavra "AJUSTE" [14].

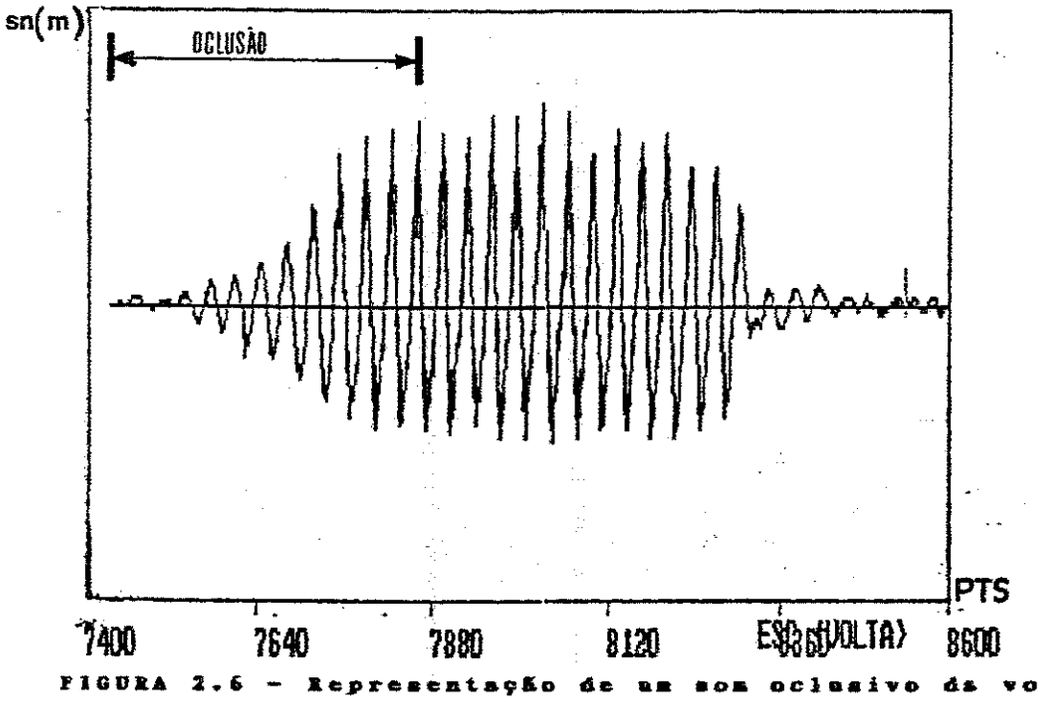


FIGURA 2.6 - Representação de um som oclusivo da voz constituído pela sílaba "te" da palavra "AJUSTE".

As regiões nas quais encontram-se os sons sonoros, são mais facilmente identificadas, devido à sua alta energia e à periodicidade da forma de onda do sinal de voz.

Os sons surdos possuem baixo valor de energia o que, em muitas ocasiões, dificulta sua deteção. Além da característica anterior, os sons surdos ocupam o espectro

de alta freqüência. São apresentados, ao longo deste trabalho, testes que possibilitam caracterizar os sons surdos, evitando-se confundi-los com o sinal de ruído.

Os sons oclusivos caracterizam-se por um alto valor do parâmetro variação de energia, no momento da liberação do ar aprisionado pelo fechamento do trato vocal.

2.4 Modelo Digital para o Mecanismo de Produção da Fala

É possível estabelecer um modelo para representar o mecanismo de produção da fala, considerando-se que as fontes de excitação e o sistema linear variante no tempo, que constitui o trato vocal, são independentes [10,12]. O modelo da forma de onda do sinal de voz é representado, considerando-se a resposta do sistema linear a essa excitação [16]. A geração de sons sonoros, surdos e oclusivos está relacionada com a forma de excitação na glote, no ponto de articulação e no modo de articulação do trato vocal [12,14]. Pode-se considerar que a geração desses sons é feita através de dois geradores distintos, sendo que um deles fornece um trem de impulsos e o outro valores aleatórios. A figura 2.7 apresenta o modelo digital para produção da voz.

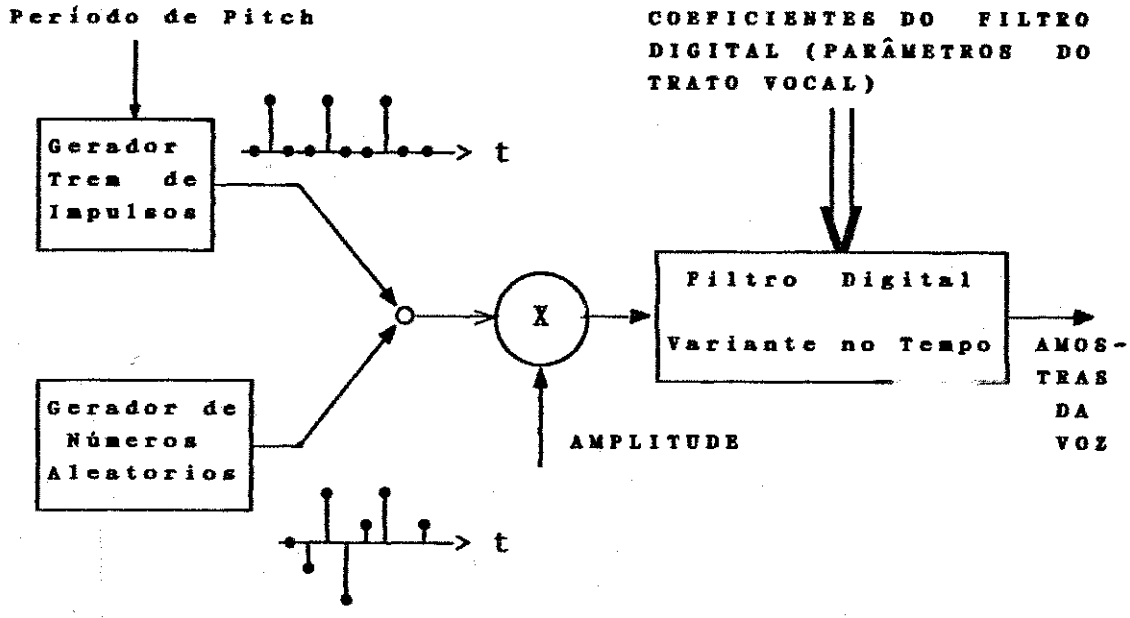


FIGURA 2.7 - Modelo Digital para Produção da Voz.

O gerador de trem de impulsos excita o filtro digital variante no tempo com impulsos quase-periódicos, caracterizando um som sonoro [17]. O espaçamento entre os impulsos corresponde ao período de *pitch* da excitação glotal, cujas características devem estar incluídas no filtro digital. No caso dos sons contínuos como as vogais, os parâmetros variam lentamente e a representação modela muito bem o que ocorre na prática [1,10,12].

O gerador de números aleatórios simula a forma de onda característica dos sons surdos produzidos tanto por turbulência quanto por oclusão, que excitarão, a seguir, o filtro digital [10,12].

O filtro digital variante no tempo é um sistema linear que simula o trato vocal, possuindo características

definidas durante o segmento de voz em anàlise. Em média, esse segmento dura em torno de 10 ms. Os coeficientes do filtro digital correspondem, portanto, aos parâmetros do trato vocal [10].

2.5 Intervalos de Atividade e Inatividade de Voz

Em uma conversação telefônica, normalmente um interlocutor fala enquanto outro escuta. Conseqüentemente, em uma direção o canal de transmissão é deixado ocioso, a menos que técnicas adequadas permitam a ocupação desse intervalo de tempo com dados ou sinais de voz de outras conversações. Essas técnicas tem o salutar efeito de aumentar a capacidade de transmissão do sistema [18,19].

Pequenas pausas resultam do processo natural da fala, ocasionadas por diminutas interrupções no fluxo da voz durante as sentenças, frases, palavras ou sílabas. Pequenas hesitações do locutor, que interrompe brevemente sua fala para pensar, também causam pausas. Esses intervalos de silêncio são, na maioria das vezes, tão curtos que o ouvinte praticamente não os percebe. Uma análise adequada, a partir da deteção da voz por meio de técnicas a serem abordadas mais à frente, possibilita a conclusão de que o número de pausas originadas pelas razões expostas anteriormente, é considerável, justificando

a utilização simultânea de um único canal de transmissão por vários usuários [18,19,20,21].

Constatada essa possibilidade, várias técnicas visando o gerenciamento do meio de transmissão têm sido desenvolvidas, com a finalidade de ocupá-lo eficientemente [22,23]. Essas técnicas não serão discutidas com detalhes, porque fogem do objetivo principal deste trabalho.

Estudos estatísticos semelhantes também podem ser feitos utilizando-se não mais uma conversação telefônica e sim um único locutor, com o objetivo de empregar os resultados obtidos, na otimização de técnicas de processamento de sinais de voz. Essa diferenciação é necessária pois as estatísticas de uma conversação telefônica não se assemelham às da fala isolada de um locutor.

2.5.1 Intervalos Básicos da Fala

Durante a fala existem períodos em que o locutor está, efetivamente, emitindo sons e outros em que êle permanece em silêncio ou produz pequenas pausas intrínsecas ao ato de falar.

Assim, é possível definir dois intervalos básicos durante a fala de um locutor:

- **Intervalo de Atividade de Voz,** em que o locutor está falando;
- **Intervalo de Inatividade de Voz,** em que o locutor está em silêncio.

Esses intervalos podem ser subdivididos, para que o processo da fala seja melhor caracterizado :

- **Intervalo de Atividade de Voz:**

  - **Surto de Voz:** intervalo de tempo em que a energia da voz está presente;
  - **Surto de Ruído:** intervalo de tempo em que a energia do ruído está presente, durante o período em que o locutor está falando;
  - **Pausa:** intervalo de tempo entre sentenças, palavras, frases ou sílabas, em que há ausência da voz; normalmente são pouco perceptíveis para o ouvinte.

- **Intervalo de Inatividade de Voz:**

  - **Surto de Ruído:** intervalo de tempo em que a energia do ruído está presente, durante o período em que o locutor não está falando;
  - **Silêncio:** intervalo de tempo em que o locutor não está falando.

Neste trabalho, todos os intervalos de voz estão referidos a estas definições apresentadas anteriormente, a menos das observações indicadas no Capítulo 4.

2.6 Estatísticas do Sinal de Voz e Suas Finalidades

Os sinais de voz são, por natureza, aleatórios e não estacionários, com significativas diferenças de níveis de amplitude e conteúdo de freqüência entre seus segmentos. Mas os segmentos de voz a curtos intervalos de tempo, podem ser considerados como estacionários, o que significa que suas propriedades estatísticas não variam com o tempo. Em geral, para que se considere o sinal de voz com características estacionárias, devem ser usados segmentos de voz inferiores a 32 ms [1,7,13,24,25].

Em termos da função distribuição de probabilidade, o sinal de voz pode ser modelado adequadamente por uma distribuição gama ou laplaciana para longos intervalos de tempo, ou por uma distribuição gaussiana para curtos intervalos de tempo [7,13].

Na figura 2.8 são apresentadas as funções densidade de probabilidade para o sinal de voz, a curtos intervalos de tempo (funções gaussiana ou normal) e

a longos intervalos de tempo (funções gama ou laplaciana) [7,13].

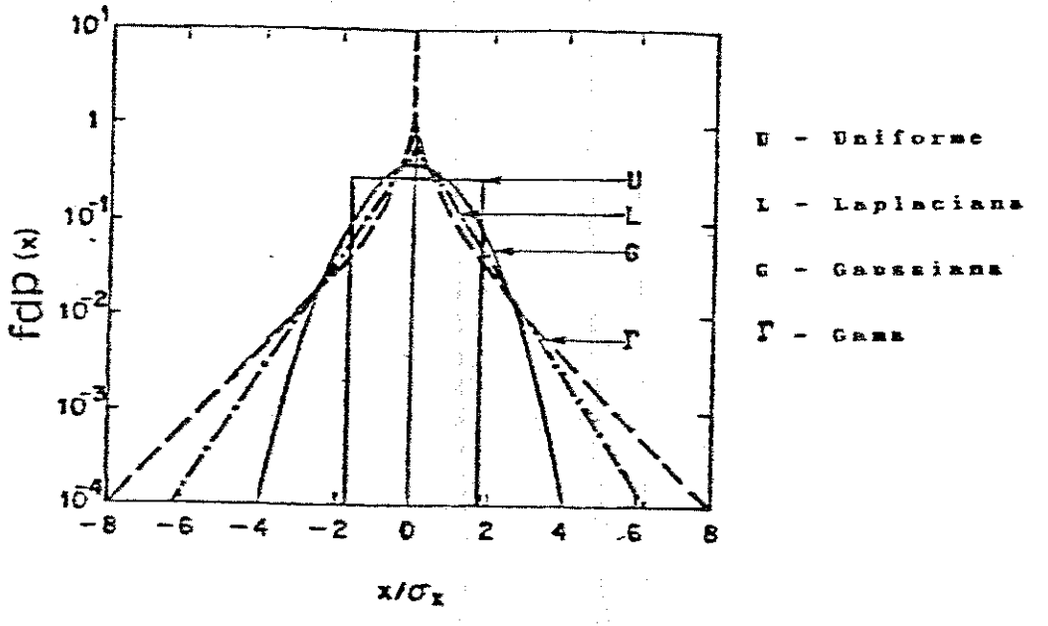


FIGURA 2.2 - Funções densidade de probabilidade para o sinal de voz a curtos e longos intervalos de tempo.

Sendo o sinal de voz descrito através de um processo estocástico, a obtenção das estatísticas desse sinal torna-se importante para a formulação de modelos de tráfego de voz, a partir da determinação de seu comportamento em uma conversação telefônica ou em uma outra aplicação de processamento de voz.

2.7 Segmentação do Sinal de Voz

A segmentação consiste em particionar o sinal de voz em segmentos, selecionados por janelas ou quadros, de duração perfeitamente definida. Cada segmento de voz é processado, a seguir, isoladamente da forma de onda original. São realizados tantos processamentos quanto necessários, para que seja possível a análise do sinal de voz como um todo. Esse método também é denominado de análise do sinal de voz a curtos intervalos de tempo [1].

O resultado do processamento do sinal de voz a curtos intervalos de tempo gera, em cada segmento analisado, um único valor ou um conjunto de valores. Esse processamento produz uma nova seqüência dependente do tempo, também utilizada como uma representação do sinal de voz [1].

A segmentação do sinal de voz pode ser obtida a partir da seguinte expressão:

$$Yn = \sum_{m=-\infty}^{\infty} \{T[sn(m)] \cdot W(n-m)\} \qquad (2.1)$$

em que o sinal de voz sn(m) é submetido a uma transformação T[sn(m)], linear ou não linear. A seqüência resultante é multiplicada pela função janela, posicionada a cada instante correspondente ao índice da enésima

amostra. Normalmente a seqüência da janela é de duração finita [1].

A expressão 2.1 representa a multiplicação da janela W(m) com a seqüência T[sn(m)]. A janela desliza ao longo da seqüência de valores representativos do sinal de voz, selecionando os intervalos envolvidos no processamento. Quanto menor o número de amostras N dentro do quadro, mais simples torna-se localizar a transição de um som surdo para sonoro, e vice-versa [10].

Alguns dos intervalos mais utilizados para as janelas tem os valores de 4 ms, 10 ms e 25 ms, como será visto no ítem 3.1.2.

A equação 2.1 pode ser representada pelo diagrama em blocos apresentado na figura 2.9.

FIGURA 2.9 - Diagrama em Blocos Representativo da Obtenção do Sinal de Voz Segmentado.

Os dois tipos mais comuns utilizados para segmentar um sinal, são a janela retangular e a janela de Hamming, cujas características são apresentadas a seguir [1]:

a) *janela* retangular:

$$W(n) = 1 \qquad 0 <= n <= N-1$$

$$W(n) = 0 \qquad \text{fora desse intervalo} \tag{2.2}$$

A janela retangular corresponde a multiplicar todas as amostras da seqüência contida no intervalo entre 0 e N-1 por 1, e multiplicar por 0 as demais amostras do sinal.

b) *janela* de Hamming:

$$W(n) = \{0,54-0,46.\cos[2\pi n/(N-1)]\} \quad 0 <= n <= N-1$$

$$W(n) = 0 \qquad \text{fora desse intervalo} \tag{2.3}$$

Comparativamente, pode-se considerar que a largura de banda da janela de Hamming é, aproximadamente, duas vezes maior que a largura de banda da janela retangular, de mesma duração. A atenuação da efetuada pela janela de Hamming fora da faixa de passagem é maior que a da janela retangular [1].

Neste trabalho será utilizada a janela retangular com duração de 4 ms, com o objetivo de aumentar a precisão na classificação do segmento em análise, pois a utilização de janelas menores possibilita a obtenção de detalhes exatos da forma de onda do sinal de voz [1]. A

figura 2.10 apresenta a divisão da seqüência de amostras do sinal de voz em segmentos para análise a curtos intervalos de tempo. Conforme será visto nos capítulos 4 e 5, a freqüência de amostragem utilizada para digitalização do sinal de voz é de 10 kHz, o que determina o processamento de 40 amostras por quadro (quadro de 4 ms / período de amostragem = 0,1 ms).

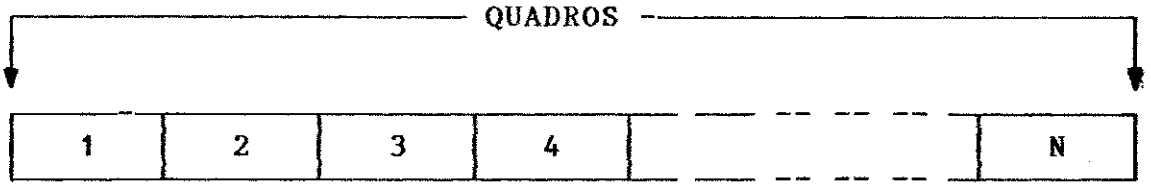


FIGURA 2.10 - Seqüência de Amostras por Quadro. Cada quadro tem a duração de 4 ms. Se a Freqüência de Amostragem é de 10 kHz, cada quadro possui 40 amostras. Em um intervalo de 1 s, o total será de 250 quadros analisados (N = 250 e 10000 amostras).

2.8 Deteção da Fala. Características ON-OFF do Sinal de Voz

O sinal de voz em uma conversação é composto, basicamente, de intervalos onde são encontrados surtos de energia, significando que o locutor está falando, e intervalos de silêncio e pausas. Dessa forma, pode-se conceber a representação do sinal de voz a partir de sua presença ou ausência em um meio de transmissão.

Um dos problemas fundamentais no processamento da voz é a determinação do início e do fim de uma palavra, discriminando-a do ruído. O aspecto de localizar as regiões que, efetivamente, contenham voz, é importante pois permite que os sistemas processem apenas as regiões que correspondam ao sinal de interesse [26].

Em ambientes de alta relação sinal-ruído, é fácil determinar a energia dos sons de voz mais fracos como as consoantes fricativas, por exemplo. Na maioria das aplicações, no entanto, é difícil de se ter condições ideais e o nível do sinal de voz se aproxima do nível do sinal de ruído [26].

As maiores dificuldades em localizar o início e o fim das palavras ocorrem nas seguintes condições [26]:

- existência de fricativas fracas no início ou no fim das palavras;
- surtos de sons explosivos fracos no início ou no fim das palavras;
- sons nasais no final das palavras;
- fricativas sonoras no final das palavras.

A deteção da presença ou ausência da voz depende do propósito da aplicação. Em telefonia, p. ex., o sinal de voz é detetado a partir de sua passagem por um dispositivo denominado detetor de atividade de voz, cuja finalidade é produzir um sinal digital representativo da forma de onda

original. O detetor transforma o sinal de voz em segmentos de duração perfeitamente definida, de modo que a forma de onda resultante seja composta de intervalos que caracterizam a presença de energia da voz (estados ON) e intervalos que caracterizam sua ausência (estados OFF). Aos intervalos ON é atribuído o nível lógico 1 e aos intervalos OFF é atribuído o nível lógico 0. O sinal resultante na saída do detetor é, portanto, uma seqüência de 1's e 0's [21,27].

Várias técnicas para construção de detetores de atividade de voz têm sido desenvolvidas. Esses detetores operam a partir de um algoritmo de decisão, que atua baseado em medidas de sinal. Na forma mais simples, a ativação do detetor é efetuada tomando-se um parâmetro relevante do sinal de voz, como a amplitude ou a energia por exemplo, e comparando-se seu valor em cada segmento, a um limiar de referência (*threshold*). Os valores do parâmetro de medição que estiverem acima do limiar, indicam a existência de surtos de voz e os que se encontrarem abaixo, configuram a presença de silêncio ou pausa. Será visto mais à frente que outras medidas de sinal, também podem ser utilizadas em conjunto para tornar o processo de deteção mais preciso [21,26,27,28].

A determinação dos intervalos de silêncio é um dos aspectos importantes na análise da voz em conversações telefônicas e no processamento de palavras isoladas. A identificação dos intervalos de silêncio existentes em uma

conversação telefônica efetuada em um meio de transmissão, permite a utilização de interpoladores de voz, cuja finalidade é aumentar o número de canais de voz transportados, aproveitando esses intervalos. Quando o sinal de voz é detetado e o canal de transmissão está ocupado, algumas técnicas desenvolvidas para o gerenciamento da rede, garantem a chegada da voz ao seu destino. Uma dessas técnicas utiliza o atraso na transmissão do surto de voz detetado, a partir de sua armazenagem transitória em um *buffer*, até a disponibilidade do canal. Outra técnica utiliza o corte de parte do sinal de voz detetado, seja no início, no meio ou no fim do surto, descartando esses fragmentos até que seja possível a transmissão do sinal desejado [28].

Em aplicações onde se utiliza palavras isoladas, a deteção do silêncio auxilia na identificação do início e do fim de uma região que, efetivamente, contenha a fala. Isto é importante para evitar o processamento em regiões do sinal, que não sejam de interesse.

A escolha adequada do valor de limiar é um ponto fundamental na determinação da eficiência e da sensibilidade do detetor de voz. Ruídos impulsivos de curta duração gerados pelo manuseio do telefone, distúrbios elétricos ou a gravação da fala em uma fita cassete, produzem surtos de energia capazes de acionar, indesejavelmente, o detetor de voz. Por outro lado, pequenas hesitações do locutor ou diminutas pausas

intersilábicas, na maioria das vezes imperceptíveis para o ouvinte, e alguns sons fricativos, podem ter energia inferior ao valor de limiar, não acionando o detetor e caracterizando um intervalo OFF em sua saída quando, na verdade, existe um segmento de voz. Estas observações sugerem os cuidados necessários na determinação do valor do limiar. Este não deve ser baixo o suficiente para permitir que sinais de ruído acionem o detetor, nem elevado a ponto de eliminar componentes importantes do sinal de voz [21,27].

No caso dos eventos de conversação, o sinal ON-OFF resultante da deteção efetuada pelo algoritmo de decisão, deve sofrer correções adequadas, efetuadas por um algoritmo corretor, que elimine os surtos de energia provenientes de ruídos impulsivos e preencha as lacunas originadas a partir de pequenas hesitações do locutor, baixo nível de sinal de voz ou pausas intersilábicas [21,26,27].

Nos eventos de locução, será visto que a aplicação desejada não exige a necessidade do algoritmo corretor.

Neste trabalho, serão utilizados outros parâmetros temporais, além da energia, como medidas de sinal do detetor proposto, além de vários níveis de limiar dos parâmetros envolvidos, buscando obter uma maior eficiência no processo de obtenção do sinal digital.

Será visto nos capítulos 4 e 5, que os detetores de voz desenvolvidos neste trabalho para os eventos de locução e conversação, utilizam estrutura semelhante para medição

do sinal. Diferenças fundamentais, no entanto, são observadas nos algoritmos construídos para as duas aplicações, devido às características específicas de cada um dos eventos. Nos eventos de locução a fala é constituída de palavras isoladas, com curtos intervalos de silêncio entremeando-as, e maior facilidade de reconhecimento de seus limites iniciais e finais. Nos eventos de conversação, ou fala contínua, é mais difícil identificar os limites de cada palavra porque existe o efeito da coarticulação, que altera a pronúncia de cada palavra de acordo com sua posição relativa às outras palavras na sentença [29].

No caso das locuções, o detetor, além de indicar a presença ou ausência do sinal de voz no segmento em análise, pode ser desenvolvido de modo a classificar os intervalos que contenham a fala, de acordo com o tipo de fonema empregado.

O conhecimento das estatísticas dos intervalos de presença e ausência da voz, possibilita uma caracterização adequada dos eventos da voz. Em conversações telefônicas, a determinação da duração dos surtos da fala e do silêncio é extremamente útil nos projetos de tráfego de voz e dados. Nas locuções, essas estatísticas podem ser utilizadas na área de comunicação homem/máquina.

CAPÍTULO 3
PARÂMETROS TEMPORAIS DO SINAL DE VOZ E EVENTOS DA VOZ

CAPÍTULO 3

PARÂMETROS TEMPORAIS DO SINAL DE VOZ E EVENTOS DA VOZ

3.1 Parâmetros Temporais do Sinal de Voz

O objetivo básico do processamento digital de sinais de voz, é o de obter uma representação mais conveniente da informação transportada por esse sinal. Um dos aspectos mais importantes na análise de um sinal de voz, consiste em determinar se o segmento do sinal em análise representa um som sonoro, um som surdo, ou indica, simplesmente, sua ausência (silêncio). Apenas a determinação dessas características da fala, possibilita a implementação de inúmeras aplicações em processamento digital de sinais de voz [1].

Neste trabalho, todo processamento utilizado para alcançar os resultados desejados, está baseado em um conjunto de métodos no domínio do tempo. Esses métodos envolvem diretamente a forma de onda do sinal de voz, produzindo informações bastante úteis acerca das características do sinal processado, a despeito da

simplicidade de suas implementações [1].

Alguns exemplos de parâmetros que representam o sinal de voz, em termos de medidas no domínio do tempo, são a energia, a taxa de cruzamento por zero, a função autocorrelação e a magnitude [1]. Esses são os parâmetros temporais mais conhecidos e utilizados em processamento digital de sinais de voz. Nem todos esses parâmetros, no entanto, serão utilizados nos algoritmos desenvolvidos neste trabalho. Além disso, alguns outros parâmetros temporais, menos conhecidos e de maior utilidade para o trabalho proposto, serão empregados.

Na determinação dos eventos da voz para locução, são utilizados os parâmetros temporais energia, taxa de cruzamento por zero, número total de picos, diferença entre o número total de picos, variação de energia e coeficiente de autocorrelação normalizado. No caso da determinação dos eventos da voz para conversação, os parâmetros utilizados são a energia, a taxa de cruzamento por zero e a variação de energia.

3.1.1 Energia a Curtos Intervalos de Tempo

A energia a curtos intervalos de tempo ou segmentar para sinais não estacionários como a voz, pode ser definida

através da seguinte expressão:

$$En = \sum_{m=0}^{N-1} [\, sn(m) \,]^2 \qquad (3.1)$$

onde N é o número de amostras na janela em análise e sn(m) representa o sinal de voz.

A energia é obtida, portanto, simplesmente somando-se os quadrados das amplitudes das N amostras do sinal contido na janela em análise, devendo refletir as variações de amplitude do sinal de voz entre quadros ou janelas [1,30,31].

A figura 3.1 apresenta o diagrama em blocos da representação do cálculo da energia a curtos intervalos de tempo, onde h(n-m) representa a segmentação do sinal de voz a curtos intervalos de tempo [1].

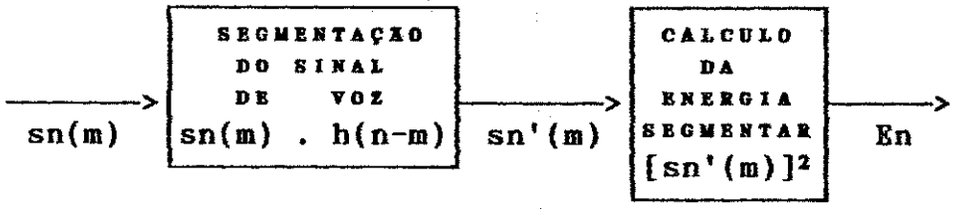


FIGURA 3.1 - Representação do calculo da energia a curtos intervalos de tempo.

A amplitude do sinal de voz varia consideravelmente com o tempo. Considerando-se que a amplitude dos segmentos surdos é muito menor que a amplitude dos segmentos sonoros, a utilização do parâmetro energia tem importância significativa na diferenciação entre os sons surdos e sonoros. Assim, pode-se afirmar que, em geral, o valor de En para os sons surdos é bem menor que para os sons sonoros [10].

A figura 3.2a apresenta um intervalo de voz de 600 ms, representando uma conversação, na qual alternam-se sons surdos e sonoros, além de intervalos de silêncio. A figura 3.2b apresenta a energia segmentar correspondente a esse intervalo. Em alguns intervalos o nível de energia da voz pode estar próximo ao nível de energia do ruído presente no intervalo de silêncio. Uma análise inicial pode estimar que este intervalo corresponde ao som de uma consoante fricativa. Mas é preciso uma abordagem mais precisa, utilizando outros parâmetros temporais, para garantir o tipo de som que corresponde a este intervalo.

Freqüentemente, a energia é maior nos sons surdos que nos intervalos de silêncio, mas, em alguns casos, essa afirmativa não é totalmente correta. Quando o segmento em análise representa um som fricativo, sua energia pode estar muito próxima do nível da energia de ruído, único sinal existente nos intervalos de silêncio, o que pode causar êrros na interpretação do sinal desejado. Outros parâmetros

temporais são utilizados, então, para dirimir essas dúvidas.

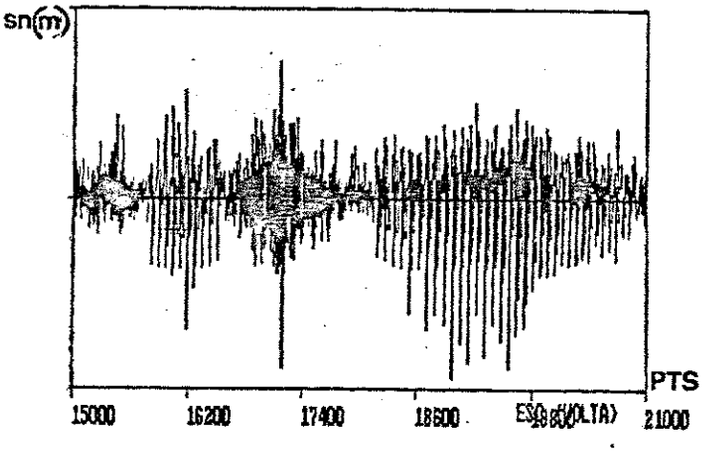


FIGURA 3.2a - Sinal de voz de uma conversação com duração de 600 ms.

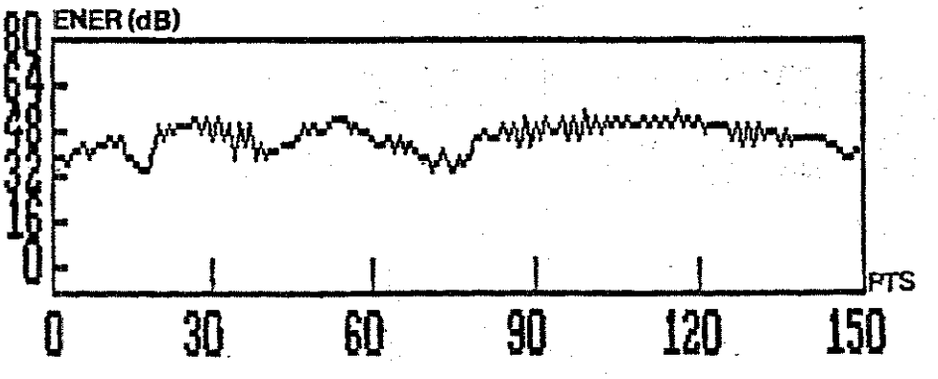


FIGURA 3.2b - Energia segmentar do sinal de voz de uma conversação telefonica com duração de 600 ms.

