UNIVERSIDADE DO PORTO

FACULDADE DE ENGENHARIA

DEPARTAMENTO DE ENGENHARIA ELECTROTÉCNICA
E DE COMPUTADORES

Dissertação de Mestrado

MODELIZAÇÃO PARAMÉTRICA DE SINAIS PARA APLICAÇÃO EM SISTEMAS DE CONVERSÃO TEXTO-FALA

ENGENHARIA ELECTROTÉCNICA E DE COMPUTADORES
ÁREA CIENTÍFICA DE INFORMÁTICA INDUSTRIAL

João Paulo Ramos Teixeira

Porto
Outubro 1995

MODELIZAÇÃO PARAMÉTRICA
DE SINAIS PARA APLICAÇÃO EM
SISTEMAS DE CONVERSÃO
TEXTO-FALA

Tese realizada sob orientação do Professor Doutor

Diamantino Rui da Silva Freitas

Professor Auxiliar do Departamento de Engenharia Electrotécnica e de Computadores da

Faculdade de Engenharia da Universidade do Porto

Aos meus pais

e esposa

RESUMO

Foi desenvolvido neste trabalho um sistema de extracção automática de parâmetros de sinais de fala recorrendo a ferramentas de análise cepstral, de predição linear quer pela matriz autocorrelação quer pela matriz covariância, e ao método de análise síncrona com o período fundamental.
Realiza-se uma segmentação e classificação dos sinal em vocalizados, não vocalizados ou silêncio. Aos segmentos com conteúdo de fala atribuem-se modelos baseados em formantes.
Os parâmetros definidos pelo modelo para a fala vocalizada são 4 formantes e respectivas larguras de banda, frequência fundamental e amplitude. Para os sons não vocalizados considerou-se um modelo com um pólo, um zero e excitação com sinal de ruído aleatório.
O método de análise cepstral segmenta o sinal com comprimentos fixos e analisa individualmente cada segmento. A análise de cada segmento consiste na separação das características do trato vocal e da fonte excitadora, recorrendo a uma função de "lifteragem" nas quefrências sendo determinada a frequência fundamental da fonte excitadora ("pitch") e alisado o espectro relativo ao trato vocal. A partir deste espectro alisado ou envelope espectral é aplicado um algoritmo de determinação dos picos para extrair as frequências formantes das ressonâncias do trato vocal obedecendo a constrições respeitantes às regiões de frequências de cada formante e às amplitudes relativas dos respectivos picos. São também determinadas as correspondentes larguras de banda a 3 dB a partir do envelope espectral.
Os métodos de predição linear analisam também individualmente cada segmento de comprimento fixo do sinal de fala, obedecendo a um modelo só com pólos, determinando os coeficientes de predição linear por multiplicação matricial. A partir destes coeficientes são determinados os pólos. Cada par de pólos complexos conjugados é considerado um possível formante, sendo posteriormente seleccionados justamente 4 formantes por um processo de eliminação das frequências formantes que não têm a correspondência de um pico na função de transferência do sistema.
O método de análise síncrona com o período fundamental determina o sincronismo com o impulso glotal segmentando o sinal em troços de duração de um período, sendo estes posteriormente analisados pelo método de predição linear (matriz covariância).
Posteriormente a sequência de parâmetros é sujeita a um alisamento não linear para corrigir eventuais pontos fora de uma linha definida pelos valores dos parâmetros anteriores e posteriores ("outliers").
Todos estes métodos determinam com razoável fidelidade as frequências formantes dos sinais de fala, contudo, as larguras de banda são mais correctamente determinadas pelo método de predição linear pela matriz covariância.
É ainda apresentado o desenvolvimento de um conversor texto-fala para o português baseado num sintetizador de formantes com o mesmo modelo usado na análise para os sinais vocalizados. Os principais resultados obtidos foram a realização acústica de uma lista de 37 fonemas fundamentais, regras de conversão grafema-som na forma tabular, um grupo de regras de concatenação para as estruturas acústica e temporal inerente aos sons, regras prosódicas elementares e, pronuncia de acrónimos e numerais.
Foram ainda desenvolvidas várias ferramentas complementares à análise dos sinais de fala como sejam um espectrógrafo e um outro sintetizador de formantes, exclusivamente computacional e para testes, baseado no modelo com os mesmos parâmetros.
Os métodos desenvolvidos foram testados com sinais de fala adequadamente seleccionada e recolhida em sala insonorizada e, registados magneticamente com aparelhagem adequada.
Os resultados atingidos satisfazem os objectivos inicialmente propostos para este trabalho.

Palavras-chave: processamento digital de sinais, análise de fala, síntese de fala, conversores texto-fala, modelização de sinais de fala, análise cepstral, predição linear.