Um inconveniente na utilização da energia está no aspecto de que este parâmetro é muito sensível a grandes níveis de sinal, uma vez que, em seu cálculo, a amplitude do sinal é elevada ao quadrado, acentuando as grandes

variações de energia entre as amostras. Em aplicações onde este problema torna-se crítico, uma alteração na expressão 3.1 pode contornar esta dificuldade, somando-se os valores absolutos da amplitude do sinal, a voz neste caso, sem elevá-los ao quadrado, obtendo-se, assim, sua magnitude [10].

3.1.2 Taxa de Cruzamento por Zero a Curtos Intervalos de Tempo

A taxa de cruzamento por zero é outro parâmetro bastante utilizado em aplicações de processamento digital de sinais de voz, que utilizam métodos de análise no domínio do tempo. Ela indica o número de vezes que as amostras de um sinal, em um determinado segmento, cruzam o zero tomado como referência, dentro de uma janela.

O cruzamento por zero ocorre entre os instantes de amostragem **m-1** e **m** se for observada a seguinte condição [10]:

$$sgn[sn(m)] \neq sgn[sn(m-1)] \qquad (3.2)$$

em que sgn = 1 se sn(m) for maior que 0 e sgn = -1 se sn(m) for menor que 0. Isto significa que a contagem é efetuada sempre que sucessivas amostras do sinal tiverem diferentes

valores algébricos [1,10,32].

Esta medida pode ser interpretada como uma forma simples de se determinar o conteúdo de freqüência de um sinal. Neste trabalho, será utilizada a expressão a seguir, para determinar a taxa de cruzamento por zero de um sinal de voz [1]:

$$ZCR = \sum_{m=1}^{N-1} | \, sgn[sn(m)] - sgn[sn(m-1)] \, | \qquad (3.3)$$

A figura 3.3 apresenta o diagrama em blocos da representação da taxa de cruzamento por zero.

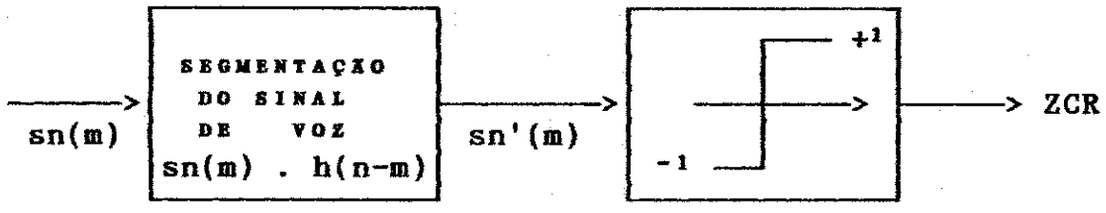


FIGURA 3.3 - Representação do cálculo da taxa de cruzamento por zero a curtos intervalos de tempo.

A energia dos sons sonoros é elevada e tende a concentrar-se nas regiões mais baixas do espectro de freqüência (abaixo de 1 kHz), enquanto que a energia dos sons surdos, de intensidade menor, ocupa as regiões mais altas (acima de 3 kHz). Ao contrário, altas taxas de cruzamento por zero caracterizam os sons surdos e taxas

mais reduzidas indicam a presença de sons sonoros. O ruído tende a produzir taxas intermediárias de cruzamento por zero. Em geral, o número de cruzamentos por zero é bastante eficaz na identificação de consoantes fricativas surdas. A definição do que é baixo ou alto para os valores de energia e taxa de cruzamento por zero é uma questão bastante discutível. Nos capítulos 4 e 5, onde será necessário definir os valores dos parâmetros, esse aspecto será discutido de forma mais abrangente [30,31,32].

A precisão da medida taxa de cruzamento por zero para sinais de banda larga, como o sinal de voz, é muito menor que para os demais sinais. Outro inconveniente está em sua susceptibilidade ao ruído de 60 Hz, ao nível DC offset e ao ruído em geral. Ainda assim, sua aplicação é freqüênte como parâmetro temporal para medidas de sinais de voz. A utilização de filtros passa-faixa e de altas taxas de amostragem no processo de digitalização do sinal de voz, garante uma boa resolução para as medidas de cruzamento por zero [1,10].

A figura 3.4 apresenta as medidas de cruzamento por zero segmentar obtidas a partir de um intervalo de voz com duração de 600 ms, extraídas de uma conversação telefônica simulada. Neste caso misturam-se sons surdos, sonoros e, possívelmente, ruídos originados a partir de pequenos intervalos de silêncio, ao longo da representação. À primeira vista, não é possível diferenciar com precisão os diferentes sons envolvidos. Mas a utilização dessas

medidas em conjunto com outros parâmetros, permitem decidir, com considerável precisão, o tipo de segmento em análise [10].

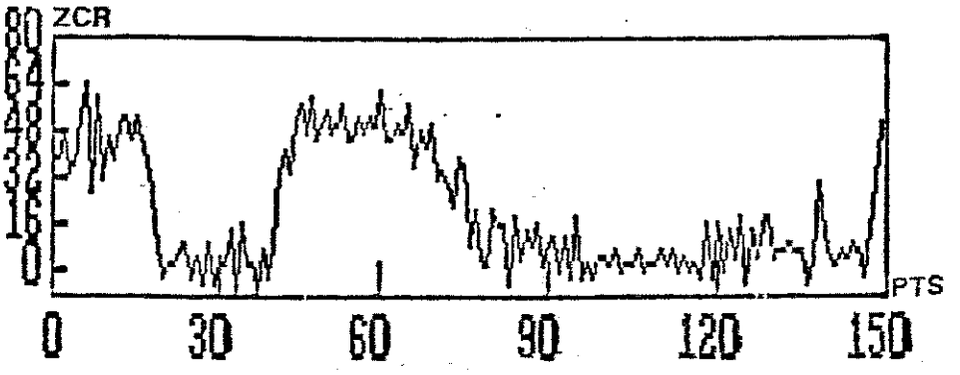


FIGURA 3.4 - Taxa de cruzamento por zero segmentar de um sinal de voz obtido a partir de uma conversação.

Em resumo pode-se atribuir as seguintes propriedades da taxa de cruzamento por zero, relativas ao sinal de voz [1]:

- sinais com conteúdo harmônico de alta freqüência (como os sons surdos) possuem altos valores de ZCR;
- sinais com conteúdo harmônico de baixa freqüência (como os sons sonoros) possuem baixos valores de ZCR;
- há forte correlação entre a taxa de cruzamento por zero e a distribuição de energia, com a freqüência;
- é preciso cuidado ao analisar os sinais em

função das afirmativas anteriores, pois as definições do que sejam valores altos ou baixos de ZCR, são um tanto imprecisas;

- em geral, os sons sonoros apresentam valores de ZCR entre 0 e 30, para quadros de 10 ms [33], com média em torno de 14 [1] ou 15 [34], ambos para um quadro de 10 ms e os sons surdos encontram-se, típicamente, na faixa de 10 a 100 [33], com média entre 48 [34] e 49 [1];
- existem casos em que a ZCR dos sons surdos pode ser tão baixa quanto a dos sons sonoros e vice-versa;
- os valores de ZCR para o ruído variam em função do ambiente, mas como o conteúdo harmônico do sinal concentra-se nas regiões de freqüências baixas e médias da banda de 0 a 5 kHz, os valores de ZCR para o ruído são menores que os dos sons surdos e comparáveis aos dos sons sonoros [33];
- alguns autores consideram que ZCR, para o ruído, está em torno de 16 para um quadro de 4 ms [35], outros estabeleçem o intervalo de 13 a 40 para um quadro de 25 ms [36].

3.1.3 Número Total de Picos

O número total de picos por quadro não é um parâmetro tradicional na análise dos sinais de voz, mas é útil na determinação de sons surdos como as consoantes fricativas de pequena intensidade [36]. Este parâmetro mede o número de picos encontrados dentro do intervalo de voz em análise.

A expressão utilizada para determinar o número total de picos, é apresentada a seguir:

$$npico = picpos + picneg \tag{3.4}$$

em que

$$picpos = picpos + 1 \ , \tag{3.5}$$

sempre que

$$\{[sn(m) >= 0] \ \underline{E} \ [sn(m) >= sn(m-1)] \ \underline{E} \ [sn(m) > sn(m+1)]\} \tag{3.6}$$

e

$$picneg = picneg + 1 \ , \tag{3.7}$$

sempre que

$$\{[sn(m) < 0] \ \underline{E} \ [sn(m) <= sn(m-1)] \ \underline{E} \ [sn(m) < sn(m+1)]\} \tag{3.8}$$

3.1.4 Diferença Entre o Número de Picos

A diferença entre o número de picos por quadro é outro parâmetro que não é tradicionalmente utilizado em processamento de voz. Este tipo de medida foi proposto para identificar os sons fricativos em relação às vogais de pequena intensidade, dadas as limitações de freqüência impostas ao sinal de voz, no processo de digitalização [36].

A expressão que determina a diferença entre o número de picos é apresentada a seguir:

$$dpic = picpos - picneg \qquad (3.9)$$

As expressões anteriores de picpos e picneg, também são válidas neste item.

3.1.5 Variação da Energia a Curtos Intervalos de Tempo

Foi observado na seção 2.1, que os sinais de voz são não estacionários e, à exceção do ruído impulsivo, o ruído ambiente é estacionário. Considerando-se que as variações de energia a curtos intervalos de tempo de sucessivos

blocos do sinal de voz, podem ser muito grandes, pode-se determinar que, caso a expressão 3.10 a seguir, seja verdadeira, o segmento em análise contêm voz [35]. No caso das locuções este parâmetro será utilizado para determinar os sons sonoros e, no caso das conversações, o estado ON ou intervalo de voz.

$$[\ En(m) / En(m-1)\ ] >= NS \qquad (3.10)$$

em que o valor de NS é definido nos capítulos 4 e 5 e depende da aplicação utilizada.

3.1.6 Coeficiente de Autocorrelação Normalizado

Este parâmetro é bastante útil na distinção entre os sons surdos e sonoros. Seus valores variam entre +1 e -1. Os sons sonoros são altamente correlacionados devido à concentração da energia do sinal que os constitui, nas baixas freqüências do espectro. Com isso, os valores do coeficiente de autocorrelação normalizado, para os sons sonoros, é muito próximo da unidade.

Para os sons surdos, o valor desse parâmetro aproxima-se de zero. Os valores típicos para os intervalos de silêncio, variam com o ambiente, mas encontram-se entre os números obtidos para os sons surdos e sonoros [33].

A expressão 3.11 é utilizada para determinar o valor do coeficiente de autocorrelação normalizado [37].

$$COR = \frac{\sum_{m=1}^{N} [sn(m).sn(m-1)]}{\left[\{ \sum_{m=1}^{N} [sn^2(m)] \} . \{ \sum_{m=0}^{N-1} [sn^2(m)] \} \right]^{1/2}} \quad (3.11)$$

3.2 Eventos da Voz

O sinal de voz pode ser caracterizado a partir de sua presença ou ausência em um meio no qual esteja sendo processado. Isto leva a definir dois eventos de voz básicos: o surto de voz e a pausa [20].

Neste trabalho serão consideradas duas situações distintas para definir os eventos da voz, observando-se que, em cada uma dessas aplicações, os eventos de voz diferem:

- locução individual;
- conversação telefônica.

3.2.1 Eventos da Voz para Locução Individual

Nas aplicações em que apenas um locutor fala, definidas como **locução individual**, os eventos de voz utilizados neste trabalho são os seguintes:

- **Número Total de Surtos de Sons Sonoros, Surdos e de Intervalos de Silêncio;**
- **Número Total de Quadros e de Quadros Sonoros, Surdos e de Silêncio;**
- **Tempo Total do Teste e Tempo dos Sons Sonoros, Surdos e dos Intervalos de Silêncio;**
- **Taxas de Surto de Voz e de Ocorrência dos Sons Sonoros, Surdos e dos Intervalos de Silêncio;**
- **Duração Média dos Sons Sonoros, Surdos e dos Intervalos de Silêncio.**

3.2.2 Eventos da Voz para Conversação

No caso das conversações telefônicas simuladas, além dos Surtos de Voz e dos Intervalos de Pausa, são determinados os seguintes eventos [18]:

- **Taxa de Surto de Voz Média (TS):**

em que

$$TS = \frac{N}{T} \qquad (3.12)$$

onde

N indica o número total de surtos de voz;

T indica o tempo total do teste;

TS indica o ciclo de repetição dos surtos de voz.

- **Duração Média das Pausas (Mp):**

em que

$$Mp = \frac{\sum Si}{P} \qquad (3.13)$$

e

$\sum$ Si = somatório dos períodos de tempo ocupados por pausas no intervalo de observação;

P é o número total de períodos de pausa no intervalo de observação.

- Duração Média do Surto de Voz (Msv):

em que

$$Msv = \frac{\sum Vi}{Nv} \qquad (3.14)$$

e

$\sum Vi$ = somatório dos períodos de tempo ocupados por surtos de voz no intervalo de observação;

Nv é o número total de períodos de voz no intervalo de observação.

Seja a seguinte expressão:

$$Msv + Mp = \frac{\sum Vi}{Nv} + \frac{\sum Si}{P} \qquad (3.15)$$

Notando-se que $\sum Vi + \sum Si = T$, em que T é o intervalo de observação, é possível obter o seguinte resultado, a partir da expressão 3.15, considerando-se que, no limite, Nv = P = N:

$$Msv + Mp = \frac{T}{N} \qquad (3.16)$$

Da equação (3.12) sabe-se que:

$$TS = \frac{N}{T} \quad \text{e, portanto, } T = \frac{N}{TS} \qquad (3.17)$$

Substituindo a expressão (3.17) em (3.16), resulta:

$$Msv + Mp = \frac{1}{TS} \qquad (3.18)$$

A expressão (3.18) também pode ser representada da seguinte forma:

$$Msv = \frac{1}{TS} - Mp \qquad (3.19)$$

A figura 3.5 apresenta o significado dos eventos definidos anteriormente.

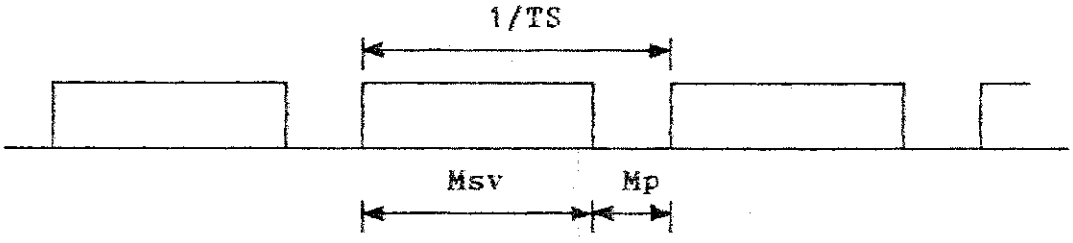


FIGURA 3.5 - Significado físico dos eventos para conversação telefônica.

- Atividade de Voz (Av):

em que

$$Av = \frac{Msv}{Msv + Mp} \qquad (3.20)$$

Outros eventos de voz para conversação telefônica podem ser definidos, além dos que serão utilizados neste trabalho [27]. Alguns deles são apresentados a seguir:

- Dupla conversação;
- Silêncio mútuo;
- Intervalos de silêncio alternados;
- Surtos de voz após interrupção;
- Surtos de voz antes da interrupção.

CAPITULO 4
ANALISE DA LOCUÇÃO

CAPITULO 4

ANALISE DA LOCUÇÃO

4.1 Características da Locução

Normalmente, quando o ser humano fala, está dirigindo-se a outra pessoa. A evolução tecnológica, no entanto, está possibilitando um outro tipo de comunicação, até então apenas imaginado em filmes de ficção científica ou em livros escritos por futurólogos. É a comunicação homem/máquina, que vem sendo perseguida pelos pesquisadores há algumas décadas, mas que apenas recentemente tem proporcionado resultados satisfatórios.

A utilização do têrmo locução resulta do fato de que não é estabelecida uma conversação entre dois ou mais interlocutores. Neste caso, apenas um locutor emite sua fala e, pode-se supor, ela é captada por um ser humano ou uma máquina.

Neste capítulo não haverá preocupação com o elemento receptor da informação, mas com as características da locução emitida e sua deteção. Neste trabalho a locução é limitada a algumas poucas palavras, constituídas de fonemas

representativos da língua portuguesa, criando-se, assim, uma frase considerada como referência.

A frase utilizada é **AJUSTE DE TEMPO**. Na seção seguinte serão apresentadas as razões para sua escolha. Essa frase é pronunciada por diversos locutores, masculinos e femininos, sendo gravada em uma fita cassete ou falada diretamente em um microfone. A digitalização do sinal que representa essa frase se dá conforme descrito na seção 4.1.3. Para o processamento é utilizado um detetor de voz, desenvolvido unicamente a partir de parâmetros temporais, cuja finalidade é discriminar os sons sonoros, os sons surdos e os intervalos de silêncio existentes nessa frase. A partir daí podem ser obtidas as médias, desvios padrões e outras estatísticas dos eventos da voz para a locução.

4.1.1 A Língua e a Teoria da Informação

O sinal de voz é uma seqüência de sons, sendo que o arranjo desses sons é governado pelas regras da *linguagem*. Os sistemas de símbolos destinados a transmitir mensagens significativas entre os seres humanos, constitui um *sistema semiótico*. A Teoria da Informação considera o sistema lingüístico como um código que gera mensagens contendo a informação que se deseja transmitir, a partir de uma

fonte (emissor) até um destinatário (receptor) [38]. Toda a mensagem lingüística oral baseia-se na emissão de certos sons pelo aparelho fonador. Um fonema é a menor unidade sonora da fala, constituindo a unidade mínima pertinente de uma dada língua [39,40]. É o som elementar e distintivo que, articulado e combinado com outros, formam as sílabas, os vocábulos e a teia de frases na comunicação oral. Os fonemas são elementos diferenciadores das palavras porque são capazes de diferenciá-las umas das outras [40].

Não se deve confundir letra com fonema. Fonema é som. Letra é o sinal gráfico que representa o som. O ideal seria que a cada fonema correspondesse uma só letra, e vice-versa. Isso não acontece porque o sistema ortográfico não é rigorosamente fonético e ainda está preso à origem das palavras, como no caso em que se escreve a palavra *exame* e não *ezame*, porque este substantivo vem da palavra latina *examen* [40].

Na língua portuguesa os fonemas são classificados da seguinte forma [40]:

- **vogais**: são os fonemas sonoros que chegam livremente ao exterior sem causar ruído;
- **semivogais**: são os fonemas */i/* e */u/* átonos que se unem a uma vogal, formando com esta uma só sílaba.

  Ex.: **vai - andei - ouro - água**;
- **consoantes**: são ruídos provenientes da

resistência que os órgãos bucais opõem à corrente de ar.

As vogais são elementos básicos e indispensáveis para a formação da sílaba. As consoantes e as semivogais só podem formar sílabas com o auxílio das vogais, sendo, portanto, fonemas dependentes. Cada língua opera com um número diferente de elementos dos códigos lingüísticos, daí sua extensa variedade. Teoricamente, existem 3876 fonemas [38]. Na língua portuguesa, desse total, existem apenas 33. Na língua inglesa existem 46. Daí resulta a restrição às regras combinatórias dos códigos lingüísticos, com limitações impostas à combinação entre os fonemas, para que haja manipulação humana desses elementos [38].

É possível estabelecer listas hierárquicas dos fonemas, em função da freqüência de suas utilizações no cotidiano [41]. A frase **AJUSTE DE TEMPO** utilizada como referência para a realização deste trabalho, foi construída a partir da análise de algumas dessas listas. A pesquisa elaborada pelo prof. Luis Carlos Cagliari da Universidade Estadual de Campinas - UNICAMP [39], cujos resultados são reproduzidos na Tabela 4.1, apresenta a Freqüência Relativa dos Fonemas em Portugues, obtida a partir de noticiários falados por estações de TV de São Paulo e do Rio de Janeiro, e foi uma das listas hierárquicas de freqüência de fonemas utilizada. Por medida de simplificação, são apresentados apenas os quinze primeiros fonemas mais

utilizados.

Ordem	Fonema	Freqüência(%)
1.	/a/	12,36
2.	/i/	10,00
3.	/s/	8,76
4.	/u/	6,45
5.	/r/	6,00
6.	/d/	5,60
7.	/t/	5,49
8.	/w/	4,45

Ordem	Fonema	Freqüência(%)
9.	/e/	4,13
10.	/j/	3,75
11.	/k/	3,47
12.	/o/	3,30
13.	/p/	2,82
14.	/m/	2,81
15.	/n/	2,61

TABELA 4.1 - Freqüência de Fonemas para a Língua Portuguesa.

Outro estudo observado foi o dos pesquisadores da Pontifícia Universidade Católica-PUC/RS, que estabeleceram um ranking com a freqüência com que os fonemas ocorrem na língua portuguesa, no intuito de desenvolver teclados para computador e máquinas de escrever adaptados à língua falada e escrita no Brasil, e não à língua inglesa como ocorre nos conhecidos teclados qwerty atuais. O ranking elaborado pela PUC/RS é apresentado a seguir [42]:

Ordem	Fonema	Frequência (%)
1.	/a/	13,66
2.	/e/	12,45
3.	/o/	11,04
4.	/s/	7,79
5.	/i/	6,71
6.	/r/	6,59
7.	/n/	5,38
8.	/d/	5,14
9.	/t/	4,61
10.	/u/	4,45

TABELA 4.2 - Freqüência de Fonemas para a Língua Portuguesa.

Outro trabalho importante no estabelecimento das listas hierárquicas de freqüência de fonemas, foi elaborado por Osvaldo Sangiorgi em seu estudo **Aspectos Quantitativos e Formais do Sistema Fonológico da Língua Portuguesa Contemporânea no Brasil**, uma tese de doutoramento realizada na Universidade de São Paulo - USP citada em [38]. Os resultados desse estudo estão apresentados a seguir, observando que, por motivo de simplificação, são apresentados apenas os primeiros quinze fonemas mais utilizados:

Ordem	Fonema	Ordem	Fonema	Ordem	Fonema
1.	/a/	6.	/d/	11.	/w/
2.	/i/	7.	/t/	12.	/k/
3.	/u/	8.	/e/	13.	/o/
4.	/s/	9.	/y/	14.	/p/
5.	/r/	10.	/l/	15.	/m/

TABELA 4.3 - Frequencia de Fonemas para a Lingua Portuguesa.

A partir desses estudos, a frase **AJUSTE DE TEMPO** foi construída, verificando-se que os fonemas utilizados neste caso, ocupam os quinze primeiros lugares em têrmos da freqüência com que ocorrem na língua portuguesa.

4.1.2 Fontes de Voz para Locução e Métodos de Gravação

Os sinais de voz utilizados para o desenvolvimento do trabalho, foram obtidos a partir de duas fontes:

- gravação, através de fita cassete;
- diretamente de um microfone.

No primeiro caso foram realizadas onze locuções com cinco homens e seis mulheres pronunciando, individualmente, a frase **AJUSTE DE TEMPO**. Da mesma forma, foram realizadas oito locuções, sendo quatro de vozes masculinas e quatro de vozes femininas, utilizando-se diretamente um microfone. A frase pronunciada pelos locutores, neste caso, também foi a mesma.

Utilizou-se o microfone e a fita cassete para obter os sinais de locução, para simular as fontes de voz utilizadas na prática.

O microfone utilizado nas gravações é da marca Dynamic, com impedância de 600 ohms e características unidirecionais, auxiliando, dessa forma, na eliminação de sons indesejáveis. O locutor foi colocado à frente de um Rádio Gravador Sanyo MCD 40, em uma sala fechada na qual só tiveram acesso o participante e o operador do aparelho, com

a finalidade de que fossem evitados quaisquer tipos de ruído ambiental. A fita cassete utilizada é da marca Sony, modelo HF60, de baixo nível de ruído de polarização.

No caso das locuções com microfone, foi utilizada a própria sala do LAPS. O microfone foi ligado diretamente ao aparelho de som Phillips FC210 / Stereo Tape Deck, que, por sua vez, estava interligado ao processador de sinais TMS 320C25, responsável pela digitalização dos sinais de voz e instalado em um microcomputador PC.

Apesar das salas de gravação não disporem de isolamento acústico adequado, a ausência de ruídos causados por conversas de fundo, motores de automóveis e equipamentos de ar condicionado foi adequada.

4.1.3 Processo de Digitalização do Sinal de Voz para Locução e Metodologia do Trabalho

Os sinais de voz analógicos foram gravados em fita cassete ou captados diretamente por um microfone, filtrados e digitalizados. Os sinais foram filtrados por um filtro Butterworth passa-baixas de 4ª ordem, desenvolvido no LAPS, com freqüência de corte em 5 kHz. Este valor foi escolhido por abranger as componentes de freqüência mais significativas da voz e por permitir que um número

inteiro de amostras seja processado dentro de um segmento [43].

Os sinais foram digitalizados utilizando-se uma placa de aquisição de dados baseada no processador de sinais TMS 320C25 da Texas Instruments [44,45], instalado em um microcomputador PC. A digitalização foi realizada utilizando-se uma freqüência de amostragem de 10 kHz, com quantização, definida no processador, em 16 bits (2^{16}=65536 níveis). A seguir, as amostras dos sinais de voz foram transferidas para o disco rígido do microcomputador, pelo próprio processador TMS. Os sinais foram gravados em arquivos, sendo que cada arquivo contêm a voz de um locutor mencionando a frase de referência.

As estatísticas dos sinais de voz em uma locução, foram obtidas a partir dos arquivos de voz gerados, os quais foram segmentados em blocos de 4 ms e processados pelo algoritmos desenvolvidos neste trabalho.

Os algoritmos foram codificados através do ambiente de programação Turbo C, versão 2.0 [46,47,48,49,50], para microcomputadores do tipo IBM PC.

4.2. Configuração do Detetor de Voz: Modelo Utilizado

A performance de um detetor de voz está ligada à sua eficiência em reconhecer a voz na presença de ruído, detetando os fonemas falados e identificando, com a maior precisão possível, o início e o fim das palavras.

Neste trabalho o detetor desenvolvido deve ser capaz de determinar se o segmento em análise representa um som **surdo**, um som **sonoro** ou **silêncio** (ausência de voz). Existem vários processos para se conseguir obter essa classificação, considerando a fonte de voz como uma frase ou um som contínuo [33,51,52,53]. Neste caso, o detetor utiliza parâmetros temporais para determinar se o segmento em análise é um som **surdo** ou **sonoro**, ou mesmo **silêncio**.

À despeito das definições apresentadas na seção 2.5, referentes aos intervalos de voz, neste Capítulo específico, as pausas intersilábicas e as pausas entre palavras serão consideradas como intervalos de silêncio, em virtude desta denominação ser utilizada com freqüência na literatura encontrada sobre este tema.

4.2.1 Parâmetros Temporais Utilizados no Detetor

Os parâmetros temporais **energia** (En), **taxa de cruzamento por zero** (ZCR), **variação de energia** (En^m/En^{m-1}), **número total de picos** (NPICO), **diferença de picos** (DPIC) e **coeficiente de autocorrelação normalizado** (COR), são utilizados como medidas de sinal para a construção do algoritmo de decisão do detetor elaborado.

O trabalho desenvolvido por Yohtaro Yatsuzuka [35], que construiu um detetor de voz altamente sensível para utilização em circuitos de conversação internacional, empregado em satélites de comunicação, foi tomado como referência, no que concerne à utilização da **energia** como parâmetro principal, com o objetivo de determinar o processamento a ser utilizado para classificação do sinal de voz. Consideràveis alterações, a partir daí, foram efetuadas, para adequar a estrutura básica do detetor tomado como referência, ao objetivo do trabalho proposto. Essas alterações foram necessárias porque, em [35], Yatsuzuka utilizou testes com parâmetros espectrais como a seqüência de bits do sinal e, neste trabalho, foram empregados apenas parâmetros temporais no algoritmo desenvolvido para a classificação sonoro-surdo-silêncio do quadro em análise.

A figura 4.1 apresenta a configuração do detetor desenvolvido neste trabalho.

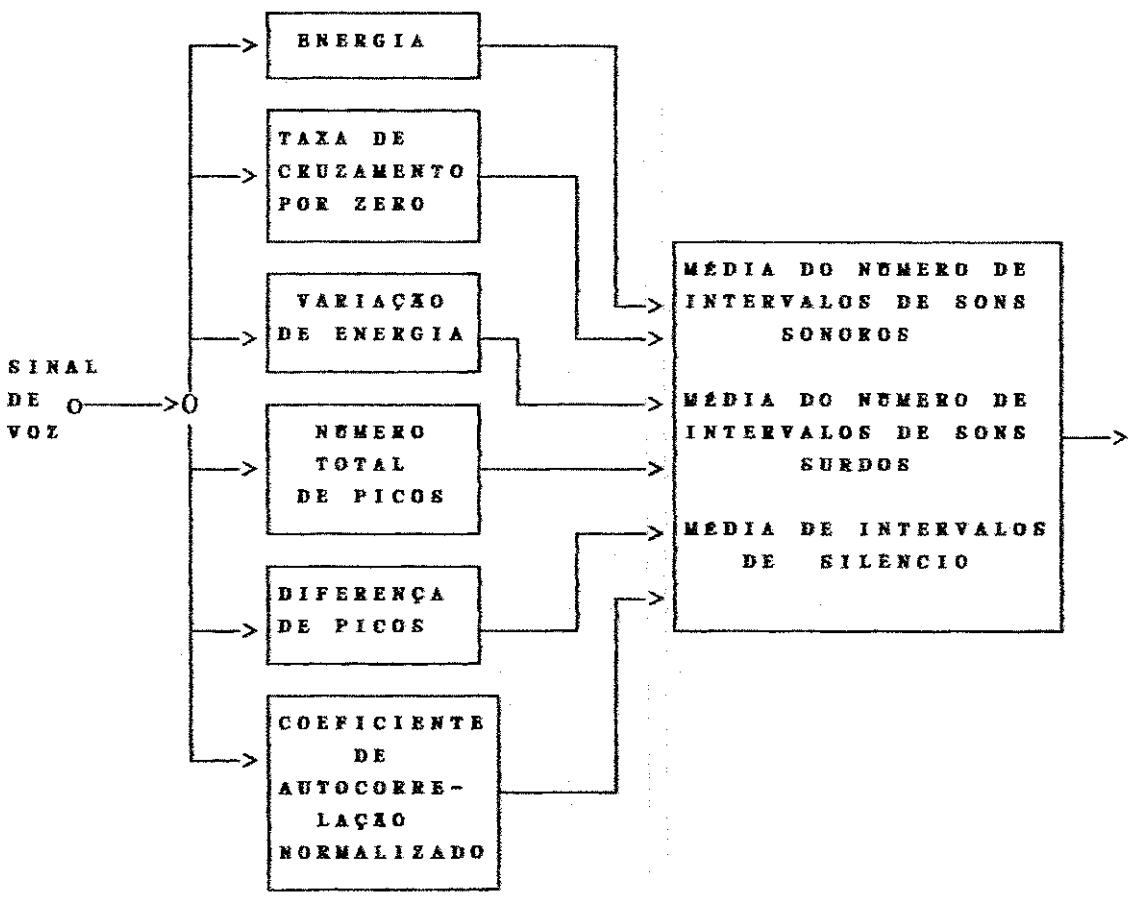


FIGURA 4.1 - Configuração do Detetor de Voz Utilizado.

4.2.2 Procedimento Utilizado para a Deteção do Sinal de Voz

O processo de deteção da voz em uma locução consiste de sua segmentação em blocos de 4 ms, calculando-se, para cada um desses segmentos os valores dos parâmetros correspondentes [54]. A seguir, o valor da energia obtido no intervalo em análise (En), deve ser comparado com três valores de limiar previamente estabelecidos (E1, E2 e E3), que delimitam quatro faixas de energia, conforme pode ser visto a seguir:

FAIXAS	LIMIARES
1	E1 > En
2	E1 <= En < E2
3	E2 <= En < E3
4	E3 <= En

TABELA 4.4 - Faixas definidas pelos Limiares de Energia.

A energia segmentar medida deve selecionar uma das quatro faixas, a partir de sua comparação com os valores de limiar. Cada uma dessas faixas possui processamentos próprios específicos, utilizando testes com os demais parâmetros temporais definidos anteriormente, com a finalidade de determinar o tipo de segmento de voz em análise.

4.2.2.1 Algoritmo Utilizado no Detetor e Medidas Físicas dos Parâmetros Temporais

Cada segmento de voz gerado na saída do detetor de voz, pode ser classificado como **som sonoro**, **som surdo** ou **silêncio**. Esses eventos também são denominados de estados de saída.

Se a **energia** segmentar (**En**) estiver situada abaixo do valor de limiar definido por **E1** (Faixa 1), o segmento na saída do detetor será considerado como **silêncio**. Se o valor de **energia** for superior ao maior valor de limiar definido por **E3** (Faixa 4), o segmento na saída do detetor será considerado como um som **sonoro**.

Caso a **energia** segmentar esteja situada entre os valores máximo e mínimo de limiar, a determinação do tipo de som em análise, dependerá dos diferentes processamentos executados nas faixas 2 e 3 de **energia**. Na faixa 2 é possível encontrar os três estados definidos anteriormente. Na faixa 3 pode-se encontrar apenas os sons **sonoros** e **surdos**, em virtude do elevado valor de limiar de **energia** definido para **E2**, em relação aos valores de ruído encontrados.

A seguir é apresentado o algoritmo elaborado para o detetor de voz, cuja finalidade é determinar os sons emitidos em uma locução:

1. FAIXA 1 [En < E1]

Silêncio

2. FAIXA 2 [E1 <= En < E2]

SE [(ZCR <= 16) E (NPICO <= 12)]

Som Sonoro

SE [(4 <= ZCR <= 20) E (12 < NPICO <= 22)]

SE [($\overline{\text{DPIC}}$ > 3.6) OU (DPIC > ZCR)]

Silêncio

CASO CONTRARIO

Som Sonoro

SE [(20 < ZCR <= 28) E (14 <= NPICO <= 26)]

Silêncio

SE [(28 < ZCR <= 50) E (14 <= NPICO)]

SE [COR > 0.4]

Silêncio

CASO CONTRARIO

Som Surdo

SE [(ZCR > 50)]

Som Surdo

SE [Indefinido]

SE [Os últimos quatro segmentos forem iguais, o segmento atual assumirá esta condição]

CASO CONTRARIO [Segmento Indefinido]

```
3. FAIXA 3 [ E2 <= En < E3 ]

      SE [ En^m > (En^(m-1)+15) ]

           Som Sonoro

      CASO CONTRÁRIO

           SE [ (ZCR <= 20) E (NPICO <= 26) ]

                Som Sonoro

           SE [ (22  < ZCR < 30 ) E

                      (14 < NPICO < 26) ]

                SE [ COR > 0.6 ]

                     Som Sonoro

                CASO CONTRÁRIO

                     Som Surdo

           SE [ (ZCR >= 30)  E (NPICO >= 14) ]

                Som Surdo

           SE [ Indefinido ]

                SE [ Os últimos quatro segmentos
                     forem iguais, o segmento atual
                     assumirá esta condição ]

                CASO CONTRÁRIO [Segmento Indefinido]
```

```
4. FAIXA 4 [ En > E3 ]

      Som Sonoro
```

Os valores de limiares foram obtidos a partir de intensivos testes realizados com os arquivos de locução. Esses testes consistiram da anâlise dos valores encontrados para os parâmetros temporais, ao longo da frase de referência, para todos os locutores, especialmente nos intervalos de transição entre os sons.

Pode-se considerar que, hierárquicamente, a energia é o principal parâmetro na decisão do estado de saída de uma locução. Sua função é a de atuar como uma chave, definindo o estado na saída do detetor, ou designando o processamento adequado para essa decisão. A taxa de cruzamento por zero, o número total de picos e a variação de energia, aparecem em um grau intermediário no processo de decisão. Finalmente, o coeficiente de autocorrelação normalizado e a diferença de picos do sinal de voz, são parâmetros utilizados em escala inferior de importância.

As faixas 1 e 4 do algoritmo elaborado, estabelecem diretamente o tipo de sinal obtido na saída do detetor. O parâmetro energia é suficiente para indicar ausência de sinal de voz (silêncio) quando seu valor é muito baixo (faixa 1), ou a presença de um som sonoro quando seu valor é consideravelmente elevado (faixa 4).

A definição do tipo de som obtido na saída do detetor de voz torna-se difícil quando a energia do sinal encontra-se nas faixas 2 e 3, sendo que a faixa 2 apresenta maior complexidade para indicar esta discriminação, em virtude da possibilidade de existência de

um dos três estados de interesse.

Os limiares de energia que definem a faixa 3 foram escolhidos de modo que, com certeza, sons surdos e sonoros estejam presentes. No caso dos arquivos gravados em fita cassete, por exemplo, a média dos valores de energia em um intervalo reconhecidamente considerado como silêncio (ausência de voz), não ultrapassou a 38,36 dB tomando-se 57 quadros de 4 ms. A energia deste sinal refere-se ao ruído que, evidentemente, não pôde ser totalmente eliminado. No caso dos sinais de voz obtidos a partir de um microfone, o valor médio da energia na região de ausência de voz, atingiu 55,18 dB considerando-se 34 quadros de 4 ms.

A análise do algoritmo estabelecido na faixa 2, pode ser iniciada pela observação de que todo quadro com baixos valores de taxa de cruzamento por zero e número total de picos, abaixo de 16 e 12 respectivamente, corresponde a um som sonoro.

A transição entre os estados é um processo complexo e de difícil definição. Alguns quadros correspondentes aos intervalos de silêncio, possuem valores de ZCR e NPICO muito próximos aos valores dos sons sonoros, nos períodos de transição. Nesses casos foi constatado experimentalmente, que ZCR é maior ou igual a 4 e menor ou igual a 20, e NPICO é maior que 12 e menor ou igual a 22. Para que se possa definir entre um som sonoro e silêncio, nesses casos, foi utilizado o parâmetro diferença de picos. Foi observado que a média dos últimos cinco valores desse

parâmetro, em módulo, é um pouco mais elevada no caso dos intervalos de silêncio, que nos demais estados. Experimentalmente observou-se que essa média está em torno de 3,6. Com isso, quadros com valores acima desse número ou com valores de DPICO maiores que ZCR, são considerados como silêncio e os demais como sons sonoros.

Foi verificado também que quadros com valores de ZCR maiores que 20 e menores que 28, e NPICO maiores ou iguais a 14 e menores ou iguais a 26, configuram estados de silêncio. Este teste é importante para detetar intervalos de silêncio que contenham sinais de ruído com níveis elevados.

A transição entre os estados de silêncio e de sons surdos, pareceu mais simples de ser feita. Quando os valores de ZCR dos quadros analisados encontram-se entre 28 e 50, e os valores de NPICO são maiores que 14, os estados de saída podem ser de silêncio ou de sons surdos. O parâmetro COR possibilita definir com segurança o estado real. Se COR for inferior a 0,4, o estado de saída é um som surdo. Caso contrário, o estado de saída é o silêncio.

Se o valor de ZCR for maior que 50, o estado de saída é um som surdo, independente de outros testes.

Existe ainda a remota possibilidade de que o segmento em análise não se enquadre em nenhum dos casos anteriores. Nesse caso, uma forma de recorrência estabelece que se os últimos quatro quadros na saída do detetor forem iguais, o atual assumirá esse valor. Se isso não for

verdadeiro, o estado atual será considerado como indefinido.

Quando a energia segmentar medida, estabelece que o processamento a ser executado para definição do estado na saída do detetor, é o contido na faixa 3, a análise se torna um pouco mais simples.

Se os valores de ZCR e NPICO forem iguais ou inferiores a 20 e 26, respectivamente, a saída do detetor indicará um som sonoro. Se, no entanto, os valores de ZCR estiverem dentro do intervalo entre 22 e 30, e os valores de NPICO ocuparem o intervalo entre 14 e 26, não haverá uma clara definição à respeito do estado na saída do detetor de voz. Utilizar novamente o parâmetro COR é uma boa maneira de dirimir esta indefinição. Se COR for inferior a 0,6, o estado na saída do detetor será um som surdo, caso contrário será um som sonoro.

Se os valores de ZCR e NPICO forem superiores ou iguais a 30 e 14, respectivamente, o estado de saída pode ser considerado como um som surdo.

A avaliação sobre a possibilidade de um quadro ser considerado como indefinido, mencionada na análise do processamento da faixa 2, também é válida neste caso.

As figuras 4.2, 4.3, 4.4, 4.5, 4.6 e 4.7 apresentam, os sinais de voz de dois dos participantes dos testes pronunciando a frase "AJUSTE DE TEMPO" e os parâmetros temporais obtidos para essas locuções. Todas as figuras são divididas nos itens a e b, concernentes,

respectivamente, aos arquivos gravados em fita cassete e aos arquivos obtidos através do microfone.

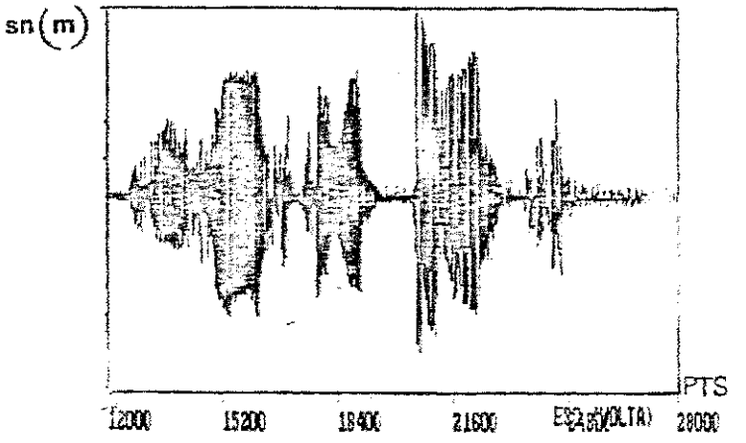


FIGURA 4.2a - Sinal de Voz gravado em fita cassete.

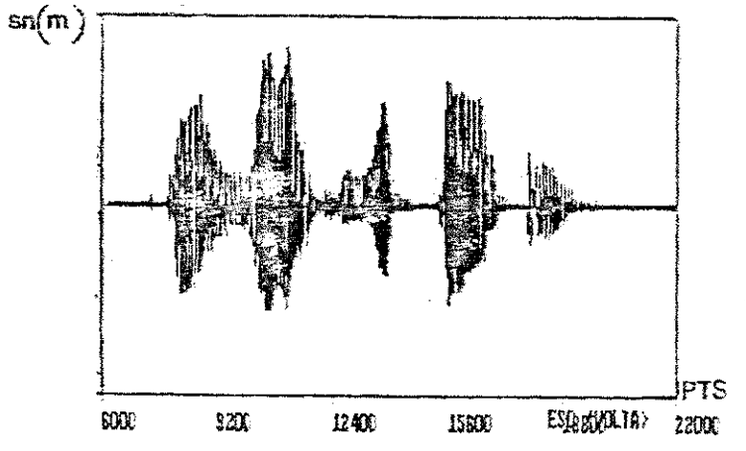


FIGURA 4.2b - Sinal de Voz obtido de um microfone.

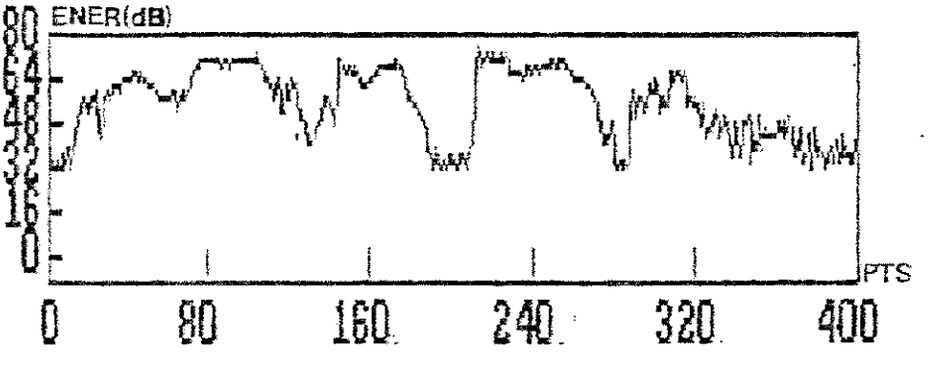


FIGURA 4.3a - Energia da locução / Tape.

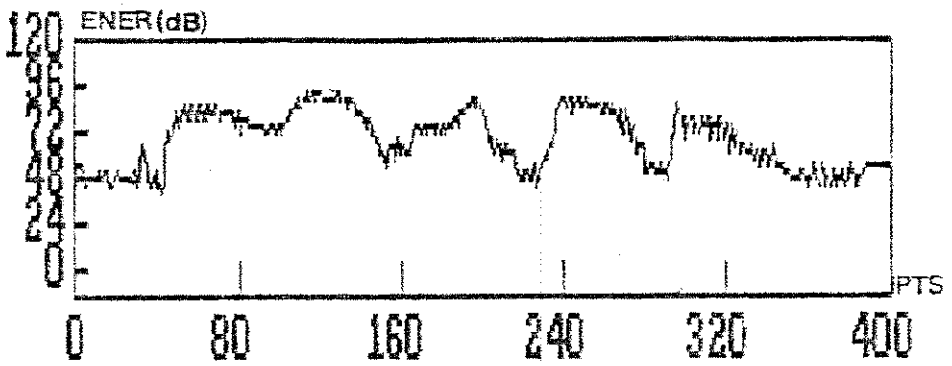


FIGURA 4.3b - Energia da Locução / Microfone.

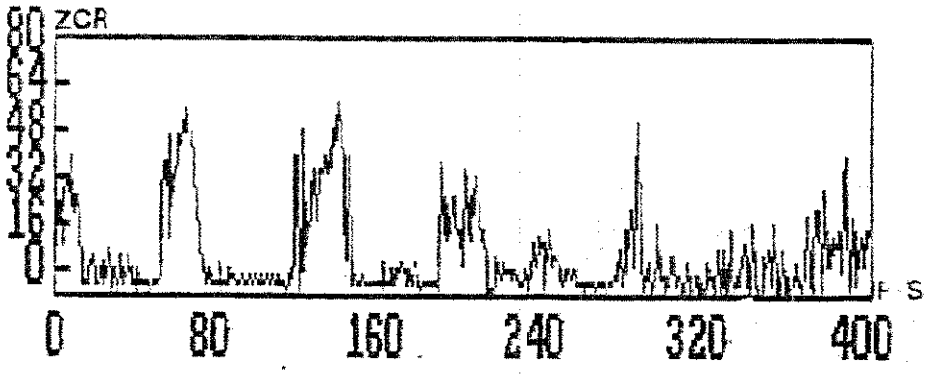


FIGURA 4.4a - Taxa de Cruzamento por Zero da Locução para Tape.

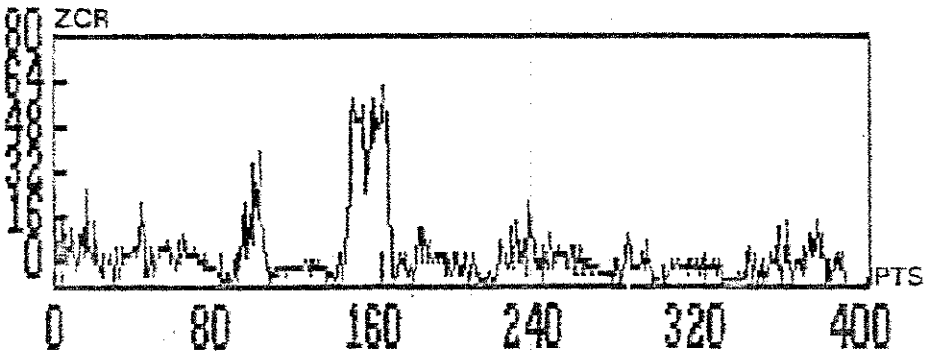


FIGURA 4.4b - Taxa de Cruzamento por Zero da Locução para Microfone.

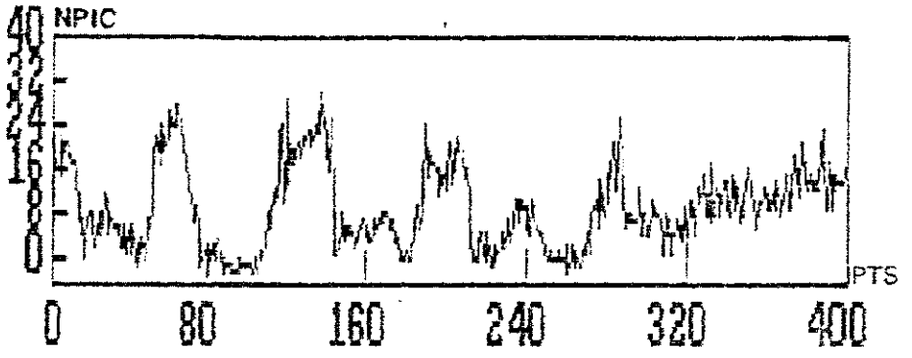


FIGURA 4.5a - Número Total de Picos da Locução para Tape.

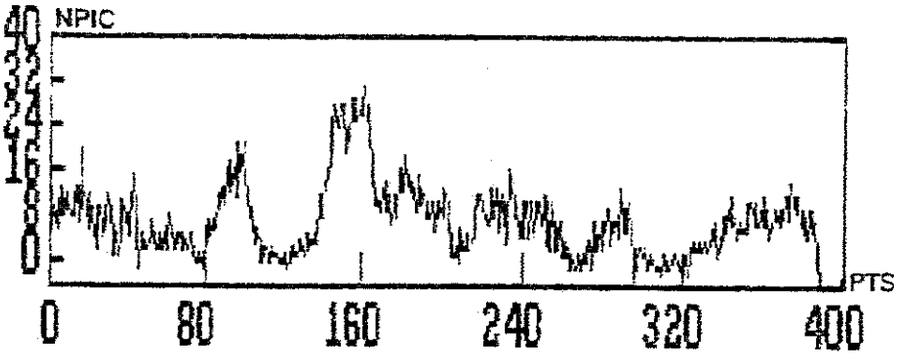


Figura 4.5b - Número Total de Picos da Locução para Microfone.

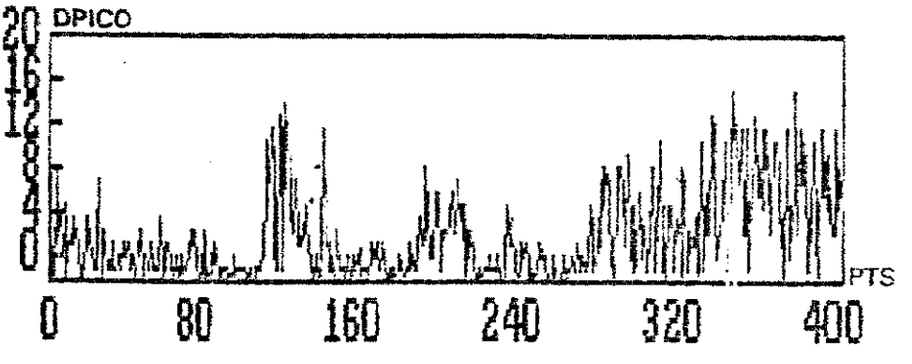


Figura 4.6a - Diferença de Picos da Locução para Tape.

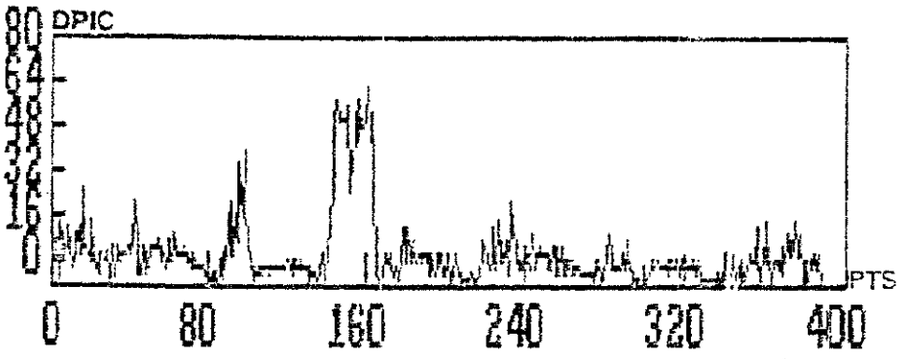


Figura 4.6b - Diferença de Picos da Locução para Microfone.

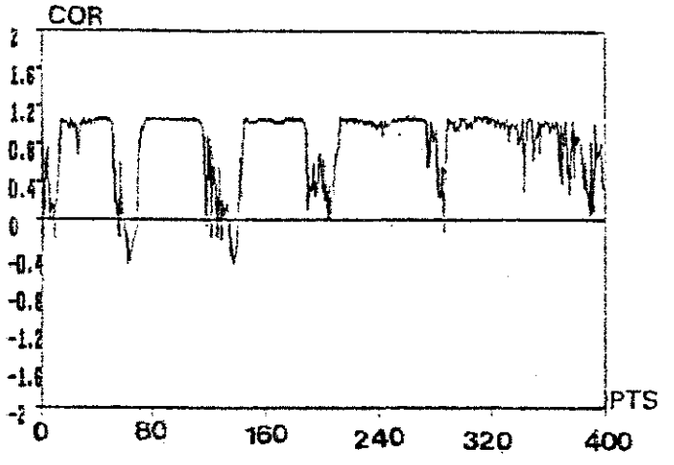


Figura 4.7a - Coeficiente de Autocorrelação Normalizado para Tape.

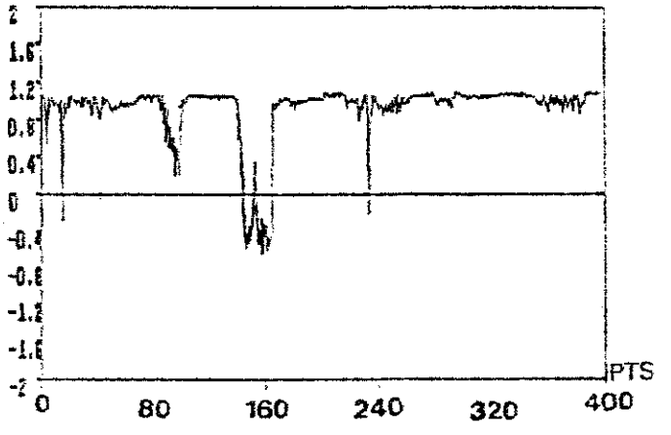


Figura 4.7b - Coeficiente de Autocorrelação Normalizado para Microfone.

Observando as figuras anteriores, pode-se verificar que a amplitude dos sinais de voz gravados em fita cassete e, conseqüentemente, sua energia, é inferior a dos sinais obtidos diretamente do microfone. Em vista disso, os valores dos limiares de energia são diferentes para cada caso.

A tabela 4.5 apresenta os valores de limiares de energia para os arquivos de voz gravados em fita cassete e os obtidos diretamente através de um microfone.

ORIGEM DO ARQUIVO	VALORES DE LIMIAR (dB)		
	E1	E2	E3
MICROFONE	62	68	87
TAPE	42	48	67

TABELA 4.5 - Limiares de Energia utilizados.

É possível notar, a partir da investigação das demais figuras, que os outros parâmetros temporais não sofrem alteração em função da origem do arquivo, daí a utilização dos mesmos valores de limiar em ambos os casos.

4.3 Anàlise Estatística dos Resultados

Os resultados apresentados a seguir, foram obtidos a partir do processamento dos sinais de voz das dezenove locuções da frase "AJUSTE DE TEMPO". O tempo de processamento variou de acordo com a duração que cada participante levou para emitir a frase. Em média, esse tempo foi de 1,18 segundos.

Os resultados são apresentados em dois grupos. O primeiro para as locuções obtidas a partir de uma fita cassete e o segundo para as locuções obtidas por um microfone. Para melhor visualização, os resultados foram tabelados, sendo que as medidas dos arquivos de voz geradas pela fita cassete, estão em tabela diferente das medidas obtidas pelos arquivos de voz conseguidas através do microfone.

4.3.1 Estatística dos Eventos para Locução

Os eventos medidos no intervalo de locução são os seguintes:

- Número Total de Surtos de Sons Sonoros: **NSN**;
- Número Total de Surtos de Sons Surdos: **NSD**;
- Número Total de Intervalos de Silêncio: **NSIL**;
- Número de Quadros: **NQ**;
- Número de Quadros com Sons Sonoros: **SN**;
- Número de Quadros com Sons Surdos: **SD**;
- Número de Quadros de Silêncio: **SIL**;
- Número de Quadros Indefinidos: **IND**;
- Tempo Total do Teste: $TL = NQ.4.10^{-3}$ (s);
- Duração dos Sons Sonoros:

$$Tsn = SN.4.10^{-3} \text{ (s)};$$

- Duração dos Sons Surdos:

$$Tsd = SD.4.10^{-3} \text{ (s)};$$

- Duração dos Intervalos de Silêncio:

$$Tsil = SIL.4.10^{-3} \text{ (s)};$$

- Taxa de Surto de Voz Média:

$$TSl = \frac{NSN + NSD}{TL} \text{ (ciclos/s)} ;$$

- Taxa de Ocorrência dos Sons Sonoros:

$$TOsn = \frac{Tsn}{TL} .100\ (\%);$$

- Taxa de Ocorrência dos Sons Surdos:

$$TOsd = \frac{Tsd}{TL} .\ 100\ (\%);$$

- Taxa de Ocorrência dos Intervalos de Silêncio:

$$TOsil = \frac{Tsil}{TL} .\ 100\ (\%);$$

- Duração Média dos Sons Sonoros:

$$Msn = \frac{Tsn}{NSN}\ (s);$$

- Duração Média dos Sons Surdos:

$$Msd = \frac{Tsd}{NSD}\ (s);$$

- Duração Média dos Intervalos de Silêncio:

$$Msil = \frac{Tsil}{NSIL}\ (s).$$

As Tabelas 4.6 e 4.7 mostram o número total de surtos de sons sonoros, de sons surdos e de intervalos de silêncio para cada locutor, além das médias encontradas para cada um desses eventos, considerando-se as duas fontes de locução utilizadas:

QUADROS	EVENTO	LOCUTOR											MÉDIA
		1	2	3	4	5	6	7	8	9	10	11	
	NSN	9	9	16	10	15	10	20	14	11	17	8	13
	NSD	5	6	8	5	6	4	17	7	11	8	12	8
	NSIL	12	2	10	10	9	7	19	7	19	18	17	12

Tabela 4.6 - Valores dos Números de Surtos para os Sons Sonoros, Surdos e Intervalos de Silêncio, para fita cassete.

QUADROS	EVENTO	LOCUTOR								MÉDIA
		1	2	3	4	5	6	7	8	
	NSN	12	19	12	8	19	20	13	9	8
	NSD	7	2	1	4	2	4	2	3	3
	NSIL	7	18	11	7	16	18	13	7	12

Tabela 4.7 - Valores dos Números de Surtos para os Sons Sonoros, Surdos e Intervalos de Silêncio, para microfone.

As Tabelas 4.8 e 4.9 mostram a soma total dos quadros, dos quadros com sons sonoros, dos quadros com sons surdos, dos quadros com silêncio e as médias desses eventos, para os arquivos de voz gerados por microfone e fita cassete.

	EVENTO	LOCUTOR											MÉDIA
		1	2	3	4	5	6	7	8	9	10	11	
QUADROS	TOTAL (NQ)	271	311	309	319	310	298	311	279	277	304	292	298
	SONORO (SN)	183	240	218	242	242	208	183	166	174	175	150	198
	SURDO (SD)	15	40	29	23	25	44	55	46	31	40	34	35
	SILENCIO (SIL)	71	34	62	54	43	46	73	67	72	88	108	65
	INDEFINIDOS (IND)	1	0	0	0	1	0	0	0	0	1	2	0,5

TABELA 4.8 - Valores dos Eventos de locução, considerando-se sinais de voz obtidos através de fita cassete.

	EVENTO	LOCUTOR								MÉDIA
		1	2	3	4	5	6	7	8	
QUADROS	TOTAL (NQ)	319	300	225	230	368	313	288	294	292
	SONOROS (SN)	241	177	163	164	245	233	187	208	202
	SURDOS (SD)	28	25	7	23	18	26	19	34	22
	SILENCIO (SIL)	50	98	55	43	105	50	82	52	67
	INDEFINIDOS (IND)	0	0	0	0	0	4	0	0	0,5

TABELA 4.9 - Valores dos Eventos de locução, considerando-se sinais de voz obtidos através de um microfone.

A tabela 4.10 apresenta o tempo médio de duração de cada evento, dentro da locução da frase utilizada como referência, obtido a partir da multiplicação dos resultados apresentados nas tabelas anteriores, por $4x10^{-3}$ s, tempo de duração de cada quadro.

FONTE DA LOCUÇÃO	EVENTO			
	TL (s)	Tsn (s)	Tsd (s)	Tsil (s)
TAPE	1,192	0,792	0,140	0,260
MICROFONE	1,168	0,808	0,088	0,268

TABELA 4.10 - Tempo médio de duração de cada evento.

A Tabela 4.11 apresenta a taxa de surto de voz média, as taxas de ocorrência e as durações médias dos sons sonoro, surdo e do intervalo de silêncio.

EVENTO	FONTE DE LOCUÇÃO	
	TAPE	MICROFONE
TSl (ciclos/s)	17,61	9,40
TOsn (%)	66,44	69,18
TOsd (%)	11,74	7,54
TOsil (%)	21,81	22,95
Msn (ms)	60,92	101,00
Msd (ms)	17,50	29,33
Msil (ms)	21,66	22,33

Tabela 4.11 - Valores obtidos para os Eventos de Locução.

4.3.2 Análise dos Resultados e Conclusões

A seguir é feita uma análise dos resultados apresentados nas tabelas 4.6, 4.7, 4.8, 4.9, 4.10 e 4.11.

Os resultados mostram que o algoritmo detetou a presença de uma maior quantidade de fonemas sonoros, que constituem a frase de referência. De fato, em "AJUSTE DE TEMPO" existe a presença de vàrios sons sonoros como o /a/ e o /u/, além de oclusiva sonora como o /d/ e oclusivas como o /t/ e o /p/, em que a pressão do ar exercida sobre a boca fechada, se transforma em um rompimento brusco, quando de sua abertura, causando o aumento súbito da energia do sinal. Isto leva à obtenção de estados de silêncio antes da oclusão e de sons sonoros após o efeito dessa "explosão". Além desses fonemas, pode ser encontrada, ainda, a nasal sonora /m/ na frase.

Os sons surdos correspondem à fricativa /s/. A fricativa /j/ tem características sonoras, mas sua classificação como um estado surdo ou sonoro, depende, basicamente, da pronúncia do locutor. Em alguns arquivos, a intensidade com que o fonema /j/ foi falado, praticamente tornou os parâmetros temporais classificadores, com características sonoras, ou seja, baixos valores de taxa de cruzamento por zero e número total de picos e elevado valor de coeficiente de autocorrelação normalizado.

Um dado interessante a observar refere-se ao fonema /p/ associado à vogal /o/ no final da frase. Alguns locutores pronunciaram a sílaba de modo decidido. Outros à pronunciaram na forma de um sopro, o que estendeu por alguns milisegundos a duração da frase. Neste segundo caso, o que se percebeu foi a alternânacia entre estados de

silêncio e de sons sonoros, esses em maior quantidade, na saída do detetor.

Foi possível notar que a transição entre os sons surdos e o estado de silêncio ou os sons sonoros, pôde ser constatada com mais precisão do que a transição entre os sons sonoros e o estado de silêncio, através da atuação do parâmetro coeficiente de autocorrelação normalizado. As figuras 4.7 a e b ilustram, com precisão, os intervalos em que COR atinge valores muito baixos, caracterizando os sons surdos, devido à descorrelação desse sinal.

A constatação da exatidão dos resultados pôde ser verificada, através do exame das listagens contendo os valores dos parâmetros temporais utilizados na classificação, bem como o estado de saída correspondente em cada segmento. Dada a grande quantidade de pontos envolvida no processamento, em função do reduzido tamanho utilizado para o quadro de voz e do número de arquivos de voz processado, o exame para verificação da confiabilidade do algoritmo classificador foi realizado apenas em alguns trechos mais críticos dos arquivos, tais como as transições de estados. Os sons sonoros e os intervalos de silêncio óbvios, foram conferidos apenas graficamente. Ao todo foram examinados 844 pontos de arquivos diferentes, através de listagens como a apresentada na Tabela 4.12.

ENER	ZCR	NPICO	DPIC	COR	SAÍDA	QUADRO
91.87	6.000000	5.000000	1.000000	0.966521	2.000000	0
91.18	6.000000	4.000000	2.000000	0.967032	2.000000	1
89.53	8.000000	3.000000	1.000000	0.952177	2.000000	2
91.14	6.000000	5.000000	1.000000	0.970456	2.000000	3
91.48	8.000000	6.000000	2.000000	0.960191	2.000000	4
92.20	4.000000	4.000000	0.000000	0.969613	2.000000	5
91.51	8.000000	5.000000	3.000000	0.960475	2.000000	6
92.33	6.000000	5.000000	1.000000	0.972400	2.000000	7
90.59	8.000000	6.000000	2.000000	0.950169	2.000000	8
92.74	6.000000	7.000000	1.000000	0.964766	2.000000	9
93.01	6.000000	6.000000	2.000000	0.971985	2.000000	10
91.83	6.000000	5.000000	1.000000	0.971692	2.000000	11
88.90	6.000000	6.000000	4.000000	0.967640	2.000000	12
90.61	2.000000	8.000000	4.000000	0.972539	2.000000	13
89.72	4.000000	6.000000	2.000000	0.968827	2.000000	14
89.67	4.000000	7.000000	3.000000	0.978837	2.000000	15
85.82	2.000000	8.000000	8.000000	0.982105	2.000000	16
87.47	2.000000	7.000000	3.000000	0.976840	2.000000	17
85.40	4.000000	7.000000	3.000000	0.968819	2.000000	18
85.85	10.000000	14.000000	2.000000	0.975623	2.000000	19
82.03	4.000000	11.000000	1.000000	0.930668	2.000000	20
86.14	2.000000	13.000000	5.000000	0.959674	2.000000	21
77.41	12.000000	16.000000	10.000000	0.538863	2.000000	22
82.30	18.000000	17.000000	7.000000	0.790315	2.000000	23
80.19	34.000000	21.000000	5.000000	-0.004211	2.000000	24
76.18	54.000000	26.000000	0.000000	-0.420923	2.000000	25
75.21	58.000000	28.000000	0.000000	-0.402042	2.000000	26
70.84	52.000000	26.000000	0.000000	-0.538202	2.000000	27
68.92	52.000000	26.000000	0.000000	-0.325299	2.000000	28
66.13	50.000000	25.000000	1.000000	-0.468117	1.000000	29

TABELA 4.12 - Valores dos Parametros Temporais e Estado de Saida do Detetor de Voz.

Os valores da Tabela 4.12 referem-se ao intervalo de voz apresentado na figura 4.8 a seguir.

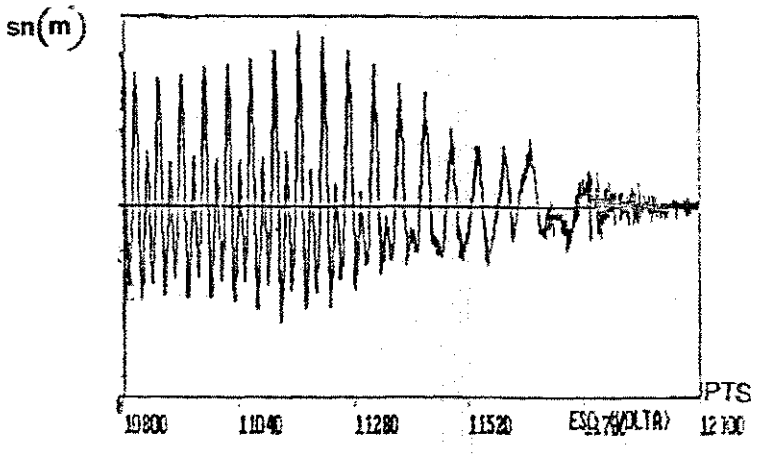
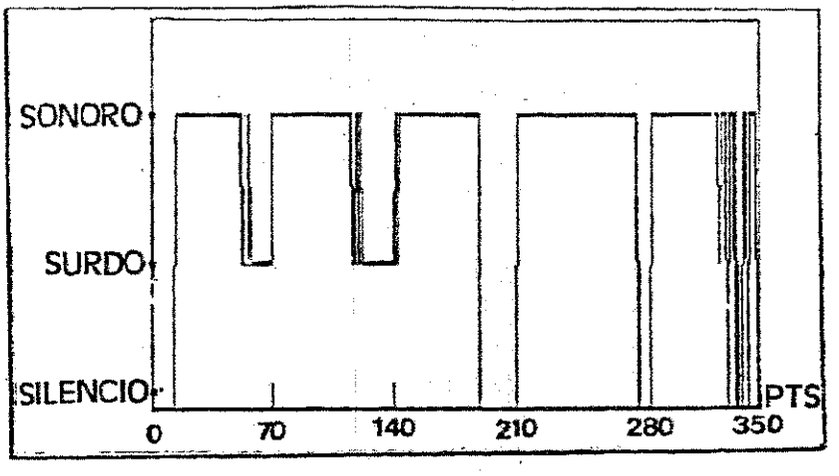


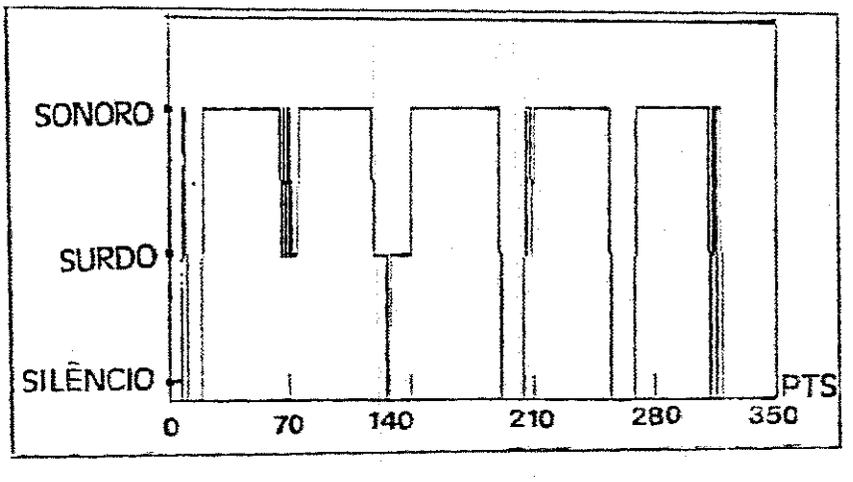
FIGURA 4.8 - Intervalo de voz utilizado para obter os parametros temporais classificadores do estado de saida do Detetor de Voz.

Pôde ser constatado nessa análise, que os êrros evidentes não ultrapassaram a 7% do número total de quadros analisados, o que garante uma boa confiabilidade do algoritmo utilizado no detetor de voz, para fins de discriminação entre os sons surdos, sonoros e o silêncio.

A figura 4.9, a seguir, apresenta os estados de saída de algumas locuções, obtidos a partir do algoritmo desenvolvido.



a) Tape.



b) Microfone

FIGURA 4.9 - Estados de saída do detetor de voz.

É importante citar que os intervalos de silêncio considerados, situam-se entre as palavras ou bastante próximas do início ou do fim da locução.

As figuras de 4.10 a 4.15 apresentam as FDP's dos eventos de voz obtidos através do processamento descrito.

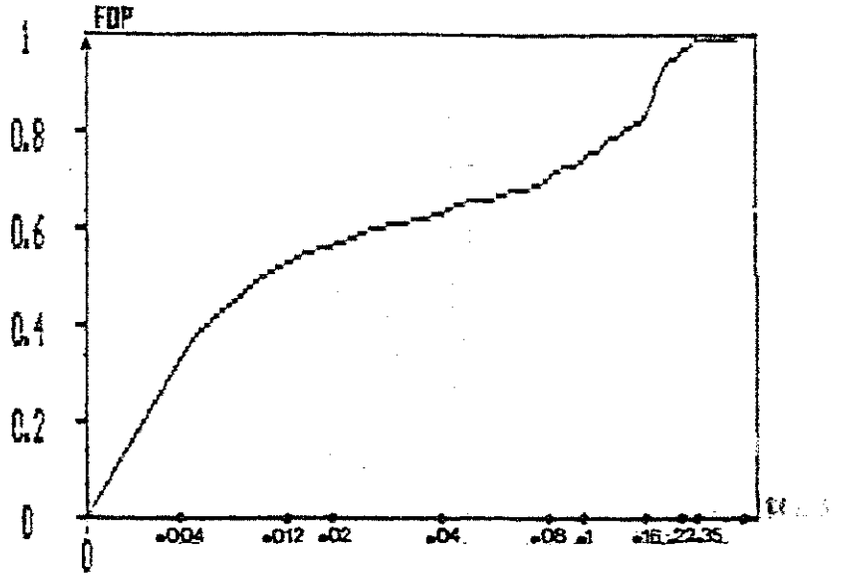


FIGURA 4.10 - FDP do evento Som Sonoro, para Ta.

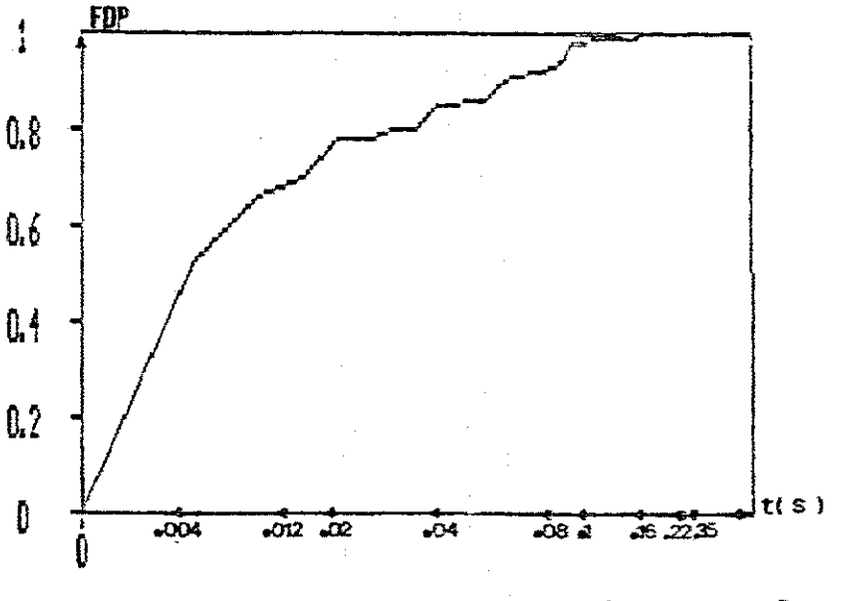


FIGURA 4.11 - FDP do evento Som Surdo, para Tapc.

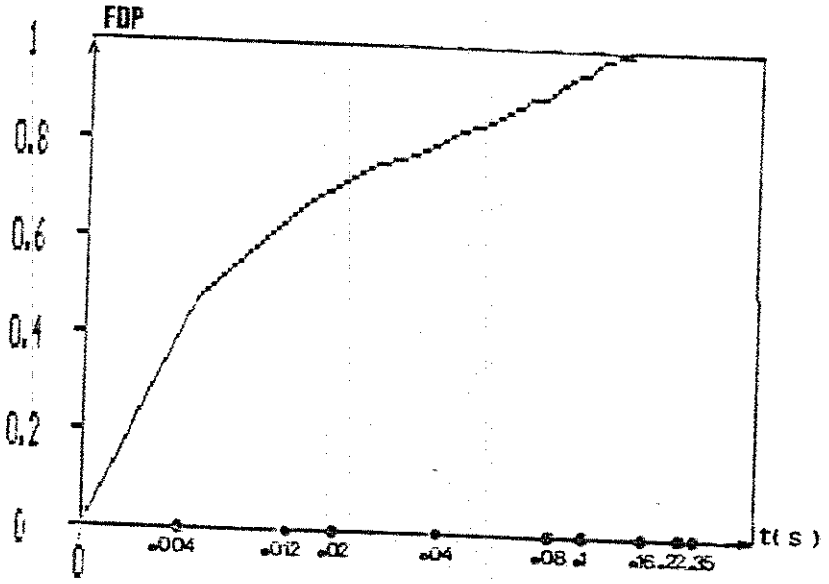


FIGURA 4.12 - FDP do evento Silencio, para Tape.

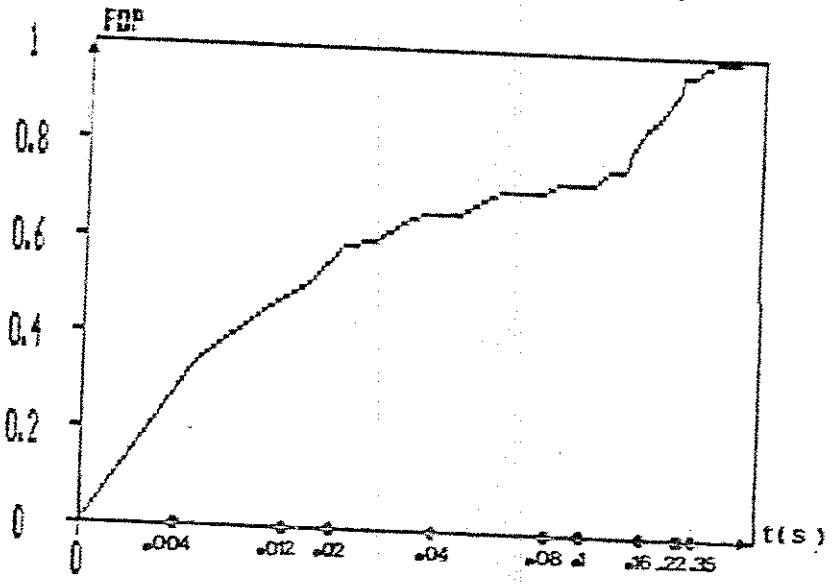


FIGURA 4.13 - FDP do evento Som Sonoro, para Microfone.

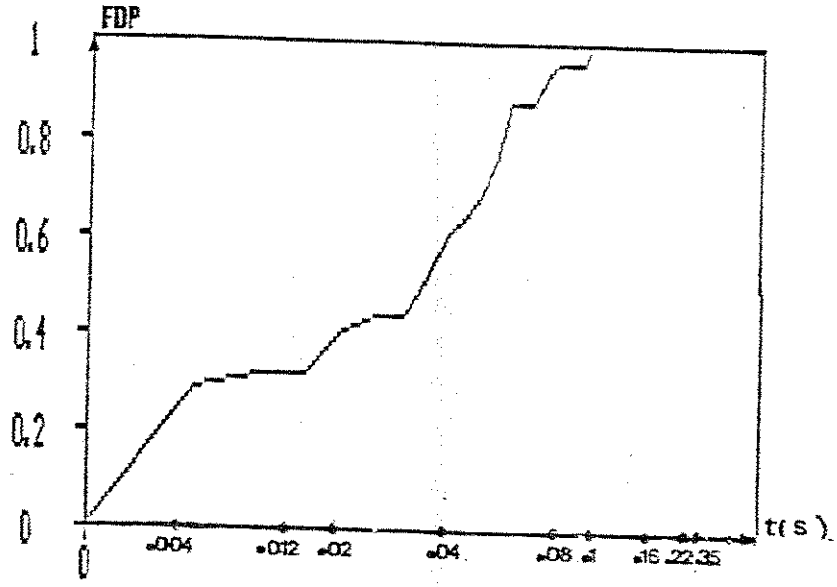


FIGURA 4.14 – FDP do evento Som Surdo, para Microfone.

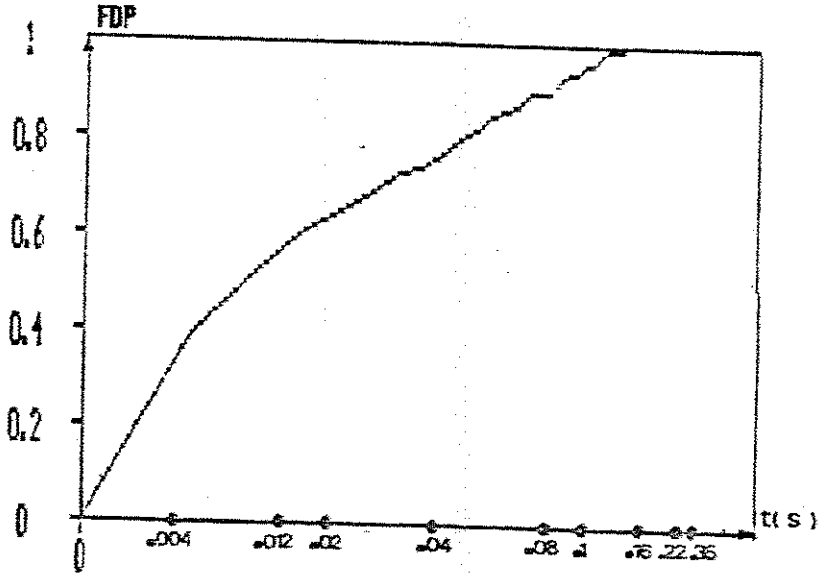


FIGURA 4.15 - FDP do evento Silencio, para microfone.

Analisando-se as FDP's apresentadas nas figuras de 4.10 a 4.15, é possível obter os seguintes resultados:

DURAÇÃO DOS SURTOS DE VOZ (T em ms)	% de t <= T	
	MICROFONE	TAPE
350	100	100
160	85	84
100	72	73
40	66	62
20	51	56
10	30	32

Tabela 4.13 - Valores obtidos a partir das FDP's dos sons sonoros.

DURAÇÃO DOS SURTOS DE VOZ (T em ms)	% de t <= T	
	MICROFONE	TAPE
100	100	100
80	96	92
40	61	85
20	40	77
12	32	68
4	26	45

Tabela 4.14 - Valores obtidos a partir das FDP's dos sons surdos.

DURAÇÃO DOS INTERVALOS DE SILÊNCIO (T em ms)	% de t <= T	
	MICROFONE	TAPE
160	100	100
100	95	96
40	76	80
20	64	70
12	58	63
4	34	40

Tabela 4.15 - Valores obtidos a partir das FDP's dos intervalos de silêncio.

4.4 Considerações Finais

Os resultados obtidos para os eventos de voz de locução, podem ser considerados satisfatórios, considerando-se que os êrros na classificação dos sons da locução da frase de referência, foram de 7% nos pontos mais críticos da transição entre os sons.

O aperfeiçoamento do algoritmo de classificação proposto, pode ser realizado através da inclusão de parâmetros espectrais, que produzem informações consideráveis a respeito dos sons da voz humana. Para isso é necessário aumentar o tamanho da janela para valores maiores ou iguais a 20 ms.

Um dos maiores problemas na deteção da voz, reside na determinação exata do início da locução. O algoritmo proposto neste trabalho procura detetar com eficiência esse processo, através da comparação dos parâmetros temporais obtidos para cada locução, com os valores de referência, medidos anteriormente e que classificam cada evento.

O maior problema encontrado para a elaboração do algoritmo, foi a definição entre quadros correspondentes a sons sonoros e o silêncio, no momento da transição. Nesses instantes os valores de ZCR e NPICO dos dois eventos, praticamente se confundem e é preciso lançar mão de outro parâmetro. O utilizado, diferença de picos, atuou razoavelmente, mas, em algumas ocasiões, gerou indefinições. A análise detalhada da forma de onda da energia do sinal de voz, permitiu concluir que a faixa de energia em que o nível de silêncio encontra-se, é reduzida, diminuindo as dificuldades em decidir se um quadro representa um som sonoro ou silêncio.

A transição entre o estado de silêncio ou os sons sonoros e os sons surdos, como foi visto anteriormente, é feita satisfatóriamente.

A maior parte dos êrros encontrados, foram originados na transição entre os quadros de silêncio e de som sonoro. Os quadros situados nas demais transições, foram classificados praticamente sem êrros.

A escassa quantidade de trabalhos publicados sobre este tema, desenvolvidos específicamente para a língua

portuguesa, impede uma comparação adequada do algoritmo aqui elaborado, para classificar os sons emitidos em uma locução.

Em [33], o êrro encontrado no trabalho desenvolvido por Atal e Rabiner, para classificação sonoro-surdo-silêncio, em aplicações de reconhecimento de voz, varia em torno de 3,5%. Considerando-se que o êrro médio de 7% encontrado neste trabalho, refere-se aos intervalos críticos do processo de classificação (transição silêncio-sonoro), se forem considerados todos os pontos obtidos na locução da frase, este êrro, na verdade, é bem menor, da ordem de grandeza do encontrado em [33].

De modo geral, é possível considerar que a utilização de quadros com duração mais curta, é capaz de perceber melhor as variações no sinal de voz [1]. O inconveniente na utilização de um quadro muito reduzido, como o utilizado neste trabalho, é a impossibilidade do uso parâmetros espectrais importantes para a classificação dos sons da voz, tais como a estimativa da freqüência fundamental *fo* e dos formantes.

Com os resultados alcançados, demonstra-se a possibilidade de se desenvolver um algoritmo eficiente para a classificação dos sons de uma locução, a partir de parâmetros temporais, utilizando nesse processo, segmentos reduzidos para aumentar a precisão das medidas dos parâmetros e, conseqüentemente, da determinação sonoro-surdo-silêncio.

CAPITULO 5

ANALISE DA CONVERSAÇÃO

CAPITULO 5

ANALISE DA CONVERSAÇÃO

5.1 Características da Conversação

A configuração de tempo de uma conversação telefônica pode ser descrita, considerando-se os períodos durante os quais a voz está fluindo dos lábios do locutor, as pausas que intercalam essa voz e os períodos após o término do fluxo de sua voz, nos quais o interlocutor prepara-se para responder. Esse processo pode ser representado pela presença ou ausência da energia da voz em uma conversação telefônica [19,21].

As pausas existentes, indicando ausência do sinal de voz, ocorrem dentro de sentenças, frases e palavras. Algumas dessas pausas não tem duração suficiente para interromper a continuidade do fluxo da voz e outras são tão curtas que nem são notadas pelo ouvinte [19].

Em uma conversação telefônica, enquanto o locutor fala, seu interlocutor, denominado de ouvinte, escuta. Isso resulta em uma ociosidade elevada do canal de

transmissão. Primeiramente porque enquanto o locutor fala, não há sinal de voz no canal de transmissão daquele que escuta; em segundo lugar porque, mesmo considerando o período de tempo em que há sinal de voz, são encontrados inúmeros intervalos de silêncio, com durações perceptíveis, originados quando o locutor interrompe sua fala para pensar por breves instantes, e de pausas entre sentenças, frases, palavras ou sílabas [28].

Em uma conversação telefônica normal, o canal de transmissão de um dos interlocutores está ativo em apenas 40% a 50% do tempo total do evento [55,56]. O tempo restante, em que o canal está inativo, consiste de silêncio, no qual o interlocutor considerado está ouvindo, e pausas.

Este capítulo abordará o levantamento estatístico do sinal ON-OFF, obtido a partir da deteção do sinal de voz produzido pela simulação de conversações telefônicas. O sinal ON-OFF gerado por um detetor de voz projetado especificamente para esta finalidade, é constituído de intervalos que representam surtos de voz e pausas. A partir do sinal ON-OFF resultante, é possível obter as funções distribuição de probabilidade do surto de voz e do silêncio, além de outras medidas de interesse [20].

5.1.1 Fontes de Voz para Conversação e e Métodos de Gravação

Os sinais de voz utilizados para o desenvolvimento do trabalho, foram obtidos através da gravação, em fita cassete, de conversações telefônicas simuladas. Foram realizadas 20 gravações, sendo que em 10 delas foram gravadas, simultâneamente, as vozes dos dois interlocutores e nas outras 10, apenas de um deles. A finalidade da utilização desses dois tipos de gravação, surge da necessidade de se comparar as estatísticas do sinal de voz em uma conversação, sob o ponto de vista unidirecional e bidirecional.

O número de gravações foi decidido a partir da observação dos artigos de Brady e Yatsuzuka, utilizados como referência para a realização deste trabalho. Brady, em seu primeiro trabalho [20], utilizou 8 conversações telefônicas e 16 em seu segundo trabalho [27]. Yatsuzuka utilizou 31 conversações telefônicas [35]. Foi considerado que 20 conversações, divididas na forma apresentada anteriormente, possibilitam, de forma adequada, a obtenção das estatísticas do sinal de voz, além de estabelecer um teste confiável da eficiência do detetor de voz elaborado.

Das 20 gravações efetuadas, 10 delas foram feitas com vozes femininas e as outras 10 com vozes masculinas. Assim, foram obtidas 5 gravações de vozes masculinas e 5

de vozes femininas apenas com um dos interlocutores e 5 gravações de conversações com pares masculinos e 5 com pares femininos.

Cada gravação durou cerca de 30 segundos e foi solicitado aos participantes que procurassem evitar longas pausas entre as palavras, comentários breves (como "sim", "hum", "claro",etc...), respiração mais intensa e hesitações naturais existentes em uma conversação [20].

Novamente é importante reportar-se aos estudos de Brady para justificar a duração do intervalo de tempo de gravação. A duração de voz contínua de cada conversação foi de 55 segundos no primeiro trabalho de Brady [20]. Neste trabalho optou-se pela limitação do tempo de gravação, para diminuir as medidas necessárias à obtenção do objetivo proposto e, também, pelo fator tempo disponível para a realização deste trabalho. Apesar disso, o tempo de 30 segundos utilizado para cada uma das conversações, pareceu adequado aos propósitos do trabalho.

O microfone utilizado nas gravações é da marca Dynamic, com impedância de 600 ohms e características unidirecionais, auxiliando, assim, a eliminação de sons indesejáveis.

No caso da gravação das vozes dos dois interlocutores, estes foram colocados à frente de um microfone e orientados a conversar da forma mais natural possível. É evidente que, sabendo que suas conversas estavam sendo gravadas, no início do processo houve

um certo constrangimento que, evidentemente, impediu o transcurso normal da conversação. Apenas quando foi percebido que o desenrolar do diálogo tornara-se mais real, sem que os interlocutores percebessem, teve início a gravação.

Na gravação de apenas um dos interlocutores, este foi colocado à frente de um microfone e de um aparelho telefônico. A seguir foi estabelecida uma conversação telefônica com uma pessoa conhecida do participante. Assim, apenas a voz do interlocutor que participava do trabalho foi gravada. Nesse caso específico, além das orientações anteriores, foi solicitado que, durante o período de gravação, a pessoa que estivesse sendo gravada falasse o maior tempo possível, permitindo apenas breves espaços de tempo para que seu interlocutor fizesse alguns comentários. Isto ocorre porque, neste trabalho, deseja-se obter as estatísticas dos surtos de voz e das pausas de curta duração, existentes em uma conversação telefônica.

Para evitar quaisquer tipos de ruído ambiental, utilizou-se uma sala fechada na qual só tinham acesso os participantes das gravações e o operador do gravador. Apesar da sala de gravação não dispor de isolamento acústico adequado, a ausência de ruídos provocados por conversas de fundo, motores de automóveis e equipamentos de ar condicionado, foi adequada. Assim, os ruídos indesejáveis foram limitados aos causados pelo microfone, pela fita cassete e pelo aparelho de gravação. O

aparelho utilizado para gravação das conversas foi um Rádio Gravador Sanyo MCD 40 F. As fitas cassete utilizadas foram da marca Sony, modelo HF60, de baixo ruído de polarização.

5.1.2 Processo de Digitalização do Sinal de Voz

As gravações realizadas foram processadas por um detetor de voz, desenvolvido a partir do algoritmo descrito na seção 5.2. Para que esse processamento fosse efetuado, os sinais de voz analógicos gravados em fitas cassete foram filtrados e digitalizados. Os sinais foram filtrados por um filtro Butterworth passa-baixas de 4ª ordem, construído no LAPS, com freqüência de corte em 5 kHz, escolhida por abranger as componentes de freqüência mais significativas da voz e por permitir que um número inteiro de amostras seja processado dentro de um segmento [43].

Os sinais foram digitalizados pelo processador de sinais TMS 320C25 da Texas Instruments [44,45], instalado em um microcomputador PC XT, a uma taxa de amostragem de 10 KHz e quantizados em 16 bits (2^{16}=65536 níveis). A seguir, as amostras dos sinais de voz foram transferidas para um disco rígido do microcomputador pelo próprio processador TMS. A duração do sinal de voz a ser digitalizado por um processador desse porte, é limitada

apenas pela memória disponível no disco rígido [45]. Neste trabalho os sinais foram gravados em arquivos com duração de 3 segundos cada um, para facilitar o processamento do sinal pelo programa desenvolvido para geração do sinal ON-OFF. Assim, os sinais de voz de cada um dos participantes, com duração média de 30 segundos, foram distribuídos em 10 arquivos de 3 segundos.

5.2 Metodologia do Trabalho

Para a obtenção das estatísticas do sinal de voz em uma conversação telefônica simulada, os sinais de voz gravados em uma fita cassete, foram filtrados, digitalizados e armazenados no disco rígido de um microcomputador PC XT. A seguir, os arquivos de voz foram segmentados em blocos de 4 ms e processados, em tempo não real, pelos algoritmos desenvolvidos neste trabalho. Os algoritmos foram codificados através do ambiente de programação Turbo C versão 2.0 [46, 47,48,49,50], compatível com um microcomputador IBM PC XT.

5.2.1 Configuração do Detetor de Voz: Modelo Utilizado

Sabe-se que o detetor de voz tem a tarefa de indicar a presença ou ausência da fala, a partir de um algoritmo de decisão que atua baseado em medidas de sinal. A performance desses sistemas está diretamente ligada à eficiência do detetor em *reconhecer* a voz na presença de ruído. Essa eficiência pode ser verificada através de sua capacidade em identificar certas características ou parâmetros do sinal de voz em uma conversação, sem que interrupções indesejáveis sejam causadas no início, meio e fim das palavras. Outras características importantes de um detetor de voz são a sua alta imunidade ao ruído e uma baixa taxa de surto de voz [35].

Para a obtenção dos dados estatísticos do sinal de voz em uma conversação, é necessário um processamento adequado do sinal obtido pelo detetor. Esse processamento envolve a captação da fala, sua conversão em um sinal ON-OFF caracterizando sua presença ou ausência e, finalmente, a alteração desse sinal para eliminar efeitos indesejáveis como interrupções bruscas no fluxo da fala, provocadas por fonemas oclusivos, que podem fazer com que um detetor bastante sensível indique um intervalo de pausa (estado OFF), quando, na verdade, existe sinal de voz nesse período.

5.2.1.1 Parâmetros Temporais Utilizados no Detetor

Os parâmetros temporais energia, taxa de cruzamento por zero e variação da energia, a curtos intervalos de tempo, são utilizados como medidas de sinal para a construção do algoritmo de decisão do detetor elaborado. A energia é um parâmetro importante porque reflete as variações de amplitude do sinal de voz [1]. O parâmetro taxa de cruzamento por zero é utilizado por ser um indicador da freqüência na qual a energia está concentrada no espectro do sinal [33], caracterizando, com razoável segurança, os fonemas surdos [35]. A variação da energia é utilizada por ser um parâmetro bastante sensível às variações relativas, que a envoltória do sinal apresenta ao longo do tempo [55].

Foi utilizado, como referência, o trabalho desenvolvido por Yohtaro Yatsuzuka [35], que construiu um detetor de voz altamente sensível para utilização em circuitos de conversação internacional, empregado em satélites de comunicação.

No entanto, algumas alterações e simplificações foram efetuadas, em relação ao detetor de Yatsuzuka, porque os objetivos na implementação de seu trabalho eram diferentes dos aqui buscados. As alterações, basicamente, restringem-se à não realização dos testes de seqüência de bits do sinal de voz, que utiliza características

espectrais do sinal.

A figura 5.1 apresenta a configuração do detetor desenvolvido neste trabalho.

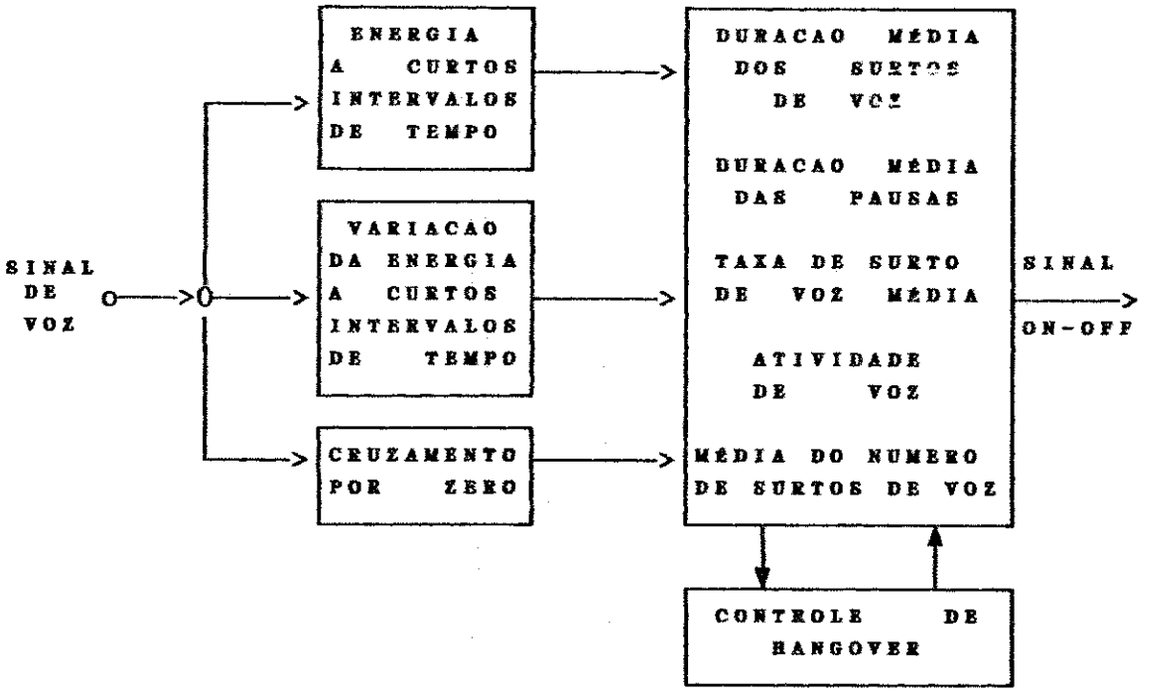


FIGURA 5.1 - Configuração do Detetor de Voz.

Um dos aspectos mais importantes em um detetor de atividade de voz refere-se ao processamento do sinal ON-OFF, obtido a partir das medidas de sinal realizadas pelo algoritmo de decisão. Sabe-se que pequenas hesitações do locutor no momento em que está falando, ou diminutas pausas intersilábicas, na maioria das vezes imperceptíveis para o ouvinte, provocam ausência de energia da voz em um meio de comunicação. Por outro lado, ruídos impulsivos de curta duração, ocasionados pelo manuseio do aparelho

telefônico ou do microfone, distúrbios elétricos ou a utilização de fitas cassete para gravação do sinal de voz, empregadas neste trabalho, produzem surtos de energia capazes de acionar o detetor em um intervalo de tempo indesejado [20,27]. Esses são alguns dos motivos pelos quais há necessidade de *corrigir* o sinal ON-OFF na saída do detetor.

Existem várias maneiras dessa correção ser efetuada. Alguns projetistas fazem essas correções utilizando parâmetros temporais como **tempo de preenchimento** (fill-in-time) e **tempo de eliminação** (throwaway time), a serem definidos a seguir. Os trabalhos de Paul Brady [20,21,27], também tomados como referência, utilizaram esses parâmetros temporais. Outros utilizam o parâmetro temporal **hangover**, do qual o tempo de preenchimento é uma variação. O trabalho de Yatsuzuka [35] utilizou o hangover para correção do sinal ON-OFF obtido a partir do algoritmo de decisão.

O tempo de eliminação é empregado para eliminar os surtos de energia que ocupam intervalos inferiores a um valor pré-definido, transformando os estados ON (nível lógico 1) em estados OFF (nível lógico 0), pois considera-se que, nessas condições, os surtos são produzidos por ruídos impulsivos. Os valores do tempo de eliminação estão em torno de 10 ms. Isto significa que qualquer surto de voz com duração inferior a esse valor, deve ser eliminado, sendo considerado como surto de ruído. Conseqüentemente,

esse segmento assume o estado OFF(nível lógico 0) [27].

O outro parâmetro temporal, denominado de tempo de preenchimento, é utilizado como referência para preenchimento das pausas de diminuta duração, transformando os estados OFF (nível lógico 0) em estados ON (nível lógico 1), ocasionadas durante o período em que o locutor está falando. Essas pausas, obviamente, diferem das existentes durante o período em que um interlocutor está ouvindo o outro, em uma conversação. Portanto, as pausas com duração inferior a um valor pré-definido, devem ser preenchidas, sendo consideradas como voz por representarem breves interrupções no fluxo da fala, quase imperceptíveis para o ouvinte. Em [20,27] foi utilizado um tempo de preenchimento de 200 ms. Assim, qualquer segmento considerado, a princípio, como pausa e que tenha duração inferior a 200 ms, deve ser transformado em surto de voz, assumindo o estado ON.

Neste trabalho, o algoritmo de correção está baseado no hangover, por ser um parâmetro que pode ser utilizado em tempo real. O propósito do hangover é o de cobrir pequenos períodos de silêncio presentes em uma conversação, criando, assim, uma quantidade menor de surtos de voz mais longos [18]. O hangover atua de modo a prolongar o surto de voz por mais algum tempo, após o seu término, evitando transições errôneas de um estado ON (presença de voz) para um estado OFF (ausência de voz) [35]. Em detetores de voz convencionais, o hangover está em torno de 150 a

250 ms. Nesses detetores a duração média do surto de voz está em torno de 1 segundo. A redução do tempo de hangover aumenta consideravelmente a sensibilidade do detetor aos sinais de voz de baixa energia, mantendo, no entanto, alta imunidade ao ruído [18,35]. Neste trabalho foram utilizados dois valores para o tempo de hangover: 20 ms (5 quadros) e 32 ms (8 quadros), considerando-se que o sinal de voz é segmentado em intervalos de 4 ms, a uma taxa de 10kHz.

A importância do hangover reside no aspecto de permitir uma redução no enfraquecimento do sinal de voz, originado pelos atrasos causados pela rede de transmissão. Esses atrasos são mais críticos para o tráfego de voz do que para o tráfego de dados, podendo modificar as características da conversação. Pode-se dividir os atrasos existentes em uma rede, em duas categorias [57]:

- **atrasos fixos**, que acontecem nas redes de circuitos comutados, à longas distâncias;
- **atrasos variáveis**, que ocorrem nas redes de pacotes comutados.

Os atrasos fixos estão relacionados ao problema do eco existente em uma transmissão de voz. Várias técnicas, que não cabem ser discutidas aqui, minimizam este problema. Os efeitos desagradáveis dos **atrasos variáveis** em um sistema de tráfego de voz, devem ser abordados a partir de dois fatores fundamentais [57]:

- a duração dos surtos de voz;
- a preservação da continuidade do surto de voz .

Os surtos de voz não devem ter duração reduzida a ponto de serem ocupados apenas por palavras ou sílabas. A duração dos surtos de voz deve ser suficiente para conter uma frase ou uma sentença. O processamento e a transmissão dos surtos de voz pela rede, devem preservar sua continuidade [57].

O hangover atua na diminuição do enfraquecimento do sinal de voz, de duas formas: primeiramente, pela redução da taxa do surto de voz e, em segundo lugar, forçando que as ocorrências de atraso causadas ao sinal, sejam mais freqüêntes nos períodos de pausa entre frases e sentenças, relativamente mais longos e cujos efeitos são menos perceptíveis que os períodos de pausa entre palavras e sílabas [18].

Os efeitos benéficos do hangover vão mais além. Alguns detetores de voz produzem cortes no início e no fim dos surtos de voz. A utilização do hangover minimiza este efeito e, ao reduzir o número de surtos de voz, garante seu fluxo mais contínuo [18,28,57].

O hangover atua vantajosamente, na deteção de intervalos de atividade de voz degradados por ruído. Os sons surdos, quando atingem níveis de energia da ordem do sinal de ruído, são recuperados através da ação do hangover. Sua ação pode ser conjugada, também, aos sistemas

de supressão de ruído [58].

A desvantagem do hangover está no fato de existir um aumento na atividade da voz, uma vez que parte dos intervalos de pausa são considerados como voz, após sua aplicação [18,57].

Experiências realizadas indicam que a utilização do tempo de preenchimento, ao invés do hangover, produz o efeito de alongar a duração média das pausas e de encurtar um pouco a duração média dos surtos de voz. As taxas de surto de voz produzidas pelo uso do hangover e do tempo de preenchimento são iguais, mas a atividade de voz gerada pela utilização do tempo de preenchimento é um pouco mais curta [27].

É importante salientar um aspecto prático na utilização do tempo de preenchimento. Sua implementação não é uma operação em tempo real, uma vez que o preenchimento a ser realizado no surto de voz em processamento, depende da duração do próximo intervalo de silêncio a ser processado [18].

Em conversações telefônicas o atraso causado pelo tempo de preenchimento, cujo valor típico é de 200 ms, é significativo. Este atraso não é importante, no entanto, em aplicações de voz em tempo não real [18]. Esta é a principal razão para a escolha do hangover como parâmetro corretor do sinal de voz originado das conversações telefônicas.

A figura 5.2 apresenta uma comparação entre os

mecanismos de tempo de preenchimento e hangover.

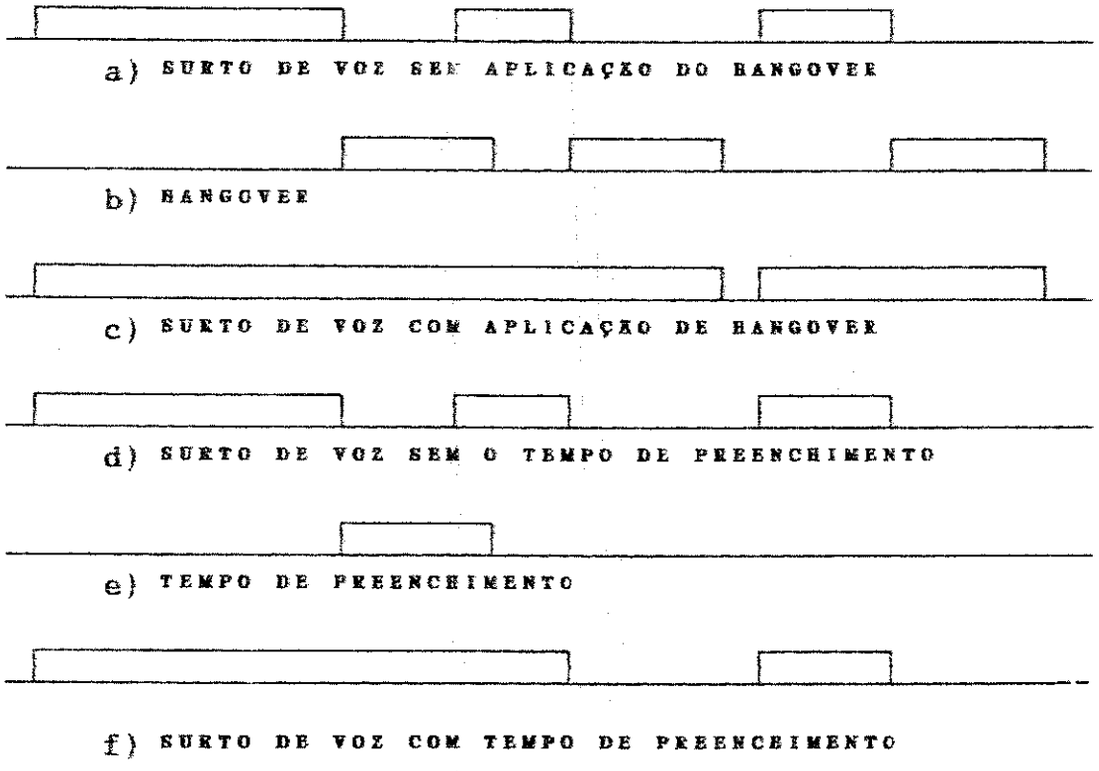


FIGURA 5.2 - Comparação entre os parametros Hangover e Tempo de Preenchimento.

5.2.2 Procedimento Utilizado para a Deteção do Sinal de Voz

O processo de deteção da voz consiste de sua segmentação em blocos de 4 ms, calculando-se, para cada um desses segmentos, a energia (En) e a taxa de cruzamento por zero (ZCR) correspondentes. A seguir, o valor da energia obtido no intervalo em análise, deve ser comparado com três

valores de limiar previamente estabelecidos (E1, E2 e E3), que delimitam quatro faixas de energia, conforme visto a seguir:

- Faixa 1 : E1 > En
- Faixa 2 : E1 <= En < E2
- Faixa 3 : E2 <= En < E3
- Faixa 4 : E3 <= En

TABELA 5.1 - Faixas Definidas pelos Limiares de Energia

A energia segmental medida, deve selecionar uma das quatro faixas, a partir de sua comparação com os valores de limiar. Cada uma dessas faixas possui processamentos específicos, utilizando valores pré-estabelecidos da taxa de cruzamento por zero e da variação de energia. Esses testes visam aumentar a sensibilidade do detetor [35].

5.2.2.1 Algoritmo Utilizado no Detetor

Cada segmento do sinal ON-OFF gerado na saída do detetor de voz, pode ser classificado em um dos três estados indicados a seguir:

- **existência de voz (EV);**
- **silêncio (SI);**
- **hangover (HO).**

Um estado adicional pode ser definido. É um estado intermediário que, nos casos mais críticos de determinação da presença do sinal de voz, utiliza um **parâmetro sinalizador** para estabelecer o estado de saída, a partir da memória de seus quatro últimos valores. Esse estado é denominado de **deteção primária (DP)** [35].

O parâmetro sinalizador utilizado é denominado de PDF^m e tem a finalidade de representar a transição de estados do detetor de voz, no emésimo segmento. Quando a energia do intervalo em análise é inferior ao menor limiar de energia estabelecido **(E1)**, PDF^m sinaliza para a saída com o nível lógico 0, indicando a existência de um segmento de **silêncio (SI)**, sem a necessidade de novos testes. Da mesma forma, quando a energia do sinal de voz no intervalo em análise é superior ao maior limiar de energia **(E3)**, PDF sinaliza para a saída com o nível lógico 1, indicando um segmento com **existência de voz (EV)** [35].

Se a energia segmental estiver situada entre os valores mínimo e máximo de limiar (faixas 2 e 3), tanto pode existir silêncio como sinal de voz no intervalo considerado. Esta decisão será tomada através da utilização dos parâmetros temporais taxa de cruzamento por zero e/ou variação de energia, que determinarão o nível lógico de

PDF^m. Nos casos em que os testes utilizados ainda apresentem incerteza quanto à presença de sinal de voz ou intervalo de silêncio, o estado de deteção primária é utilizado. Nesse caso, avalia-se a combinação dos três estados anteriores e o atual de PDF (PDF^{m-3}, PDF^{m-2}, PDF^{m-1} e PDF^m), sendo possíveis 16 combinações (2^4). Dependendo da combinação obtida, a saída indicará um sinal ON (existência de voz) ou OFF (silêncio) [35]. Isto é feito para que sejam evitadas transições errôneas de intervalos de surto de voz para silêncio e vice-versa. Os resultados dessas combinações são obtidos a partir da utilização de um Mapa de Karnaugh para 4 variáveis. A tabela das combinações de PDF^m é apresentada na seção 5.2.2.2.

A determinação do parâmetro sinalizador PDF e a classificação do estado gerado na saída do detetor de voz, dependem dos parâmetros temporais energia (En), taxa de cruzamento por zero (ZCR) e variação de enrgia.

A seguir é apresentado como são determinados os valores de PDF e do sinal ON-OFF, denominado de VDF, considerando-se as quatro faixas de energia delimitadas pelos valores de limiar:

```
FAIXA 1 ( En < E1 ):                PDF^m = 0; VDF^m = 0;
                                     Silêncio.
```

```
FAIXA 2 ( E1 <= En < E2 ):

  Se [ En^m > NS+En^(m-1)
       E VDF^(m-1) = 0 ]             PDF^m = 1; VDF^m = f(PDF);
                                      Sinal de saída dependerá
                                      da combinação dos quatro
                                      valores anteriores de PDF

  Senão
      Se [ ZCR^m > Z1 ]              PDF^m = 1; VDF^m = 1;
                                      Som fricativo surdo

      Senão
          Se [ ZCR^m > Z2
               E ZCR^m < Z1 ]        PDF^m = 0; VDF^m = 0;
                                      Silêncio

      Senão                         PDF^m = 1; VDF^m = f(PDF);
                                      Sinal de saída dependera
                                      da combinação dos quatro
                                      valores anteriores de PDF
```

```
FAIXA 3 ( E2 <= En < E3 ):          PDF^m = 1; VDF^m = f(PDF);
                                     Sinal de saída dependerá
                                     da combinação dos quatro
                                     valores anteriores de PDF
```

```
FAIXA 4 ( En <= E3 ):               PDF^(m-1) =PDF^m =1; VDF^m=1.
                                     Sinal de voz de alta
                                     energia.
```

OBS.: Valor de NS = 0,477 dB.

5.2.2.2 Diagrama de Estados

A figura 5.3 apresenta o diagrama com as transições de estado do detetor de voz utilizado. Sabe-se que o estado de **silêncio (SI)** indica ausência de voz e que o estado de **existência de voz (EV)** indica sua presença, no segmento em análise. Em certas circunstâncias, é relativamente simples identificar o estado de um segmento. Considerando-se que a energia segmental é o parâmetro de primeira instância no processo de decisão para obtenção do sinal ON-OFF, definidos seus limiares inferior (**E1**, e superior (**E3**), é possível considerar, com bastante segurança, o segmento em análise como **silêncio** se **En < E1**, ou **voz** se **En > E3** e, conseqüentemente, definir seu estado. O maior problema nesta decisão existe quando o valor da energia encontra-se entre os valores de limiar **E1 < En < E3**. Neste caso, o parâmetro sinalizador PDF auxilia na decisão do estado do segmento em análise, através da combinação de seus últimos quatro estados [35].

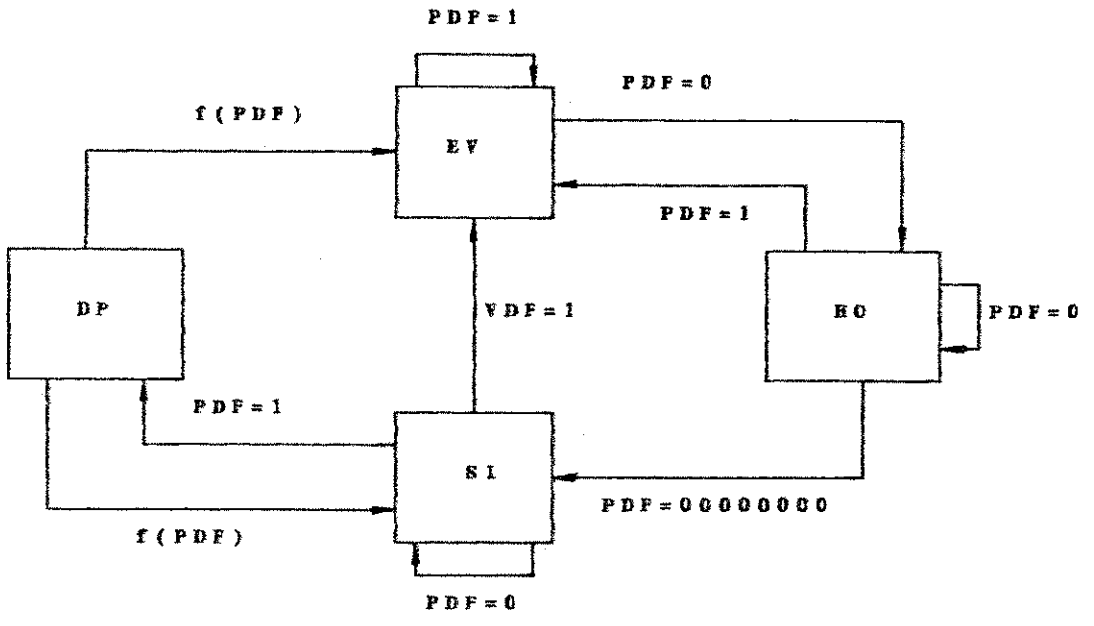


FIGURA 5.3 - Diagrama de Estados do Detetor de Voz.

A análise do diagrama de estados pode ser iniciada a partir da determinação do início da fala em uma conversação. O estado SI é estabelecido quando o sinal de voz é de baixa intensidade (En < E1). O estado permaneçe em SI enquanto En for menor que E1 (PDF^m = 0). O estado muda de SI para DP apenas quando PDF^m = 1. Isto significa que o segmento em análise será considerado como silêncio ou voz e o estado resultante como SI ou EV, respectivamente, dependendo das combinações dos quatro últimos estados de PDF. Isto é feito para evitar que um segmento seja definido errôneamente como voz ou silêncio.

A mudança de SI para EV ocorre diretamente, apenas se En for maior que E3. Se o estado EV for

estabelecido, o detetor de voz apresentará nível lógico 1 (estado ON) em sua saída. Enquanto PDF assumir o estado lógico 1, **EV** é mantido.

O processo de determinação do final de uma declaração é complexo pois foi visto que podem ser cometidos êrros. O **hangover** é utilizado nesse caso, para evitar transições errôneas de **EV** para **SI**. Durante a ação do **hangover**, o estado estabelecido é **HO** e a saída é mantida com nível lógico 1. Não há, portanto, transições diretas de **EV** para **SI**. Para que o estado se desloque de **EV para HO**, basta que PDF^m assuma nível lógico 0.

O estado é transferido de **HO** para **SI** apenas se ocorrerem sucessivos valores de PDF = 0 , correspondentes ao tempo de **hangover**. Aqui são necessários 5 quadros (20 ms) ou 8 quadros (32 ms) seguidos com PDF = 0, dependendo do **hangover** utilizado. A transição de **HO** para **SI** é efetuada somente após decorrido o tempo de **hangover**. Se, durante o tempo de **hangover**, PDF assumir o nível lógico 1, o estado deve retornar a **EV** e o processo anterior é retomado.

A tabela 5.2 apresenta as combinações possíveis de PDF,bem como a transição de estados e a saída associadas a essas combinações. A partir das combinações de PDF, é possível obter uma função que produza a saída desejada, utilizando-se o Mapa de Karnaugh para quatro variáveis. A figura 5.4 mostra o Mapa com as combinações possíveis.

As letras A, B, C e D representam,

respectivamente, PDF^{m}, PDF^{m-1}, PDF^{m-2} e PDF^{m-3}. A função obtida pelas combinações de PDF, utilizando-se a técnica dos Mapas de Karnaugh, é apresentada na equação 5.1:

$$edf = \{[PDF^{m} \underline{E} (PDF^{m-1} + PDF^{m-2} + PDF^{m-3})] + [PDF^{m-1} \underline{E} PDF^{m-2} \underline{E} PDF^{m-3}]\} \quad (5.1)$$

PDF^{m}	PDF^{m-1}	PDF^{m-2}	PDF^{m-3}	SA
0	0	0	0	0
0	0	0	1	0
0	0	1	0	0
0	0	1	1	0
0	1	0	0	0
0	1	0	1	0
0	1	1	0	0
0	1	1	1	1
1	0	0	0	0
1	0	0	1	1
1	0	1	0	1
1	0	1	1	1
1	1	0	0	1
1	1	0	0	1
1	1	0	1	1
1	1	1	1	1

TABELA 5.2 - Combinação dos Estados de PDF, Transição dos Estados e Nível Lógico na Saída do Detetor.

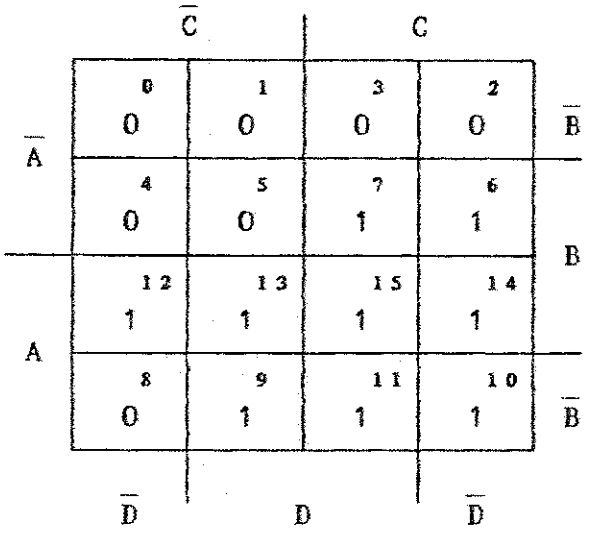


FIGURA 5.4 Mapa de Karnaugh com as combinaçoes de PDF.

5.2.2.3 Valores dos Limiares de Energia a Curtos Intervalos de Tempo

Os valores de limiar de energia foram estabelecidos a partir da anàlise da energia a curtos intervalos de tempo, medida ao longo de todo o sinal de voz. Na figura 5.5a é apresentado um intervalo que contêm sinal de voz e ruído. A figura 5.5b mostra o nível de energia do ruído obtido em um intervalo de silêncio, onde apenas o ruído está presente, e a figura 5.5c apresenta o nível de energia de uma região que contêm sinal de voz.

Yatsuzuka utilizou os seguintes valores de limiar de energia [35]:

E1 = - 51 dBm;

E2 = - 39 dBm;

E3 = - 30 dBm;

onde a diferença entre o maior e o menor valor de limiar de energia é de 21 dB. Neste trabalho, a excursão de amplitude do sinal de voz não é tão grande, o que impede a utilização dos valores apresentados anteriormente. As diferenças utilizadas entre os valores de limiar são de 3 dB entre E3 e E2 e 4.2 dB entre E2 e E1.

Três conjuntos de dados foram utilizados para estabelecer os valores de limiar de energia. O primeiro conjunto de dados, o mais sensível, utilizou o valor do limiar inferior de energia (E1), próximo ao nível de ruído. Será visto que, neste caso, os surtos de voz tem duração maior que nos demais conjuntos. É possível que alguns surtos de voz tenham sido obtidos, a partir da ação do sinal de ruído no detetor.

Os segundo e terceiro conjuntos de dados foram estabelecidos tomando-se um valor 3 dB e outro 5 dB, respectivamente, acima do valor de E1 no primeiro conjunto.

Os valores de limiar de energia utilizados são os seguintes:

LIMIAR	1º CONJUNTO	2º CONJUNTO	3º CONJUNTO
E1	40 dB	43 dB	45 dB
E2	44.2 dB	47.2 dB	49.2 dB
E3	47.2 dB	50.2 dB	52.2 dB

Tabela 5.3 - Limiares de Energia Utilizados

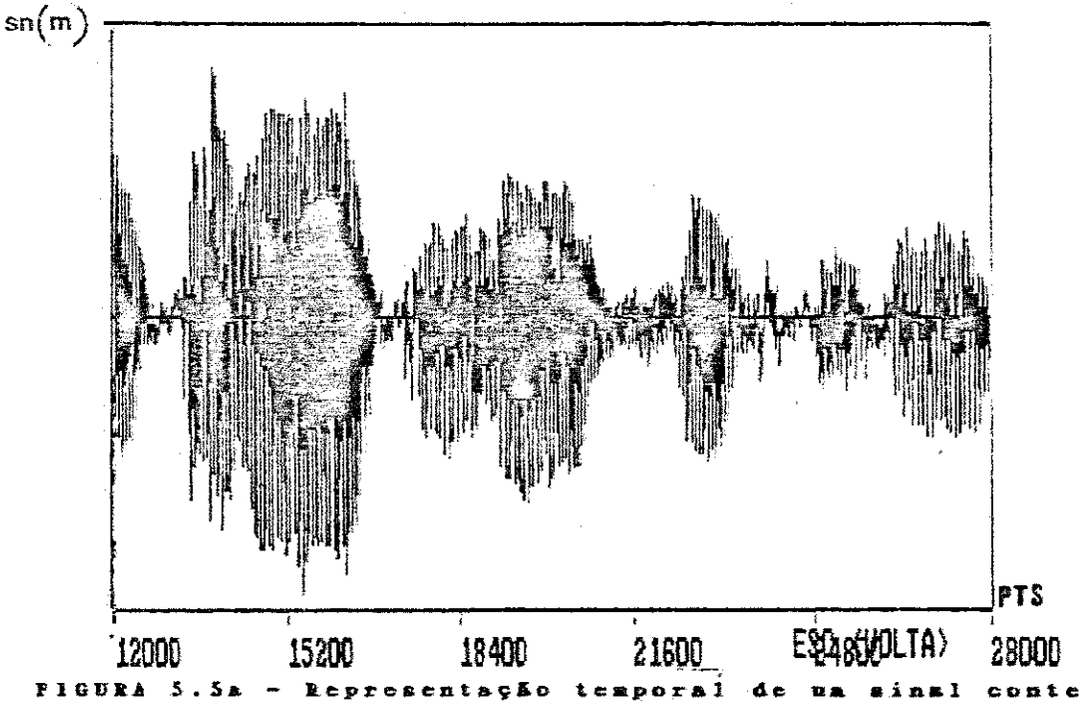


FIGURA 5.5a - Representação temporal de um sinal contendo voz e ruído, em uma conversação.

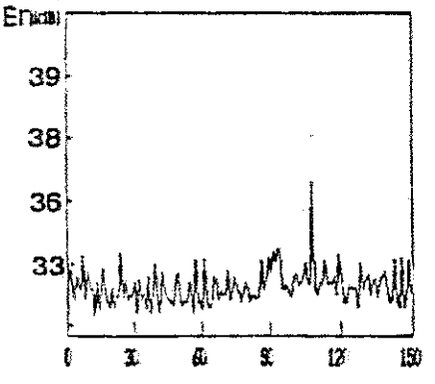


FIGURA 5.5b - Energia do sinal de ruído a curtos intervalos de tempo em 600 ms.

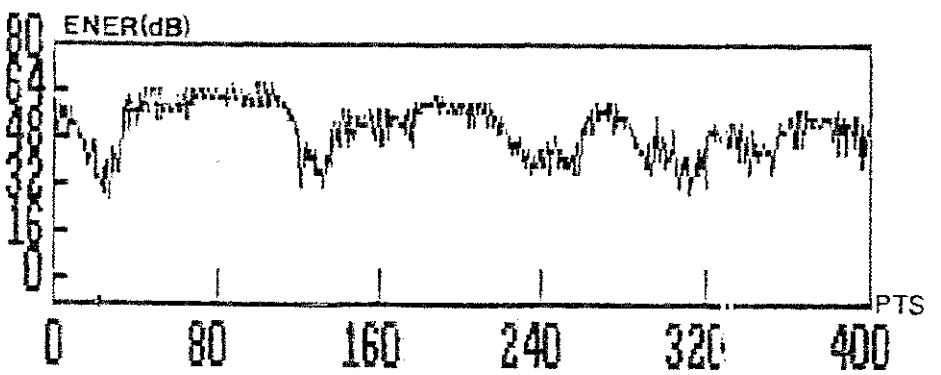


FIGURA 5.5c - Energia do sinal de voz a curtos intervalos de tempo em um período de 1,6 s.

Os valores de limiar foram obtidos a partir da análise das medidas de energia, procurando-se tirar uma proporcionalidade entre os limiares utilizados por Yatsuzuka. É fácil observar que sinais com níveis de energia acima de E3, indicam a existência de sinais de voz. Abaixo desse valor prepondera a incerteza quanto à existência de um segmento de voz ou ruído. Daí a utilização dos testes com os demais parâmetros temporais.

É óbvio que estes valores de limiar de energia podem ser otimizados, mas os testes efetuados parecem indicar que são adequados aos objetivos desejados.

5.2.2.4 Valores da Taxa de Cruzamento por Zero a Curtos Intervalos de Tempo

Os valores da taxa de cruzamento por zero foram obtidos a partir da análise de intervalos de silêncio e de voz. A figura 5.6 mostra o gráfico deste parâmetro, considerando-se um intervalo de tempo de 1,6 segundos. Analisando-se vários intervalos de silêncio, onde apenas o sinal de ruído está presente, pôde ser verificado que ZCR varia em torno de 27.

Após análise da taxa de cruzamento por zero em vários segmentos contendo sinal de voz e silêncio, optou-se pela utilização do intervalo 27±9 para os limiares superior e inferior, que resulta nos valores Z1 = 36 e Z2 = 18. Assim, valores de ZCR maiores que 36 indicam a presença de um segmento contendo um sinal de voz constituído por um som fricativo. Se ZCR for inferior a 18, pode-se considerar que o segmento em análise é constituído de um som sonoro ou nasal. Se o valor de ZCR estiver entre 18 e 36, considera-se que o segmento em análise é um sinal de ruído.

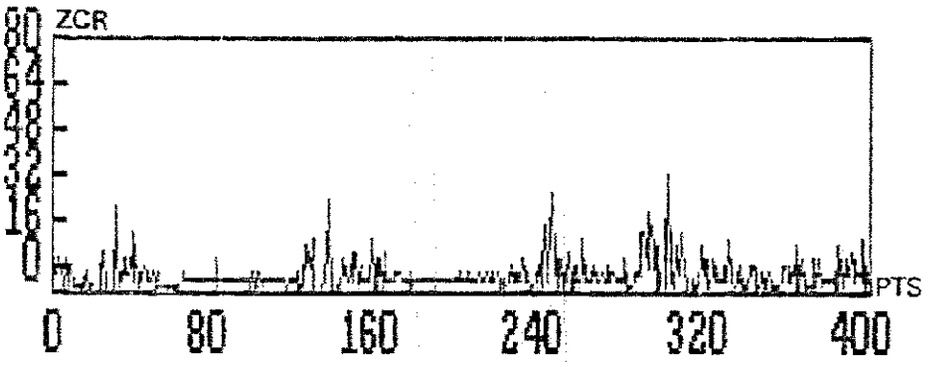


FIGURA 5.6 - Taxa de Cruzamento por Zero a Curtos Intervalos de Tempo em um período de 1,6 s.

É importante considerar ainda que a taxa de cruzamento por zero e a variação de energia, são parâmetros temporais que irão atuar de forma subseqüente à ação da energia, na definição da existência de voz ou pausa no segmento em análise.

5.3 Análise Estatística dos Resultados

Os resultados apresentados a seguir, foram obtidos a partir do processamento dos sinais de voz das vinte conversações telefônicas simuladas, pelo algoritmo que determina o sinal ON-OFF. O processamento utilizou os valores de 20 ms e 32 ms para o hangover, sendo que em cada um desses casos foram utilizados os três conjuntos de limiar de energia definidos anteriormente. Ao todo foram utilizadas, portanto, seis situações de processamento.

São apresentadas as médias dos resultados dos eventos para cada um dos grupos participantes. Como foi visto anteriormente, as vinte conversações utilizadas estão divididas em quatro grupos dispostos da seguinte forma:

GRUPO I - Gravações Unilaterais Femininas.
GRUPO II - Gravações Unilaterais Masculinas.
GRUPO III - Dupla Conversação Feminina.
GRUPO IV - Dupla Conversação Masculina.

Inicialmente são apresentados os resultados obtidos pelo processamento das conversações, utilizando-se o hangover de 20 ms.

Posteriormente são apresentados os resultados obtidos com a utilização do hangover de 32 ms.

Foram efetuadas medidas apenas nos intervalos de atividade de voz. Os resultados foram tabelados, para melhor visualização, sendo que as medidas efetuadas nos intervalos de surto de voz são apresentadas em tabelas diferentes das realizadas nos intervalos de pausa, em virtude da grande quantidade de dados obtida. As conclusões e os comentários posteriores, apresentados em seção específica, procuram dar um sentido aos resultados obtidos.

5.3.1 Estatísticas dos Eventos para Surtos de Voz e Pausas

Os eventos medidos nos intervalos de voz são os seguintes:

- Número Total de Surtos de Voz : **N**;
- Soma Total dos Quadros de Voz : **Sv** (quadros);
- Duração Total dos Surtos de Voz :

  $$Tsv = \Sigma Vi = Sv \cdot 4 \cdot 10^{-4} \quad (s);$$

- Duração Média dos Surtos de Voz : **Msv = Tsv/N** (s);
- Média do Número de Quadros por Surto de Voz :

  MNs = Sv/N;

- Desvio Padrão dos Quadros por Surto de Voz : **DPv**.

As tabelas de 5.4 a 5.9 apresentam os resultados obtidos para os eventos medidos nos intervalos de voz, sendo utilizados os valores de 20 ms e 32 ms para o hangover, considerando-se os três conjuntos de limiares:

GRUPO	N	Sv quadro	Tsv s	Msv s	MNs	DPv
CONVERSA UNILATERAL FEMININA	50	4945	19,77	0,395	99	152
CONVERSA UNILATERAL MASCULINA	52	5748	22,99	0,449	112	194
DUPLA CONVERSA FEMININA	44	5007	20,03	0,449	112	163
DUPLA CONVERSA MASCULINA	55	5376	21,50	0,401	100	178

TABELA 5.4 - Valores dos Eventos, obtidos a partir do uso do Conjunto I de Limiares de energia e hangover de 20 ms.

GRUPO	N	Sv quadro	Tsv s	Msv s	MNs	DPv
CONVERSA UNILATERAL FEMININA	69	4330	17,32	0,316	82	104
CONVERSA UNILATERAL MASCULINA	71	5287	19,15	0,358	89	116
DUPLA CONVERSA FEMININA	56	4872	19,48	0,467	117	130
DUPLA CONVERSA MASCULINA	73	4932	19,72	0,322	81	109

TABELA 5.5 - Valores dos Eventos, obtidos a partir do uso do Conjunto II de Limiares de energia e hangover de 20 ms.

GRUPO	N	Sv quadro	Tsv s	Msv s	MNs	DPv
CONVERSA UNILATERAL FEMININA	89	3820	15,28	0,189	47	66
CONVERSA UNILATERAL MASCULINA	95	4484	17,93	0,196	49	73
DUPLA CONVERSA FEMININA	81	4246	16,98	0,218	55	70
DUPLA CONVERSA MASCULINA	100	3979	15,92	0,161	40	58

TABELA 5.6 - Valores dos Eventos, obtidos a partir do uso do Conjunto III de Limiares de energia e hangover de 20 ms.

GRUPO	N	Sv quadro	Tsv s	Msv s	MNs	DPv
CONVERSA UNILATERAL FEMININA	50	5110	20,44	0,950	185	170
CONVERSA UNILATERAL MASCULINA	35	5854	23,41	0,768	192	241
DUPLA CONVERSA FEMININA	29	5174	20,69	0,756	191	199
DUPLA CONVERSA MASCULINA	36	5373	21,49	0,782	196	222

TABELA 5.7 - Valores dos Eventos, obtidos a partir do uso do Conjunto I de Limiares de energia e hangover de 32 ms.

GRUPO	N	Sv quadro	Tsv s	Msv s	MNs	DPv
CONVERSA UNILATERAL FEMININA	36	4671	18,68	0,642	160	157
CONVERSA UNILATERAL MASCULINA	33	5495	21,98	0,765	191	189
DUPLA CONVERSA FEMININA	29	4867	19,47	0,927	232	156
DUPLA CONVERSA MASCULINA	36	5118	20,47	0,716	179	185

TABELA 5.8 - Valores dos Eventos, obtidos a partir do uso do Conjunto II de Limiares de energia e hangover de 32 ms.

GRUPO	N	Sv quadro	Tsv s	Msv s	MNs	DPv
CONVERSA UNILATERAL FEMININA	60	4151	16,60	0,290	72	94
CONVERSA UNILATERAL MASCULINA	59	4813	19,25	0,349	87	114
DUPLA CONVERSA FEMININA	53	4518	18,07	0,372	93	99
DUPLA CONVERSA MASCULINA	62	4434	17,73	0,298	75	93

TABELA 5.9 - Valores dos Eventos, obtidos a partir do uso do Conjunto III de Limiares de energia e hangover de 32 ms.

Os eventos medidos nos intervalos de pausa são os seguintes:

- Número Total de Pausas : **P**;
- Soma Total dos Quadros de Pausa : **Sp** (quadros);
- Duração Total das Pausas :

$$Tp = \sum Si = Sp \cdot 4 \cdot 10^{-4} (s);$$

- Duração Média das Pausas : **Mp** = **Tp/P** (s);
- Média do Número de Quadros por Pausa : **MNp** = **Sp/P**;
- Desvio Padrão dos Quadros por Pausa : **DPp**.

As tabelas de 5.10 a 5.15 apresentam os resultados obtidos para os eventos medidos nos intervalos de pausa, sendo utilizados os valores de 20 ms e 32 ms para o hangover, considerando-se os três conjuntos de limiares.

GRUPO	P	Sp quadro	Tp s	Mp s	MNp	DPp
CONVERSA UNILATERAL FEMININA	43	878	3,51	0,074	18	27
CONVERSA UNILATERAL MASCULINA	46	911	3,65	0,076	19	29
DUPLA CONVERSA FEMININA	38	819	3,28	0,084	21	40
DUPLA CONVERSA MASCULINA	51	717	4,07	0,072	18	25

TABELA 5.10 - Valores dos Eventos, obtidos a partir do uso do Conjunto 1 de Limiares de energia e hangover de 20 ms.

GRUPO	P	Sp quadro	Tp s	Mp s	MNp	DPp
CONVERSA UNILATERAL FEMININA	67	1298	5,193	0,1137	28	56
CONVERSA UNILATERAL MASCULINA	68	1247	4,988	0,0897	22	42
DUPLA CONVERSA FEMININA	53	1151	4,604	0,1182	30	49
DUPLA CONVERSA MASCULINA	71	1377	5,508	0,0903	23	43

TABELA 5.11 - Valores dos Eventos, obtidos a partir do uso do Conjunto II de Limiares de energia e hangover de 20 ms.

GRUPO	P	Sp quadro	Tp s	Mp s	MNp	DPp
CONVERSA UNILATERAL FEMININA	85	1851	7,405	0,1048	26	46
CONVERSA UNILATERAL MASCULINA	96	1964	7,856	0,0823	21	38
DUPLA CONVERSA FEMININA	80	1636	6,545	0,0840	21	43
DUPLA CONVERSA MASCULINA	101	2190	8,759	0,0896	22	42

TABELA 5.12 - Valores dos Eventos, obtidos a partir do uso do Conjunto III de Limiares de energia e hangover de 20 ms.

GRUPO	P	Sp quadro	Tp s	Mp s	MNp	DPp
CONVERSA UNILATERAL FEMININA	22	728	2,912	0,1036	26	34
CONVERSA UNILATERAL MASCULINA	28	842	3,367	0,1103	27	39
DUPLA CONVERSA FEMININA	21	728	2,913	0,1346	34	54
DUPLA CONVERSA MASCULINA	31	981	3,922	0,1353	34	41

TABELA 5.13 - Valores dos Eventos, obtidos a partir do uso Conjunto I de Limiares de energia e hangover de 32 ms.

GRUPO	P	Sp quadro	Tp s	Mp s	MNp	DPp
CONVERSA UNILATERAL FEMININA	32	1045	4,181	0,1803	45	64
CONVERSA UNILATERAL MASCULINA	31	1131	4,526	0,1719	43	53
DUPLA CONVERSA FEMININA	27	920	3,681	0,1843	46	53
DUPLA CONVERSA MASCULINA	35	1279	5,116	0,1787	45	53

TABELA 5.14 - Valores dos Eventos, obtidos a partir do uso do Conjunto II de Limiares de energia e hangover de 32 ms.

GRUPO	P	Sp quadro	Tp s	Mp s	MNp	DPp
CONVERSA UNILATERAL FEMININA	59	1553	6,213	0,1216	30	53
CONVERSA UNILATERAL MASCULINA	60	1778	7,114	0,1260	31	47
DUPLA CONVERSA FEMININA	51	1232	4,926	0,1005	25	45
DUPLA CONVERSA MASCULINA	63	1941	7,766	0,1360	34	49

TABELA 5.15 - Valores dos Eventos, obtidos a partir do uso do Conjunto III de Limiares de energia e hangover de 32 ms.

As tabelas 5.16 e 5.17 apresentam os resultados obtidos para os seguintes eventos de voz, considerando-se um hangover de 20 ms e outro de 32 ms, respectivamente:

- Intervalo de Observação : T = Tsv + Tp (s);
- Taxa de Surto de Voz Média :

  TS = N / T (ciclos/s);
- Atividade de Voz : Av.

GRUPO	T (s)			TS (ciclos/s)			Av		
	L I M I A R			L I M I A R			L I M I A R		
	I	II	III	I	II	III	I	II	III
CONVERSA UNILATERAL FEMININA	23,3	22,5	22,7	2,15	3,06	3,92	0,84	0,73	0,64
CONVERSA UNILATERAL MASCULINA	26,6	24,1	25,8	1,95	2,94	3,68	0,86	0,80	0,70
DUPLA CONVERSA FEMININA	23,3	24,0	23,5	1,89	2,33	3,44	0,84	0,80	0,72
DUPLA CONVERSA MASCULINA	25,6	25,2	24,7	2,15	2,89	3,65	0,85	0,78	0,64

TABELA 5.16 - Valores dos Eventos de Voz (hangover =20 ms).

GRUPO	T (s)			TS (ciclos/s)			Av		
	L I M I A R			L I M I A R			L I M I A R		
	I	II	III	I	II	III	I	II	III
CONVERSA UNILATERAL FEMININA	23,4	22,9	22,8	2,14	1,57	2,63	0,90	0,78	0,70
CONVERSA UNILATERAL MASCULINA	26,8	26,5	26,4	1,31	1,24	2,24	0,87	0,82	0,74
DUPLA CONVERSA FEMININA	23,6	23.2	23,0	1,23	1,68	2,74	0,85	0,83	0,79
DUPLA CONVERSA MASCULINA	25,4	25,6	25,5	1,42	1,41	2,43	0,85	0,80	0,69

TABELA 5.17 - Valores dos Eventos de Voz (hangover =32 ms).

5.3.2 Análise dos Resultados e Conclusões

As quatorze tabelas apresentadas na seção 5.3.1, forneçem os resultados dos eventos de voz de interesse, obtidos em uma conversação telefônica simulada. A seguir é feita uma análise dos resultados obtidos.

I) À medida em que os limiares de energia são elevados, tornando o detetor menos sensível, os eventos de voz tem sua duração diminuída, uma vez que os surtos de voz são divididos em outros de menor duração; observando-se os valores encontrados para os eventos **Soma Total dos Quadros de Voz (Sv)**, **Duração Total dos Surtos de Voz (Tsv)** e **Duração Média dos Surtos de Voz (Msv)** nas Tabelas 5.4, 5.5 e 5.6 (**hangover** de 20 ms), tomando o caso da **conversação unilateral feminina**: para os Conjuntos I, II e III de limiares de energia, os valores de **Sv** são 4945, 4330 e 3820 quadros; os de **Tsv** são 19,77 s; 17,32 s e 15,28 s e os de **Msv** são 0,395 s; 0,316 s e 0,189 s, respectivamente. O mesmo efeito pode ser observado nos demais grupos participantes dos testes.

II) Considerando ainda os eventos **Soma Total dos Quadros**, **Duração Total dos Surtos de Voz** e **Duração Média dos Surtos de Voz**, os resultados apresentados nas tabelas

5.7, 5.8 e 5.9 (**hangover** de 32 ms), para o caso da **conversação unilateral feminina**, comprovam a diminuição dos eventos de voz, à medida em que os limiares de energia são elevados: para os Conjuntos I, II e III de limiares, os valores de **Sv** são 5110, 4671 e 4151 quadros; os de **Tsv** são 20,44 s; 18,68 s e 16,60 s e os de **Msv** são 0,950 s; 0,642 s e 0,290 s, respectivamente. O mesmo efeito pode ser observado nos demais grupos participantes dos testes.

III) Comparando as tabelas 5.4, 5.5 e 5.6 com as tabelas 5.7, 5.8 e 5.9, respectivas, considerando-se os diferentes valores de **hangover**, pode-se observar, como era de se esperar, que os resultados obtidos para os eventos **Soma Total dos Quadros (Sv)**, **Duração Total dos Surtos de Voz (Tsv)** e **Duração Média dos Surtos de Voz (Msv)**, são maiores quando se aplica o **hangover** de 32 ms, em relação ao de 20 ms.

IV) Para o caso do evento **Sv**, é possível considerar que, em média, os valores obtidos são 1,021 (Conjunto I de limiares), 1,038 (Conjunto II de limiares) e 1,084 (Conjunto III de limiares) vezes maiores quando se utiliza o **hangover** de 32 ms, em relação ao de 20 ms.

V) Para o evento **Tsv**, os valores obtidos são, em média, 0,435 s (Conjunto I de limiares), 1,232 s (Conjunto II de limiares) e 1,887 s (Conjunto III de limiares) maiores quando se utiliza o hangover de 32 ms, em relação ao de 20 ms.

VI) Para o evento **Msv**, os valores obtidos são, em média, 0,390 s (Conjunto I de limiares), 0,397 s (Conjunto II de limiares) e 0,136 s (Conjunto III de limiares) maiores quando se utiliza o hangover de 32 ms, em relação ao de 20 ms.

VII) A respeito da utilização de um hangover menor (150 ms a 250 ms nos detetores convencionais contra 20 ms e 32 ms neste trabalho), verifica-se, através das tabelas de 5.4 a 5.9, que houve uma redução considerável nos valores dos **surtos de voz**, conforme previsto. Nos detetores convencionais a duração média desses eventos está em torno de 1 segundo. Neste trabalho, considerando a média dos valores obtidos para a **duração média dos surtos de voz** dos grupos, em cada conjunto de limiares de energia, os resultados foram os seguintes:

- Hangover de 20 ms:

  - Conjunto I - $\overline{Msv} = 0,423$ s;
  - Conjunto II - $\overline{Msv} = 0,365$ s;
  - Conjunto III - $\overline{Msv} = 0,191$ s.

- Hangover de 32 ms:
  - Conjunto I - $\overline{Msv}$ = 0,814 s;
  - Conjunto II - $\overline{Msv}$ = 0,762 s;
  - Conjunto III - $\overline{Msv}$ = 0,327 s.

Obtêm-se com isso, o objetivo de reduzir a duração média dos surtos, visando a diminuição do enfraquecimento do sinal de voz.

A despeito dos parâmetros utilizados por Brady, serem diferentes dos aqui utilizados, pode-se concluir que os valores da duração média dos surtos de voz (Msv) obtidos em seus trabalhos [20,27], são maiores. Brady obteve os valores médios de Msv iguais a 1,366 s para o menor limiar de energia, 1,197 s para o limiar intermediário e 0,980 s para o maior valor.

Em [18], Gruber encontrou os valores de 2,360 s e 2,157 s para a duração média dos surtos de voz, utilizando, respectivamente, hangover e tempo de preenchimento. Os valores desses parâmetros utilizados em [18], foram os mesmos e iguais a 202,5 ms.

VIII) O aumento do Número Total de Surtos de Voz (N) é uma tendência observada, à medida em que os limiares de energia são elevados. Este efeito é mais pronunciado quanto menor for o valor do hangover. Comparando-se as tabelas 5.7 e 5.8 (hangover de 32 ms), é possível observar que, no caso das conversações unilaterais femininas e masculinas

apesar do aumento dos limiares de energia, o valor do **Número Total de Surtos (N)** no Conjunto II (Tabela 5.8), é inferior ao obtido no Conjunto I (Tabela 5.7). Este fato ocorre porque o sinal de voz de alguns participantes é mais intenso que o de outros e o aumento dos valores dos limiares de energia causam apenas a eliminação dos surtos de voz de menor intensidade. Quando esses surtos de níveis menos intensos são de curta duração, o efeito causado é o da diminuição do número de surtos, pela eliminação desses mais fracos. No sentido contrário, os surtos de voz mais intensos, com níveis acima dos valores dos limiares de energia, constituem os surtos de maior duração, não sendo particionados, como normalmente ocorre, criando-se, assim, uma menor quantidade de número de surtos. Pode-se notar, no entanto, que a regra de aumento do número de surtos é seguida, quando há uma elevação considerável dos valores de limiar de energia (Conjunto III, por exemplo), conforme é observado na Tabela 5.9. Analisando-se as colunas referentes ao número de surtos de voz nas Tabelas 5.4, 5.5 e 5.6, no caso do hangover igual a 20 ms, verifica-se que a elevação dos limiares de energia produz o conseqüente acréscimo no número de surtos de voz, pelo efeito do seu fracionamento.

IX) Ao contrário, o **Número Total de Surtos de Voz (N)** diminui com o aumento do valor do hangover, o que já era esperado. Comparando as Tabelas 5.4, 5.5 e 5.6 com as

Tabelas 5.7, 5.8 e 5.9, respectivamente, é possível observar que, em média, os valores de **N** são 0,746 (Conjunto I de limiares), 0,499 (Conjunto II de limiares) e 0,642 (Conjunto III de limiares) menores quando se utiliza um **hangover** de 32 ms, em relação ao de 20 ms.

X) A análise da **Média do Número de Quadros por Surto de Voz (MNs)**, mostra a diminuição de seu valor com o aumento dos valores de limiares de **energia**. Isto também era esperado porque, quando o detetor de voz torna-se menos sensível, restringe-se o número de quadros considerados como voz, ocasionando a diminuição do evento **Soma Total dos Quadros de Voz (Sv)** e o aumento do **Número Total de Surtos (N)**. Como o valor de **MNs** é a relação entre **Sv** e **N**, e os valores do evento do denominador desta relação aumentam mais rapidamente que os do numerador, a tendência é, portanto, de uma diminuição de **MNs**. Uma exceção ocorre nos testes efetuados com **dupla conversação feminina**, comparando-se especificamente os Conjuntos I e II (Tabelas 5.4 e 5.5 para **HO** = 20 ms e Tabelas 5.7 e 5.8 para **HO** = 32 ms) de limiares de **energia**; neste caso, o número de surtos de voz **N** cresce mais lentamente que **Sv**, pelas razões expostas no item VII, ocasionando um aumento no valor de **MNs**, mesmo com o aumento dos limiares de **energia**. A tendência, no entanto, é, nitidamente, contrária.

XI) Comparando-se os resultados de **MNs**, verifica-se que há um aumento na **Média do Número de Quadros por Surto de Voz**, com o aumento do **hangover**. Em média, pode-se afirmar que os valores de **MNs** são 1,748 (Conjunto I de limiares), 2,072 (Conjunto II de limiares) e 1,718 (Conjunto III de limiares) vezes maiores quando se utiliza o **hangover** de 32 ms, em relação ao de 20 ms. Isto ocorre porque quando se utiliza valores maiores para o **hangover**, a **Soma Total de Quadros de Voz (Sv)** aumenta e o **Número Total de Surtos de Voz (N)** diminui. Como o valor de **MNs** é obtido através da relação entre **Sv** e **N**, a afirmativa anterior fica clara.

XII) Observando-se os resultados obtidos, nota-se uma quantidade ligeiramente maior de quadros de surtos de voz no caso de **conversações unilaterais masculinas**, do que nas femininas, independentemente do valor do **hangover**. A **Duração dos Surtos de Voz (Tsv)**, a **Duração Média dos Surtos (Msv)** e a **Média do Número de Quadros por Surto de Voz (MNs)** também são maiores para o caso masculino. Isto não significa, no entanto, que se possa concluir como uma tendência, o fato de eventos com vozes masculinas terem atividades maiores que os eventos com vozes femininas. Os resultados podem ser explicados levando-se em conta que as gravações das vozes masculinas foram mais aproveitadas que as das vozes femininas. Para que se tenha certeza dessa afirmação, basta verificar os valores obtidos nos

Intervalos de Observação, apresentados nas Tabelas 5.16 (**hangover** de 20 ms) e 5.17 (**hangover** de 32 ms), e que serão abordados mais à frente. Pode-se concluir o mesmo para o caso de **dupla conversação masculina** ou **feminina**.

XIII) É importante mencionar que as medidas efetuadas nos intervalos que não contêm voz, referem-se aos eventos de pausa. Compreeende-se com isto que os eventos de interesse, neste caso, são as pausas intersilábicas ou entre palavras, que não tenham sido eliminadas pela ação do **hangover**. Deste modo, apenas após o primeiro surto de voz ter sido computado, foi iniciada a contagem das pausas. É evidente que este processo acarreta diminuição no número de quadros de pausa, e até mesmo êrros, mas foi necessário ser adotado para que não houvessem dúvidas quanto ao fato das pausas anteriores ao primeiro surto serem realmente pausas existentes em uma conversação, ou terem sido originadas no processo de digitalização, uma vez que o tempo de resposta do processador TMS, a partir de seu acionamento, é bastante superior ao tempo de resposta do aparelho reprodutor da fita cassete que armazena as conversações.

XIV) Analisando-se os eventos medidos nos intervalos de pausa, é possível constatar, a partir da investigação das Tabelas 5.10, 5.11 e 5.12, correspondentes aos resultados obtidos utilizando-se **hangover** de 20 ms, e das

Tabelas 5.13, 5.14 e 5.15, utilizando-se **hangover** de 32 ms, o efeito contrário ao obtido nos intervalos de surto de voz. O **Número Total de Pausas (N)**, a **Soma Total dos Quadros de Pausa (Sp)** e a **Duração das Pausas (Tp)** crescem com o aumento dos valores dos limiares de energia.

XV) Comparando o evento **Duração Média das Pausas (Mp)**, é possível notar que a tendência é de seu aumento, à medida em que os limiares de energia são elevados. Investigando as Tabelas 5.11 e 5.12 (**hangover** de 20 ms) e as Tabelas 5.13 e 5.14 (**hangover** de 32 ms), verifica-se que, nesses casos, o aumento dos valores de limiares de energia ocasionou efeito contrário ao esperado. Como **Mp** é obtido a partir da relação entre **Tp** e **P**, observando esses valores conclui-se que os intervalos médios de pausa são menores, mas o número desses intervalos é maior no Conjunto III de limiares, em relação ao Conjunto II.

XVI) No caso do evento **Média do Número de Quadros por Pausa (MNp)**, seus resultados são determinados através da relação entre **Sp** e **P**. Observando os dados obtidos em ambos os casos de **hangover**, verifica-se que, após um crescimento nos valores de **MNp** ao passar do Conjunto I para o Conjunto II de limiares, os resultados decrescem a seguir. Isto ocorre porque, após serem elevados os limiares de energia do Conjunto II para o Conjunto III, **Sp** cresce menos rapidamente que **P**, o que torna a **Média do Número de**

Quadros por Pausa menor.

XVII) Analisando as Tabelas 5.16 e 5.17, verifica-se que o **Intervalo de Observação** médio é de 24 segundos, em ambos os casos de **hangover**. A **Taxa de Surto de Voz Média (TS)** é crescente com o aumento dos limiares de **energia**. No caso do uso do **hangover** de 32 ms, quando se passa do Conjunto I para o Conjunto II de limiares, há uma diminuição no valor de **TS** para as **conversações unilaterais masculina e feminina**. Isso pode ser explicado pelo fato do **Número Total de Surtos de Voz (N)** ser menor no Conjunto II que no Conjunto I, conforme explicado no item VII.

XVIII) Analisando-se o evento **Atividade de Voz** é possível comprovar que há uma diminuição do número de surtos de voz detetados em uma conversação, à medida em que os limiares de **energia** são elevados. A diminuição do valor do **hangover** também causa redução na **atividade de voz**. Utilizando-se o **hangover** de 20 ms, as médias dos valores de **Atividade de Voz**, percentualmente, foram de 84%, 78% e 67% para os Conjuntos I, II e III, respectivamente. No caso do **hangover** de 32 ms, a **atividade de voz** aumentou um pouco e os valores médios encontrados foram de 86%, 80% e 73%, para os Conjuntos I, II e III, respectivamente. As médias foram obtidas tomando-se os valores encontrados para cada um dos grupos de teste.

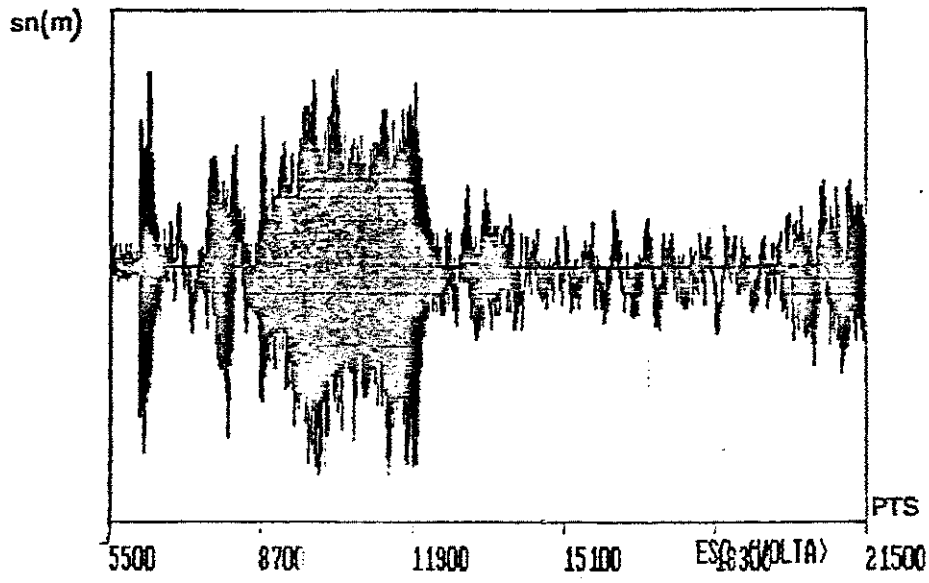
Em [27], Brady, empregando o tempo de preenchimento como fator de correção do sinal de voz, não determinou o evento **Atividade de Voz**, mas um cálculo simples utilizando os valores das durações médias do surto de voz e do silêncio, medidos em seu trabalho, permite obter os percentuais de 43%, 39% e 36% para o maior, o intermediário e o menor valor de limiar de energia, respectivamente.

Em [18], o evento **Atividade de Voz** determinado por Gruber, foi de 79,56% e 72,73% para o hangover e o tempo de preenchimento, respectivamente.

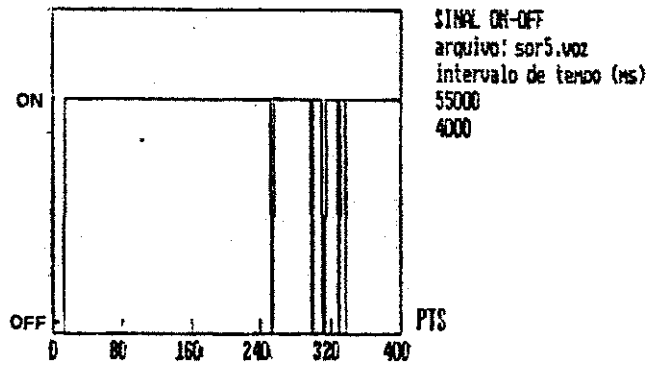
É importante mencionar que Brady em [27] e Gruber em [18] utilizaram os intervalos de silêncio em suas medidas e não pausas.

XIX) É importante mencionar ainda, que os testes com **dupla conversação**, não referem-se apenas aos intervalos em que ambos interlocutores estão falando, mas a toda conversação.

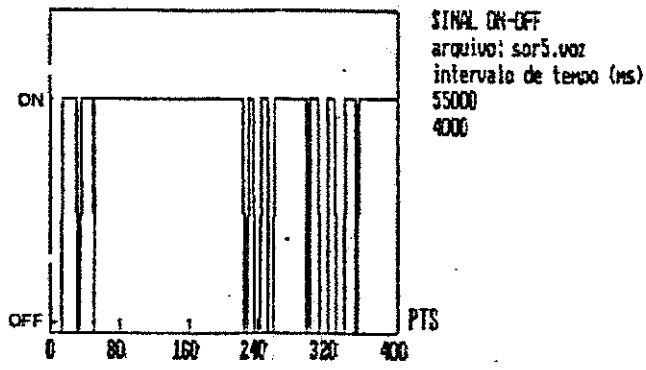
XX) A figura 5.7 apresenta o sinal de voz de uma conversação simulada, utilizado nos testes realizados neste trabalho, e os sinais ON-OFF obtidos a partir de seu processamento pelo detetor de voz proposto. A amostra de voz utilizada tem a duração de 1,6 segundos.



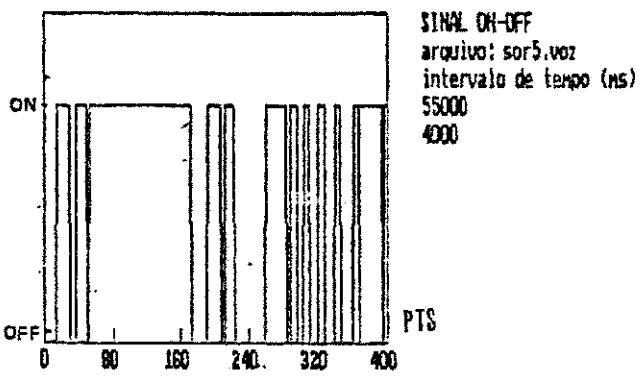
a) Sinal de Voz.



b) Sinal ON-OFF (Conjunto I de Limiares de Energia).



c) Sinal ON-OFF (Conjunto II de Limiares de Energia).



d) Sinal ON-OFF (Conjunto III de Limiares de Energia).

FIGURA 5.7 - Sinal de Voz e Sinais ON-OFF resultantes na saída do detetor. Hangover de 20 ms.

XXI) Analisando-se os valores médios de surto de voz encontrados nas tabelas de 5.4 a 5.9, pode-se verificar que os resultados obtidos para o Conjunto III de limiares de energia, menos sensível ao sinal de voz, é muito reduzido, especialmente quando se utiliza um hangover de 20 ms. Considerando-se que o surto de voz deve ter a duração de uma frase ou sentença, e que este dado é desconhecido para a língua portuguesa, pode-se estimar que o valor ideal de surto de voz, está situado entre os obtidos utilizando-se os Conjuntos I e II de limiares, para o hangover de 32 ms, e o Conjunto I de limiares, para o hangover de 20 ms.

5.3.2.1 Funções Distribuição de Probabilidade

São apresentadas, a seguir, algumas figuras referentes às funções distribuição de probabilidade - FDP's dos surtos de voz, representativas de cada um dos quatro grupos utilizados para os testes, tanto para o caso do hangover de 20 ms, quanto para o de 32 ms. Em cada uma das figuras são apresentadas três FDP's, referentes aos três conjuntos de limiares de energia. A curva indicada pelo número 1, representa a FDP para o Conjunto I de Limiares de energia e assim sucessivamente.

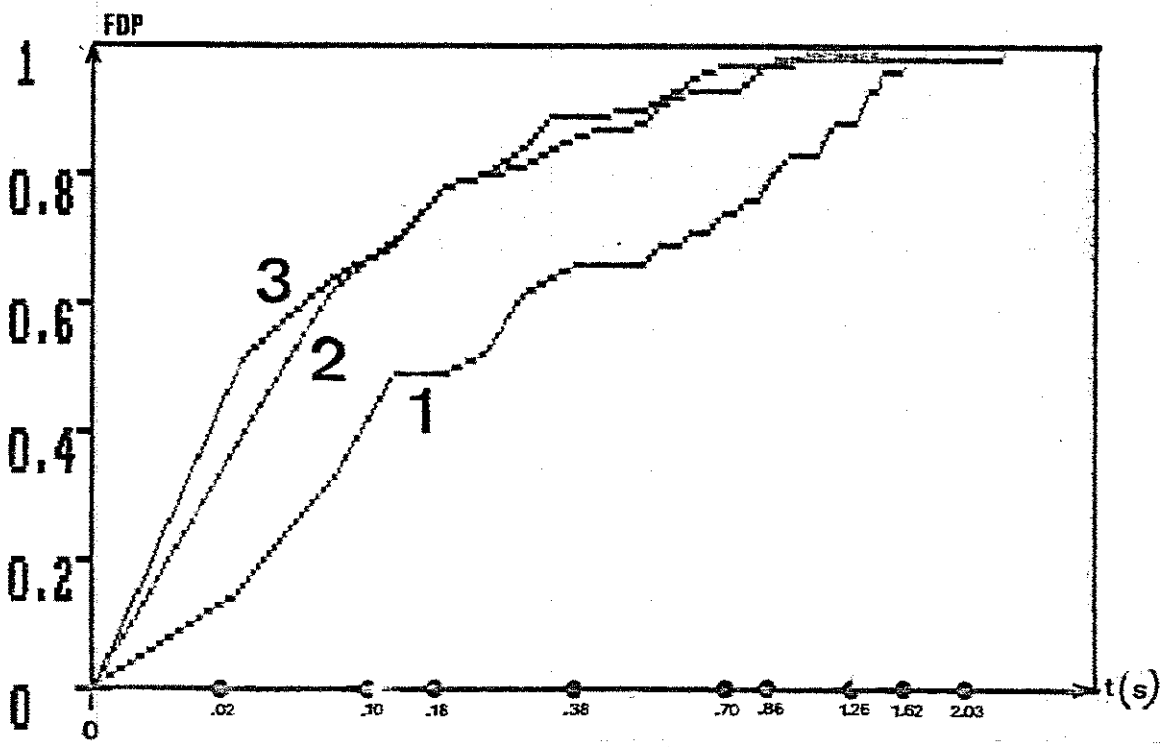


FIGURA 5.8 - FDP de uma conversação telefonica simulada para o grupo "conversação unilateral feminina". Hangover de 20 ms.

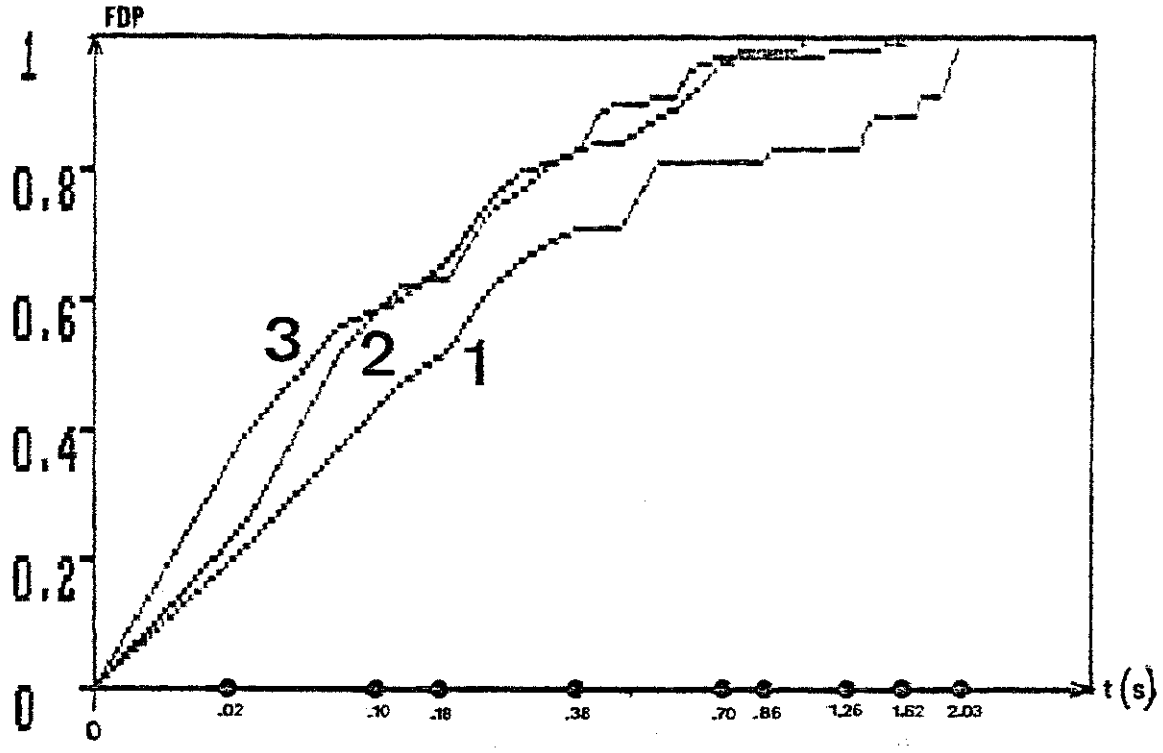


FIGURA 5.9 - FDP de uma conversação telefonica simulada para o grupo "conversação unilateral masculina". Hangover de 20 ms.

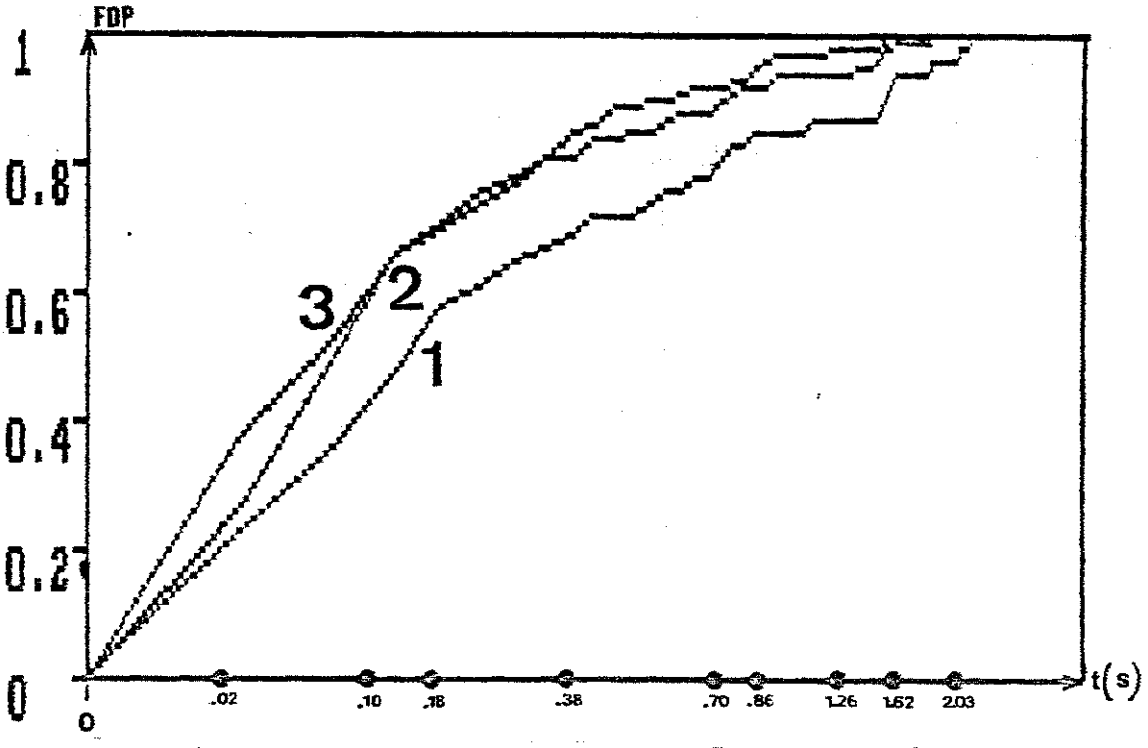


FIGURA 5.10 - FDP de uma conversação telefonica simulada para o grupo "dupla conversação feminina". Hangover de 20 ms.

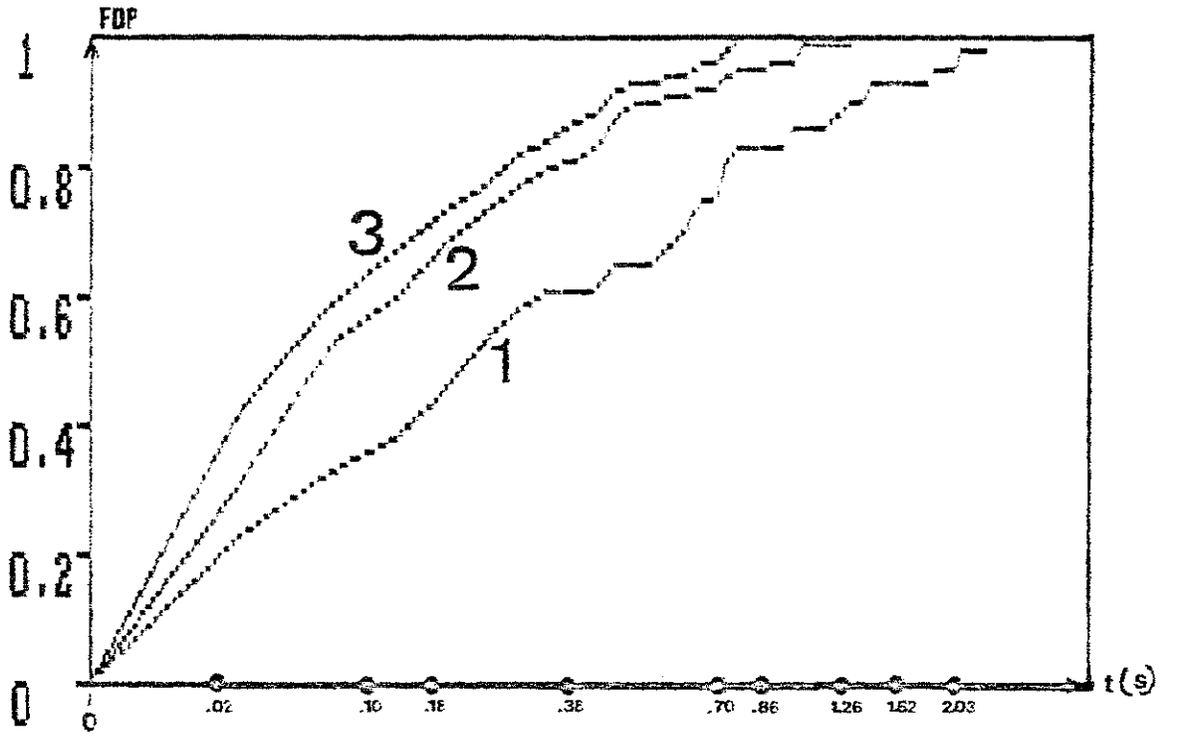


FIGURA 5.11 - FDP de uma conversação telefonica simulada para o grupo "dupla conversação masculina".

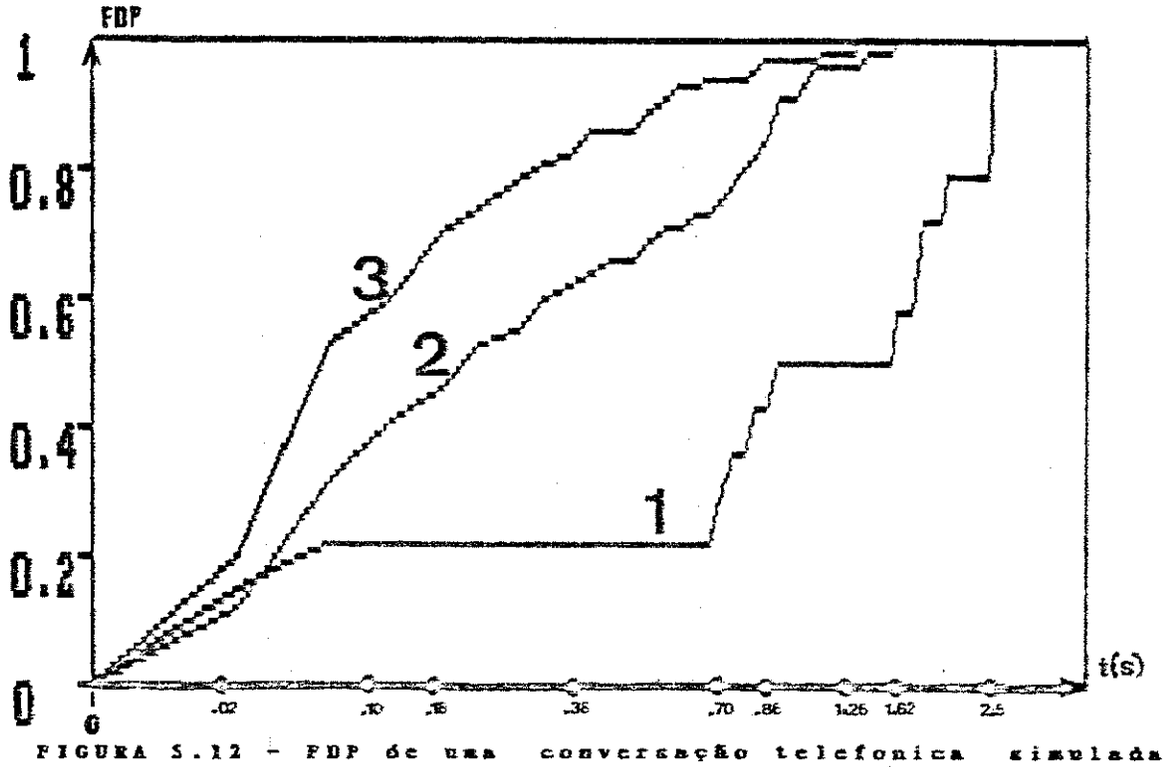


FIGURA 5.12 - FDP de uma conversação telefonica simulada para o grupo "conversação unilateral feminina". Rangover de 32 ms.

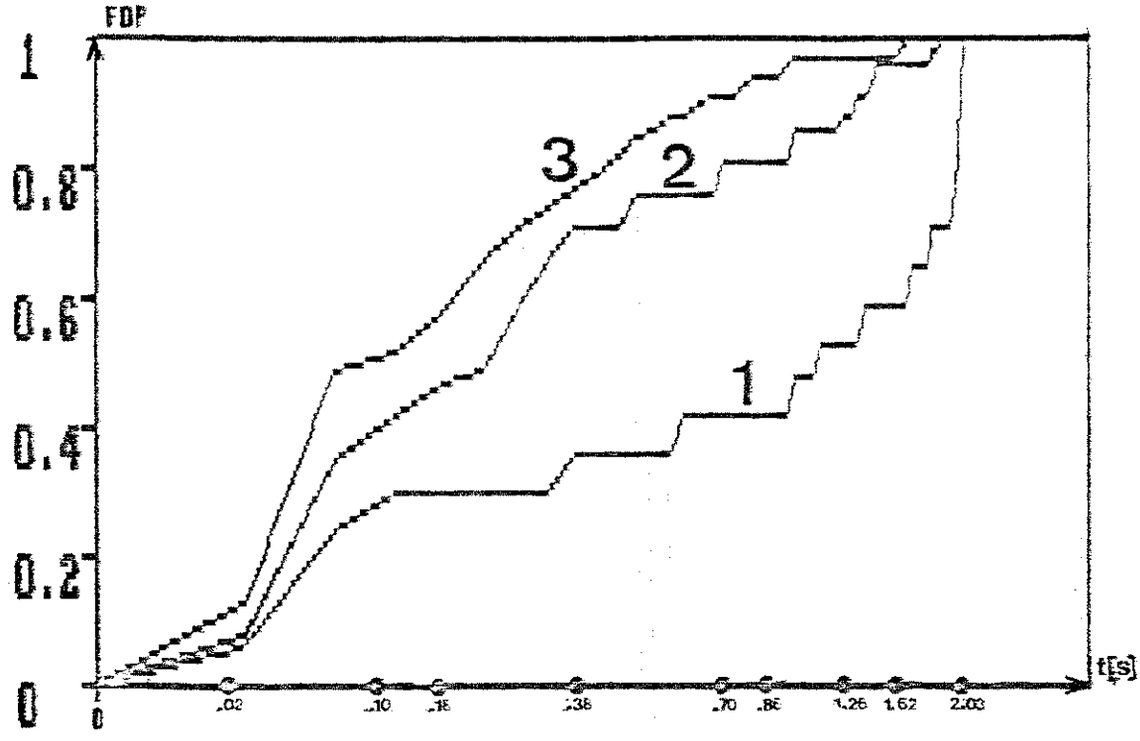


FIGURA 5.13 - FDP de uma conversação telefonica simulada para o grupo "conversação unilateral masculina". Hangover de 32 ms.

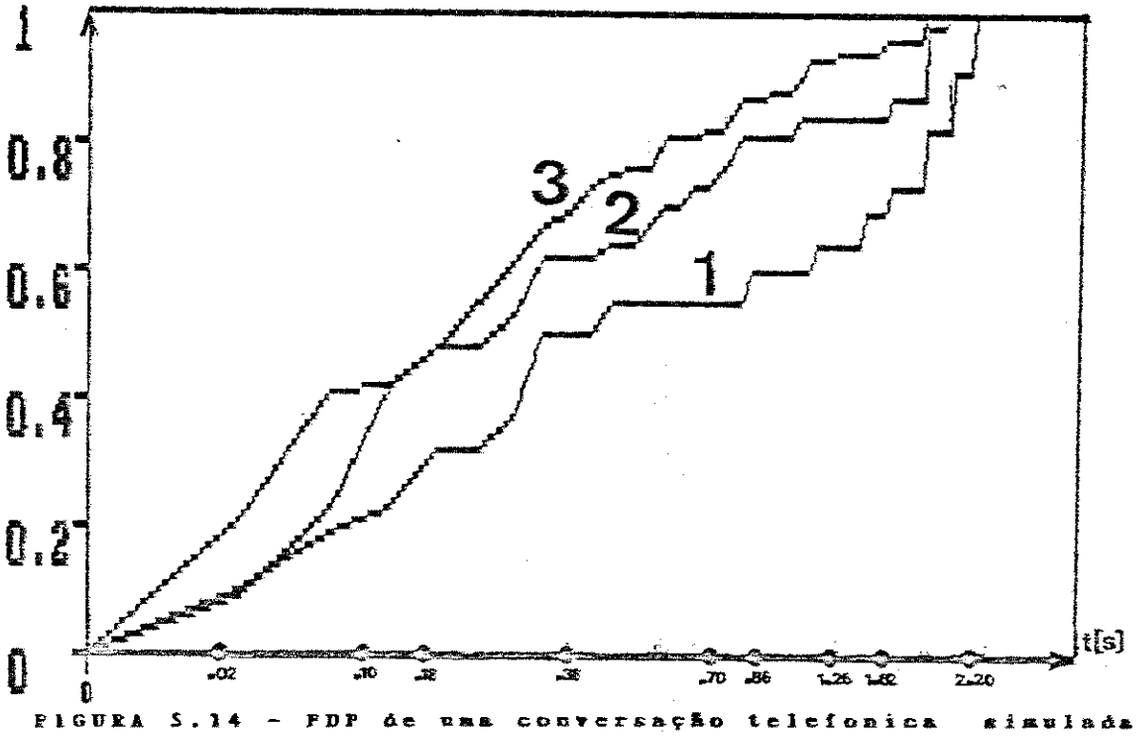


FIGURA 5.14 - FDP de uma conversação telefonica simulada para o grupo "dupla conversação feminina". Hangover de 30 ms.

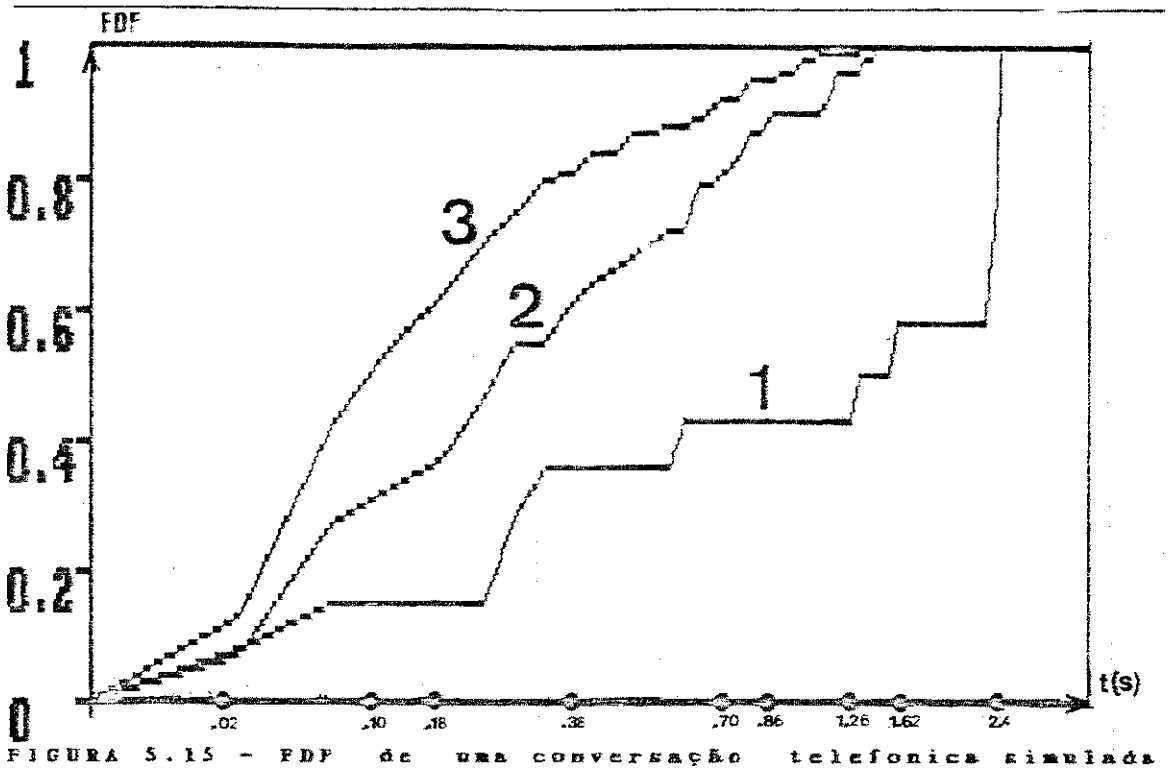


FIGURA 5.15 - FDP de uma conversação telefonica simulada para o grupo "dupla conversação masculina". Hangover de 32 ms.

5.4 Considerações Finais

Pode-se considerar que os resultados obtidos para os eventos de voz de conversações telefônicas simuladas, são satisfatórios, estando de acordo com a teoria formulada.

O algoritmo desenvolvido parece dar bons resultados, na busca de evitar que haja transições errôneas de intervalos de voz para intervalos de pausa.

Na tentativa de aperfeiçoamento do detetor, pode-se prosseguir em seu desenvolvimento, incluindo parâmetros espectrais não utilizados neste trabalho. Em conversações

telefônicas não há a exigência e o rigor em se determinar o tipo de som emitido por um interlocutor, no quadro em análise, mas existe o máximo interesse em garantir se há ou não voz humana neste intervalo. A certeza de que este quadro é constituído de um som gerado pela fala, é fundamental para o sistema onde está inserido o detetor. Assim, justifica-se um esforço no desenvolvimento da robustez do detetor de voz empregado, com a utilização de parâmetros espectrais.

Dois aspectos importantes dificultaram a realização deste trabalho: o primeiro refere-se à falta de trabalho semelhante para a língua portuguesa, que pudesse ser utilizado como referência; o segundo trata da questão da duração média de sentenças ou frases pronunciadas na língua portuguesa. No primeiro caso, seria interessante a realização deste trabalho, utilizando não mais o **hangover** para correção do sinal de voz detetado, mas os **tempos de preenchimento e de eliminação**, para fins comparativos. Outra alternativa seria a realização do mesmo experimento, utilizando-se outro detetor de voz, podendo, neste caso, ser utilizado um detetor desenvolvido a partir de parâmetros temporais e espectrais. No segundo caso, foi visto anteriormente que os surtos de voz não devem ser ocupados apenas por palavras ou sílabas, mas sim por uma frase ou uma sentença. Torna-se importante, portanto, conhecer o tamanho médio de uma frase ou uma sentença emitida na língua portuguesa.

Esta informação pode ser fundamental para definir os parâmetros do detetor de voz. Aqui, esse dado pode auxiliar na determinação ideal dos limiares de energia a serem utilizados pelo detetor, visando estabelecer surtos de voz em sua saída, com a duração de uma frase ou sentença. Seria uma otimização importante [59].

Outro trabalho interessante e que pode ser feito a partir dos resultados obtidos nessa dissertação, refere-se ao modelamento das FDP's dos eventos da voz. As figuras de 5.8 a 5.15 mostram as FDP's obtidas. O trabalho de modelagem é complexo e importante na análise de sistemas de voz.

CAPÍTULO 6
CONCLUSÕES

CAPITULO 6

CONCLUSÕES

6.1 Resumo do Trabalho Realizado

O trabalho apresentado trata do levantamento estatístico de sinais de voz, utilizando parâmetros temporais, tendo como fonte dois processos distintos:

- a locução de uma frase perfeitamente definida;
- a conversação telefônica simulada entre diversos participantes.

No caso das locuções, foi utilizada a frase "AJUSTE DE TEMPO", como referência, para que dezenove locutores a pronunciassem. Essa locução, obtida diretamente de um microfone ou gravada em uma fita cassete, foi digitalizada e processada por um detetor de voz, construído por um algoritmo capaz de detetar os sons surdos e sonoros, além dos intervalos de silêncio, encontrados na frase.

A frase de referência foi elaborada, a partir da utilização de diversos estudos de pesquisadores brasileiros, que formularam tabelas de freqüência dos fonemas mais utilizados na língua portuguesa, falada no Brasil.

O objetivo da elaboração da frase de referência foi o de permitir o conhecimento prévio dos fonemas pronunciados, comparando os resultados encontrados pelo algoritmo, com os resultados corretos, conhecidos à priori.

As conversações foram realizadas a partir de gravações efetuadas entre interlocutores previamente escolhidos, conforme definido no Capítulo 5, digitalizadas e processadas através de um algoritmo desenvolvido específicamente para discriminar os segmentos contendo voz e pausas.

Os resultados obtidos foram satisfatórios, considerando que os êrros encontrados na classificação sonoro-surdo-silêncio, no caso das locuções, resumiram-se aos intervalos de transição entre os sons sonoros e os segmentos de silêncio, sendo inferiores a 7%. No caso das conversações telefônicas simuladas, os resultados almejados como a redução do valor médio dos surtos de voz em relação aos valores típicos, em torno de 1 s a 2 s, encontrados na prática, foram obtidos.

6.2 Apreciação dos Resultados

A seguir são feitos os comentários finais sobre os resultados obtidos no trabalho desenvolvido, tanto para os eventos de locução, quanto para os eventos de conversação.

6.2.1 Considerações sobre os Resultados Obtidos para os Eventos de Locução

O detetor de voz desenvolvido para determinar, a partir da análise segmental do sinal de voz obtido a partir de uma locução, a classificação de cada um dos segmentos da frase utilizada como referência, em sons sonoros, sons surdos ou silêncio, foi construído com parâmetros temporais da voz.

Os parâmetros utilizados foram a energia, o principal deles, a taxa de cruzamento por zero, o número total de picos, a variação de energia, o coeficiente de autocorrelação normalizado e a diferença entre o número de picos do sinal de voz.

Os resultados obtidos comprovam a eficiência do detetor em classificar os intervalos contendo esses eventos.

A janela utilizada para segmentar o sinal de voz, foi a mais reduzida possível, com duração de 4 ms, no intuito de permitir que os parâmetros temporais empregados, refletissem adequadamente as alterações nas propriedades do sinal original.

Das dezenove locuções, as onze obtidas a partir de gravações em fita cassete e as oito geradas diretamente de um microfone, resultaram em arquivos digitalizados, nos quais o intervalo de duração da locução foi, em média, de 1,18 segundos, ou 295 segmentos, lembrando que cada segmento dura 4 ms.

Pela própria frase utilizada, observa-se que os sons sonoros são predominantes. Em média, existiram 200 segmentos com sons sonoros, totalizando 800 ms. No caso dos sons surdos, a quantidade de segmentos é bem menor, com média de 29 quadros, ou seja, 116 ms.

Os intervalos de silêncio, compreendidos no início, no fim e entre as sílabas e palavras que constituem a frase de referência, totalizaram 66 segmentos, com a duração de 264 ms.

Esses valores são as médias das médias dos resultados encontrados para as locuções obtidas através de fita cassete e microfone.

Os segmentos indefinidos podem ser considerados como irrelevantes, pois duraram menos de um segmento, em média.

A taxa de surto de voz, considerando os sons sonoros e surdos, foi de 17,61 c/s e 9,4 c/s para as locuções

obtidas através de fita cassete e microfone, respectivamente.

Os valores da taxa de ocorrência dos sons sonoros, encontrados para tape e microfone foram de 66,44% e 69,18%, respectivamente. Da mesma forma, os valores da taxa de ocorrência dos sons surdos foram de 11,74% e 7,54%. No caso dos intervalos de silêncio, esses valores foram de 21,81% e 22,95%, na mesma ordem anterior dos arquivos de locução.

Foi possível constatar que a maior dificuldade encontrada para se determinar, com exatidão, o tipo de som emitido no segmento em análise, foi na região de transição entre os sons sonoros e os intervalos de silêncio. Nesses casos, os parâmetros utilizados assumem valores muito próximos, o que causa dubiedade. Apesar disso, a margem de êrro estimada foi mínima, sendo da ordem de 7% nos intervalos mais críticos.

Auxiliou muito na diminuição de êrros, o fato de ser utilizado um segmento muito pequeno, da ordem de 4 ms.

As demais transições não foram tão complexas, graças à utilização do coeficiente de autocorrelação normalizado, considerando o fraco correlacionamento existente em um sinal surdo e o alto correlacionamento de um sinal sonoro.

A obtenção dos dados referentes à quantidade de segmentos de sons sonoros, sons surdos e intervalos de silêncio, existentes na locução da frase de referência, possibilitou também o levantamento da Função

Distribuição de Probabilidades de cada um desses eventos, apresentadas no Capítulo 4.

6.2.2 Considerações sobre os Resultados Obtidos para os Eventos de Conversação

A anàlise dos eventos de voz para conversação, foi realizada a partir do processamento desses sinais, por um detetor cujo algoritmo foi desenvolvido a partir da utilização de parâmetros temporais.

Os parâmetros temporais utilizados foram a energia, a taxa de cruzamento por zero e a variação de energia.

O trabalho procurou caracterizar o sinal de voz obtido em uma conversação telefônica simulada, a partir da deteção de sua presença (estado ON) ou ausência (estado OFF) no segmento em análise, utilizando o hangover como parâmetro corretor do sinal na saída do detetor.

Cada segmento teve a duração de 4 ms e um dos objetivos do trabalho foi o de obter surtos de voz, com duração média inferior aos valores encontrados nos detetores convencionais, que estão na faixa de 1 segundo. Para isso, foram utilizados dois valores de hangover: 20 ms e 32 ms.

Tendo em vista que os valores de limiares utilizados para os parâmetros temporais do detetor, são

fundamentais na determinação dos eventos para conversação, foram utilizados três conjuntos de teste com diferentes valores de limiares de energia.

Foi constatado que à medida que os limiares de energia aumentam, o detetor torna-se menos sensível e há uma redução dos eventos de voz, com o conseqüente aumento dos eventos de pausa. Isto significa que a atividade de voz diminui com o aumento dos valores de limiares de energia.

Os eventos de voz também são maiores, quando se utiliza valores maiores para o hangover. Em média, a atividade de voz ocupou 84%, 78% e 67% do tempo total de observação, referente aos Conjuntos I, II e III dos limiares de energia, respectivamente, para o hangover de 20 ms. Para o hangover de 32 ms, os resultados para a atividade de voz foram 86%, 80% e 73%, para os Conjuntos I, II e III, respectivamente. Os Conjuntos I, II e III indicam o aumento respectivo dos limiares de energia utilizados no detetor de voz.

Cálculos simples utilizando o trabalho de Brady [27], permitem obter os percentuais de 43%, 39% e 36% para o evento atividade de voz, considerando os Conjuntos I, II e III respectivamente, com aplicação do tempo de preenchimento igual a 200 ms. Os valores encontrados por Gruber em [18], para a atividade de voz, foram de 79,56% utilizando o hangover, e 72,73% empregando o tempo de preenchimento. Ambos referem-se aos intervalos de silêncio e não de pausa como os utilizados neste trabalho.

No caso do hangover de 20 ms, a duração média do surto de voz situou-se entre 420 ms e 191 ms, para os maiores e os menores valores de limiares de energia, respectivamente. No caso do hangover de 32 ms, a duração média do surto de voz situou-se no intervalo de 814 ms e 327 ms, para os maiores e os menores valores de limiares de energia, respectivamente.

Não foi percebida diferença considerável nos eventos de voz de conversações unilaterais masculinas e femininas, ou mesmo nos eventos de dupla conversação masculinas e femininas.

Tendo em vista que os surtos de voz devem ser suficientes para conter a locução de uma frase ou de uma sentença, e que este dado é desconhecido para a língua portuguesa, pode-se estimar que a duração média ideal desses surtos, deve estar situada entre os valores encontrados ao serem utilizados os Conjuntos I e II de limiares de energia, respectivamente 814 ms e 762 ms, para o hangover com duração de 32 ms, e o Conjunto I de limiares de energia, em torno de 423 ms, para o hangover de 20 ms. Ressalte-se que essas conclusões são estimativas, e que um estudo mais profundo, estabelecendo um valor adequado da duração média de uma frase ou sentença na língua portuguesa, permitirá uma definição mais precisa dos valores ideais de limiares de energia, para o algoritmo elaborado neste trabalho.

6.3 Possíveis Aperfeiçoamentos

Apesar do esforço deste trabalho na busca da elaboração de detetores de voz robustos e de se concluir que os resultados alcançados foram bastante positivos, a experiência obtida permite concluir que pesquisas adicionais, podem levar à obtenção de um aprimoramento dos dispositivos desenvolvidos.

No caso da locução, um aprimoramento deve ser buscado no algoritmo, para uma melhor discriminação entre os sons sonoros e os intervalos de silêncio, nos instantes de transição entre esses estados.

No caso das conversações, outros parâmetros, que não o hangover, devem ser testados para a obtenção do sinal ON-OFF.

Os tempos de preenchimento e de eliminação, definidos no Capítulo 5, podem ser utilizados como parâmetros temporais para *correção* do sinal obtido na saída do detetor de voz, em função da existência de pausas intersilábicas e pequenas interrupções que podem causar fracionamento excessivo no sinal ON-OFF resultante.

Testes realizados com esses parâmetros seriam interessantes, pois seus resultados poderiam ser comparados com os obtidos através deste trabalho.

Os valores de limiar, apesar de criteriosamente escolhidos, podem ser revistos para que novos testes os

otimizem. É importante, nesse caso, conhecer a duração média de uma frase ou sentença, para a língua portuguesa, outro trabalho interessante a ser desenvolvido, este em conjunto com profissionais da área de ligüística.

O modelamento das FDP's obtidas para os eventos de voz utilizados neste trabalho, também pode ser outro tema a ser desenvolvido a partir dos resultados aqui obtidos.

De modo geral, deve-se ressaltar o pioneirismo do trabalho para a língua portuguesa, o que causou dificuldades em muitos aspectos, dado o desconhecimento de algumas informações necessárias para a comparação dos resultados aqui encontrados. No caso da locução, o algoritmo desenvolvido utilizou a inovação de uma janela muito curta para segmentar o sinal de voz, buscando maior precisão na determinação dos parâmetros temporais e, conseqüentemente, na classificação sonoro-surdo-silêncio, em uma frase de referência, o que foi conseguido.

No caso das conversações telefônicas, inúmeros dados foram obtidos, além do desenvolvimento de um detetor robusto, capaz de evitar, adequadamente, transições errôneas entre os intervalos de voz e silêncio.

REFERÊNCIAS BIBLIOGRÁFICAS

REFERENCIAS BIBLIOGRAFICAS

[1] Rabiner, Lawrence R. & Schafer, Ronald W., "Digital Processing of Speech Signals", Prentice-Hall, 1978.

[2] Oppenheim, Alan V., Willsky, Alan S. & Young, Ian T., "Signals and Systems", Prentice-Hall,1983.

[3] Strathmeyer, Carl, "Voice Computing: An Overview of Avaiable Technologies", Computer, Vol 23, nº 8, Agosto de 1990.

[4] Flanagan, James L. & Del Riesgo, Charles J., "Speech Processing: A Perspective on the Science and its Application", AT&T Technical Journal, Vol 69, nº 5, Setembro/Outubro de 1990.

[5] de Marca, José Roberto B., "Representação Digital de Sinais de Voz", Anais do 5º Simpósio Brasileiro de Telecomunicações, p. 5-19, Campinas, SP, 1987.

[6] Hamsher, Donald H., "Communication System Engineering Handbook", McGraw-Hill Book Company, 1967.

[7] Aguiar Neto, Benedito G., "Notas de Aula da Disciplina Processamento Digital de Sinais de Voz", Mestrado em Engª Elétrica / UFPB, 1989.

[8] Kunt, Murad & Hugli, Heinz, "An Overview of Digital Techniques for Processing Speech Signal", Nato Asi Series Ftringer-Verlag, Berlim, 1985.

[9] Atkinson, J., "Telephony - General Principles and Manual Exchange Systems", Capítulo 1, Sound and Speech, Pitman Publishing, 1970.

[10] Schafer, Ronald W. & Rabiner, Lawrence R., "Digital Representations of Speech Signals", Proceedings of the IEEE, Vol 63, nº 4, p. 662-677, Abril de 1975.

[11] O'Malley, Michael H., "Text-to-Speech Conversion Technology", Computer, Vol. 23, nº 8, p. 17-23, Agosto de 1990.

[12] Rabiner, Lawrence R. & Gold, Bernard, "Theory and Application of Digital Signal Processing", Prentice-Hall International, 1975.

[13] Jayant, N. S. & Noll, Peter, "Digital Coding of Waveforms - Principles and Applications to Speech and Video", Prentice-Hall, Inc., 1984.

[14] Atal, Bishnu S., "Automatic Recognition of Speakers from Their Voices", Proceedings of the IEEE, Vol 64, nº 4, p. 460-475, Abril de 1976.

[15] Sambur, M. R. & Rabiner, L. R., "A Speaker-Independent Digit-Recognition System", The Bell System Technical Journal, p. 81-102, January, 1975.

[16] Oppenheim, A. V. & Schafer, R. W., "Digital Signal Processing", Prentice-Hall, 1975.

[17] Pereira, Marco A. T. & Ferreira, F. A. G., "Simulação de um Vocoder Digital", Revista da Sociedade Brasileira de Telecomunicações, Volume 2, nº 1, p. 49-66, dezembro de 1987.

[18] Gruber, John G., "A Comparision of Measured and Calculated Speech Temporal Parameters Relevant to Speech Activity Detection", IEEE Transactions on Communications, Vol Com-30, nº 4, p. 728-738, Abril de 1982.

[19] Norwine, A. C. & Murphy, J., "Characteristic Time Intervals in Telephonic Conversation", The Bell System Technical Journal, Vol 17, p. 281-291, 1938.

[20] Brady, Paul T., "A Technique for Investigating On-Off Pattern of Speech", The Bell System Technical Journal, Vol XLIV, nº 1, p. 1-22, Janeiro de 1965.

[21] Brady, PaulT., "A Model for Generating On-Off Speech Patterns in Two-Way Conversation", The Bell System Technical Journal, Setembro de 1969.

[22] Minoli, D., "Issues in Packet Voice Communication", Proceedings of the IEEE, Vol. 126, nº 8, p. 729-740, Agosto de 1979.

[23] Jaffe, Joseph, Cassota, Louis & Feldstein, Stanley, "Markovian Model of Time Patterns of Speech", Science, Vol 144, p. 884-886, Maio de 1964.

[24] Papoulis, Athanasios, "Probability, Randon Variables and Stochastic Process", MacGraw-Hill Book Company, 2ª edição, 1984.

[25] Richards, D. L., "Telecommunication by Speech", Halsted Press Division, John Wiley & Sons, Inc., 1971.

[26] Rabiner, L. R. & Sambur, M. R., "An Algorithm for Determining the Endpoints of Isolated Utterances", The Bell System Technical Journal, Vol 54, nº 2, p. 297-315, Fevereiro de 1975.

[27] Brady, Paul T., "A Statiscal Analysis of On-Off Patterns in 16 Conversations", The Bell System Technical Journal, p. 73-91, Janeiro de 1968.

[28] Gruber, John & Strawczynski, Leo, "Judging Speech in Dynamically-Managed Voice Systems", Telesis, p. 30-34, 1983.

[29] Peacocke, Richard D. & Graf, Daryl H., "An Introduction to Speech and Speaker Recognition", Computer, p. 26-32 Agosto de 1990.

[30] Wasson, Douglas A. & Donaldson, Robert W., "Speech Amplitude and Zero Crossing for Automated Identification of Human Speakers", IEEE Transactions on Acoustics, Speech, and Signal Processing, p. 390-392, Agosto de 1975.

[31] Vieira, Maurílio Nunes, "Segmentação da Fala em Categorias Fonéticas", Anais do 8º Simpósio Brasileiro de Telecomunicações, p. 30-35, Florianópolis, Setembro de 1989.

[32] Mantegassi, Roberta Abreu, "Discriminador de Dados/Voz", Anais do 8º Simpósio Brasileiro de Telecomunicações, p. 128-132, Florianópolis, Setembro de 1989.

[33] Atal, Bishnu S. & Rabiner, Lawrence R., "A Pattern Recognition Approach to Voiced - Unvoiced - Silence Classification with Applications to Speech Recognition", IEEE Transactions on Acoustic, Speech, and Signal Processing, Vol ASSP-24, nº 3, p. 201-212, Junho de 1976.

[34] Sanbuichi, C. Akira & Aguiar Neto, Benedito G., "Reconhecimento de Palavras Isoladas Independente de Locutor para a Língua Portuguesa do Brasil", Anais do 9o Simpósio Brasileiro de Telecomunicações.

[35] Yatsuzuka, Yohtaro, "Highly Sensitive Detector and High-Speed Voiceband Data Discriminator in DSI-ADPCM Systems", IEEE Transactions on Communications, Vol COM-30, nº 4, p. 739-750, Abril de 1982.

[36] Vieira, Maurílio Nunes, "Módulo Frontal para um Sistema de Reconhecimento Automático de Voz", Dissertação de Mestrado, UNICAMP, Dezembro de 1989.

[37] de Souza, Peter, "A Statistical Approach to the Design of an Adaptive Self-Normalizing Silence Detector", IEEE Transactions on Acoustic, Speech, and Signal Processing, Vol ASSP-31, nº 3, Junho de 1983.

[38] Biderman, Mª Teresa C., "Teoria Lingüística: Lingüistica Quantitativa & Computacional", Livros Técnicos e Científicos, 1978.

[39] Genouvier, E. & Peytard, J., "Lingüistica e Ensino de Portugues", Coimbra:Almedina, 1975.

[40] Cegalla, Domingos Paschoal, "Novíssima Gramática da Língua Portuguesa", Companhia Editora Nacional, 27ª Edição, 1985.

[41] Costa, Amasile C. L., Morais, F. F., Almeida, Rossana R. & Candido, S. Mª, "Anàlise de Fonética Articulatória", Trabalho Executado na Disciplina Lingüistica II, UFPB, 1989.

[42] O Estado de São Paulo, "Pesquisadores Criam Novo Teclado", S.A. O Estado de São Paulo, 1891-Diária, Caderno Ciência e Tecnologia, São Paulo, Terça-Feira, 14 de Maio de 1991.

[43] Rodrigues, Silvio L., "Implementação e Avaliação do Desempenho de um Sistema Automático de Reconhecimento de Locutor pela Análise de Frases Curtas", Anais do Simpósio Brasileiro de Telecomunicações, p. 44-49.

[44] Chassaing, Rulph, Peterson, Wayne A. & Horning, Darrell W., "A TMS3020C25-Based Multirate Filter", IEEE Micro, Vol 10, nº 5, p. 54-61, Outubro de 1990.

[45] (Manual), "TMS 320C25 PC System Board - User Manual", Loughborough Sound Images Ltd., Edição 4.1, Novembro de 1988.

[46] Schildt, Herbert, "C The Complete Reference", Osborne McGraw Hill, 2ª Edição, 1990.

[47] Hergert, Douglas, "O ABC do Turbo C", Makron - McGraw-Hill, 1991.

[48] Bastos, Alberto. M., Lima Filho, Álvaro da Silva, Nery, Fernando, Mannheimer, Paulo H., Pi Farias & Alexandre S., "Linguagem C Programação e Aplicações", Livros Técnicos e Científicos Editora, 3ª Edição, 1989.

[49] Schildt, Herbert, "Linguagem C - Guia Prático e Interativo", McGraw-Hill, 1989.

[50] Arakaki, R., Arakaki, J., Angerami, Paulo M., Aoki, O. L. & Salles, D. de S., "Fundamentos de Programação C - Técnicas e Aplicações", Livros Técnicos e Científicos Editora, 2ª Edição, 1990.

[51] Kobatake, Hidefumi, "Optimization of Voiced/Unvoiced Decisions in Nonstationary Noise Environments", IEEE Transactions on Acoustic, Speech, and Signal Processing, Vol ASSP-35, nº 1, p. 9-18, Janeiro de 1987.

[52] Li, Kung-Pu, Hughes, George W. & Snow, Thomas B., "Segment Classification in Continuous Speech", IEEE Transactions on Audio and Electroacoustics, Vol AU-21, nº 1, Fevereiro de 1973.

[53] Pinto, Neal B. & Childers, Donald G., "Formant Speech Synthesis: Improving Production Quality", IEEE Transactions on Acoustic, Speech, and Signal Processing, Vol 37, nº 12, Dezembro de 1989.

[54] Reddy, D. Raj, "Speech Recognition by Machine: A Review", Proceedings of the IEEE, Vol 64, nº 4, Abril de 1976.

[55] Drago, P. G., Molinari, A. M. & Vagliani, F. C., "Digital Dynamic Speech Detectors", IEEE Transactions on Communications, Vol COM-26, nº 1, Janeiro de 1978.

[56] Langenbucher, Gerhard G., "Efficient Coding and Speech Interpolation: Principles and Performance Characterization", IEEE Transactions on Communications, Vol COM-30, nº 4, Abril de 1982.

[57] Gruber, John, "Performance Considerations for Integrated Voice and Data Networks", Computer Communications, Abril de 1981.

[58] Aguiar Neto, Benedito G., "Melhoramento de Sinais de Voz Degradados por Ruído Acústico Utilizando Balanceamento Espectral", 6º Simpósio Brasileiro de Telecomunicações, Campina Grande, Setembro de 1988.

[59] Olson, Harry F., Belar, Herbert & Rogers, Edward S., "Speech Processing Techniques and Applications", IEEE Transactions on Audio and Eletroacoustics, Vol AU-15, nº 3, Setembro de 1967.

ANEXOS

Listagem do programa que codifica o detetor de voz, utilizado para locução:

```
#include <stdlib.h>
#include <math.h>
#include <stdio.h>
#include <alloc.h>


/*PROGRAMA PRINCIPAL*/

int func_discrim(vdf,ener,zcr,npic,dpico,cross,lx,lj)
float *vdf,*ener,*zcr,*npic,*dpico,*cross;
int lx,lj;
{
   register int j;

   for(j=0;j<lx/lj;j++)
   {
     if(ener[j] <= S1)
     {
       vdf[j] = 0.0;     /*SILENCIO*/
     }
     else if((ener[j] > S1) && (ener[j] < S2))
     {
       fun_class(vdf,zcr,npic,dpico,cross,j);
     }
     else if((ener[j] >= S2) && (ener[j] <= S3))
     {
       if(ener[j] > (ener[j-1]+15.0))
       {
         vdf[j] = 2.0;  /*SOM SONORO*/
       }
       else
       {
         full_eve(vdf,zcr,npic,cross,j);
       }
     }
     else
     {
       vdf[j] = 2.0; /*SOM SONORO*/
     }
   }
}
```

```
/*CALCULO DAS FUNÇÕES AUXILIARES*/

void fun_class(vdf,zcr,npic,dpico,cross,j)
float *vdf,*zcr,*npic,*dpico,*cross;
int j;
{
  float med = 0.0;

  med=(((dpico[j-4]) + (dpico[j-3]) + (dpico[j-2])

      + (dpico[j-1]) + (dpico[j]))/5);
  {
    if((zcr[j] <= Z2) && (npic <= N1))
    {
      vdf[j] = 2.0; /*SOM SONORO*/
    }
    else if(((zcr[j] >= Z1) && (zcr[j] <= Z3)) &&
          ((npic[j] > N1) && (npic[j] <= N3 )))
    {
      if((med > 3.6) || (dpico[j] > zcr[j]))
      {
        vdf[j] = 0.0; /*SILENCIO */
      }
      else
      {
        vdf[j] = 2.0; /*SOM SONORO*/
      }
    }
    else if(((zcr[j] > Z3) && (zcr[j] < Z5)) &&
          ((npic[j] >= N2) && (npic[j] <= N4 )))
    {
      vdf[j] = 0.0; /*SILENCIO*/
    }
    else if(((zcr[j] >= Z5) && (zcr[j] <= Z7)) &&
          (npic >= N2))
    {
      if(cross[j] > 0.3)
      {
        vdf[j] = 0.0; /*SILENCIO*/
      }
      else
      {
        vdf[j] = 1.0; /*SOM SURDO*/
      }
      else if(zcr[j] >= Z7)
      {
        vdf[j] = 1.0; /*SOM SURDO*/
      }
      else
      {
        if(((vdf[j-4]) && (vdf[j-3]) && (vdf[j-2]) &&

          (vdf[j-1])) == 0.0)
```

```
        {
          vdf[j] = 0.0;
        }
        else if(((vdf[j-4]) && (vdf[j-3]) && (vdf[j-2]) &&
              (vdf[j-1])) == 1.0)
        {
          vdf[j] = 1.0; /*SOM SURDO*/
        }
        else if(((vdf[j-4]) && (vdf[j-3]) && (vdf[j-2]) &&
              (vdf[j-1])) == 2.0)
        {
          vdf[j] = 2.0; /*SOM SONORO */
        }
        else
        {
          vdf[j] = 3.0; /*INDEFINIDO */
        }
      }
    }
}
/*-------------------------------------------------------*/

void full_eve(vdf,zcr,npic,cross,j)
float *vdf,*zcr,*npic,*cross;
int j;
{

  if((zcr[j] <= Z3) && (npic[j] <= N4))
  {
    vdf[j] = 2.0; /*SOM SONORO */
  }
  else if(((zcr[j] >= Z4) && (zcr[j] <= Z6) &&
        ((npic[j] > N2) && (npic[j] <= N4)))
  {
    if(cross[j] > 0.6)
    {
      vdf[j] = 2.0; /*SOM SONORO*/
    }
    else
    {
      vdf[j] = 1.0; /*SOM SURDO */
    }
  }
  else if((zcr[j] >= Z6) && (npic[j] >= N2))
  {
    vdf[j] = 1.0; /*SOM SURDO*/
  }
  else
  {
    if(((vdf[j-4]) && (vdf[j-3]) && (vdf[j-2]) &&
      (vdf[j-1])) == 1.0)
    {
      vdf[j] = 1.0; /*SOM SURDO*/
    }
```

```
    else if(((vdf[j-4]) && (vdf[j-3]) && (vdf[j-2]) &&
           (vdf[j-1])) == 2.0)
    {
      vdf[j] = 2.0; /*SOM SONORO*/
    }
    else
    {
      vdf[j] = 3.0; /*INDEFINIDO*/
    }
  }
}
/*-------------------------------------------------------------*/
```

Listagem do programa que codifica o detetor de voz, utilizado para locução:

```
#include <stdlib.h>
#include <math.h>
#include <alloc.h>
#include <stdio.h>


func_onoff(vdf,ener,zcr,lx,lj)
float *vdf,*ener,*zcr;
int lx,lj;
{
  register int j;
  int edf;
  int *pdf;

  for(j=0;j<8;j++)
  {
    if(ener[j] < SA)
    {
      pdf[j] = 0;
      vdf[j] = 0.0;  /*AUSÊNCIA DE VOZ*/
    }
    else if((SA < ener[j]) && (ener[j] < SB))
    {
      if((zcr[j] >= 36.0) || (zcr[j] < 18.0))
      {
        pdf[j] = 1;
        vdf[j] = 1.0;
      }
      if((18.0 <= zcr[j]) && (zcr[j] < 36.0))
      {
        pdf[j] = 0;
        vdf[j] = 0.0;  /*AUSÊNCIA DE VOZ*/
      }
    }
    else
    {
      pdf[j]= 1;
      vdf[j]= 1.0; /*PRESENÇA DE VOZ*/
    }
  }
  for(j=8;j<lx/lj;)
  {
    if(ener[j] < SA)
    {
      pdf[j] = 0;
      vdf[j] = 0.0;  /*AUSÊNCIA DE VOZ*/
      j++;
    }
```

```
else if((SA <= ener[j]) && (ener[j] < SB))
{
  if((ener[j] >= (ener[j-1]*NS)) && (vdf[j-1] == 0))
  {
    pdf[j] = 1;
    edf = ((pdf[j]&&(pdf[j-1]]||pdf[j-2]||pdf[j-3]))||
          (pdf[j-1]&&pdf[j-2]&&pdf[j-3]));
    switch(edf)
    {
      case 0:
      {
        vdf[j] = 0.0; /*AUSÊNCIA DE VOZ*/
        j++;
        break;
      }
      case 1:
      {
        do
        {
          if(ener[j] < SA)
          {
            pdf[j] = 0;
            vdf[j] = 1.0; /*PRESENÇA DE VOZ*/
          }
          else if((SA <= ener[j]) && (ener[j] < SB))
          {
            if((zcr[j] > 36.0) || (zcr[j] < 18.0))
            {
              pdf[j] = 1;
              vdf[j] = 1.0; /*PRESENÇA DE VOZ*/
            }
            else if((18.0 <= zcr[j]) && (zcr[j] <=
                    36.0))
            {
              pdf[j] = 0;
              vdf[j] = 1.0; /*PRESENÇA DE VOZ*/
            }
          }
          else
          {
            pdf[j] = 1;
            vdf[j] = 1.0; /*PRESENÇA DE VOZ*/
          }
          j++;
        }
        while((pdf[j-1]||pdf[j-2]||pdf[j-3]||pdf[j-4]||
               pdf[j-5])==1);
        break;
      }
    }
  }
  else
  {
    if(zcr[j] > 36.0)
```

```
{
  pdf[j] = 1;
  vdf[j] = 1.0; /*PRESENÇA DE VOZ*/
  j++;
}
else if((18.0 <= zcr[j]) && (zcr[j] <= 36.0))
{
  pdf[j] = 0;
  vdf[j] = 0.0; /*AUSÊNCIA DE VOZ*/
  j++;
}
else
{
   pdf[j] = 1;
   edf = ((pdf[j] && (pdf[j-1]] || pdf[j-2] ||
           pdf[j-3])) || (pdf[j-1] && pdf[j-2]
           && pdf[j-3]));
   switch(edf)
   {
     case 0:
     {
       vdf[j] = 0.0; /*AUSÊNCIA DE VOZ*/
       j++;
       break;
     }
     case 1:
     {
       do
       {
         if(ener[j] < SA)
         {
           pdf[j] = 0;
           vdf[j] = 1.0; /*PRESENÇA DE VOZ*/
         }
         else if((SA <= ener[j]) && (ener[j] < SB))
         {
           if((zcr[j] > 36.0) || (zcr[j] < 18.0))
           {
             pdf[j] = 1;
             vdf[j] = 1.0; /*PRESENÇA DE VOZ*/
           }
          else if((18.0<=zcr[j]) && (zcr[j]<=36.0))
           {
             pdf[j] = 0;
             vdf[j] = 1.0; /*PRESENÇA DE VOZ*/
           }
         }
         else
         {
           pdf[j] = 1;
           vdf[j] = 1.0;  /*PRESENÇA DE VOZ*/
         }
         j++;
       }
```

```
            while((pdf[j-1] || pdf[j-2] || pdf[j-3]
                   || pdf[j-4] || pdf[j-5]) == 1);
            break;
          }
        }
      }
    }
}
else if((SB <= ener[j]) && (ener[j] < SC))
{
  pdf[j] = 1;
  edf = ((pdf[j]&&(pdf[j-1]||pdf[j-2]||pdf[j-3]))||
        (pdf[j-1]&&pdf[j-2]&&pdf[j-3]));
  switch(edf)
  {
    case 0:
    {
      vdf[j] = 0.0; /*AUSÊNCIA DE VOZ*/
      j++;
      break;
    }
    case 1:
    {
      do
      {
        if(ener[j] < SA)
        {
          pdf[j] = 0;
          vdf[j] = 1.0; /*PRESENÇA DE VOZ*/
        }
        else if((SA <= ener[j]) && (ener[j] < SB))
        {
          if((zcr[j] > 36.0) || (zcr[j] < 18.0))
          {
            pdf[j] = 1;
            vdf[j] = 1.0; /*PRESENÇA DE VOZ*/
          }
          else if((18.0 <= zcr[j]) && (zcr[j] <= 36.0)
          {
            pdf[j] = 0;
            vdf[j] = 1.0; /*PRESENÇA DE VOZ*/
          }
        }
        else
        {
          pdf[j] = 1;
          vdf[j] = 1.0;  /*PRESENÇA DE VOZ*/
        }
        j++;
      }
      while((pdf[j-1]||pdf[j-2]||pdf[j-3]||pdf[j-4]||
             pdf[j-5])==1);
      break;
    }
```

```
      }
    }
    else
    {
      do
      {
        if(ener[j] < SA)
        {
          pdf[j] = 0;
          vdf[j] = 1.0;  /*PRESENÇA DE VOZ*/
        }
        else if((SA <= ener[j]) && (ener[j] < SB))
        {
          if((zcr[j] > 36.0) || (zcr[j] < 18.0))
          {
            pdf[j] = 1;
            vdf[j] = 1.0; /*PRESENÇA DE VOZ*/
          }
          else if((18.0 <= zcr[j]) && (zcr[j] <= 36.0))
          {
            pdf[j] = 0;
            vdf[j] = 1.0; /*PRESENÇA DE VOZ*/
          }
        }
        else
        {
          pdf[j] = 1;
          vdf[j] = 1.0; /*PRESENÇA DE VOZ*/
        }
        j++;
      }
      while((pdf[j-1]||pdf[j-2]||pdf[j-3]||pdf[j-4]||
             pdf[j-5])==1); }
  }
}
```