ABSTRACT

It has developed in this work a system for automatic feature extraction of speech signal parameters using tools like cepstral analysis, linear prediction by autocorrelation and covariace matrix and the pitch synchronous analysis method.
A segmentation and classification of this signal in voiced, unvoiced or silence as done. The segments with speech content are considered as a formant based model.
The parameters defined by the model for voiced speech consist in a 4 formants and corresponding bandwidths, fundamental frequency or pitch and magnitude. For the unvoiced sounds was considered a model with one pole, one zero and white noise excitation source signal.
In the cepstral analysis method the signal are segmented in units of fixed duration to be individually analysed. The analysis of each segment consist in the separation of the vocal tract characteristics from the excitation source characteristics applying a lifter function to the quefrency domain signal resulting in the smoothed spectrum of the vocal tract, meanwhile the pitch are resolved. To the smoothed spectrum or spectral envelop are applied an algorithm to extract the formant frequencies corresponding to the vocal tract resonance's obeying at a some constraints related to the frequency regions of each formant and magnitude of the picks. Also using the spectral envelop are resolved the corresponding bandwidths at 3 dB.
The LPC methods also individually analyze each segment of speech signal with fixed duration obeying a all poles model, determining the LPC coefficients by matricial multiplication. The poles are resolved using this coefficients. Each pair of conjugated complex poles is considered as a possible formant, being latter selected just 4 formants, by a process of eliminating formant frequencies which do not have the corresponding pick in the system transfer function.
The pitch synchronous analysis method adjust the synchronism with glotal pulse doing the segmentation of signal in frames of one period duration, to future analyze by the LPC method (covariance).
After the extraction, the parameters sequence are passed through a non linear smoothing in order to correct eventual outlier points.
All of this methods resolve with reasonable fidelity the formant frequencies of speech signals, although, the bandwidths are most correctly resolved by the LPC covariance method.
Also is presented the text-to-speech converter system for Portuguese development in a formant synthesizer using the same model that as used in voiced speech analysis. The principal results was an acoustic realization of a set of 37 fundamental phonemes, a grapheme-sound conversion rules in a tabular format, a set of concatenation rules for the inherent time and acoustic structures of sounds, elementary prosodic rules and, numerals and acronyms pronunciation.
Also have been developed complementary tools to speech analysis like a spectrograph and another formant synthesizer, exclusively computational for testing, based on same model.
The developed methods was tested with a selected speech signals recorded in an appropriated room and magnetically registered with adequate devices.
The achieved results corresponds to the initially proposed objectives of this work.

Keywords: digital signal processing, speech analysis, speech synthesis, text-to-speech systems, speech signals modeling, cepstral analysis, linear prediction.

AGRADECIMENTOS

Dirijo o meu primeiro agradecimento ao orientador deste trabalho, Prof. Diamantino Rui da Silva Freitas, pela forma cativante e amigável com que sempre se dispôs a discutir assuntos relacionados com esta dissertação, incentivando ao rigor científico e propondo novos desafios e ideias que resultaram em contribuições valiosas para este trabalho. Agradeço-lhe ainda o "empréstimo" da sua voz no processo de recolha dos sinais de fala.

Aos Prof. Espain Oliveira, Rui Araújo e à Anabela, ao Rui Pessegueiro e ao Manuel Cardoso agradeço as constantes trocas de impressões de matérias relacionadas com o tema da dissertação.

Aos Prof. Géza Németh e Doutor Gábor Olaszy agradeço o apoio e boas condições criadas no decurso do trabalho em Budapest, tendo resultado num forte estímulo para a prossecução deste.

Aos colegas e amigos Jorge Reis, Avelino Marques e Henrique Cabral entre outros, o meu muito obrigado pelo ambiente de boa disposição e entre ajuda que souberam criar.

Agradeço aos meus pais todo o apoio que sempre me deram até ao atingir desta etapa.

Por último mas não menos importante quero deixar uma palavra de carinho muito especial para a minha esposa pelo apoio, confiança e amor incondicionais que me dedicou fazendo com que este trabalho fosse possível.

O meu obrigado à RDP Porto pelas instalações e condições técnicas cedidas para a recolha de sinais de fala.

O desenvolvimento desta dissertação foi subsidiada com uma bolsa de Mestrado pelo programa PRAXIS XXI, facto pelo qual manifesto a minha gratidão.

Índice

Lista de Figuras

Capítulo 2

Capítulo 3

Capítulo 4

Capítulo 5

Capítulo 6

Capítulo 7

Capítulo 8

Lista de Tabelas

Capítulo 2

Capítulo 5

Capítulo 7

Lista de Abreviaturas e Símbolos

ABU - Acoustic Building Unit

ADM - Adaptative Delta Modulation

ADPCM - Adaptative Differential Pulse Code Modulation

APCM - Adaptive Pulse Code Modulation

CZT - Chirp Z Transform

DM - Delta Modulation

DPCM - Differential Pulse Code Modulation

e - número de elementos abaixo de LSES

F0 - Frequência Fundamental

FFT - Fast Fourier Transform

Fi - formante de ordem i

Fimax - limite superior da região de frequências para o formante Fi

Fimin - limite inferior da região de frequências para o formante Fi

Fn - Frequência de Nyquist

Fs - Frequência de amostragem

G(z) - Função de transferência do impulso glotal

IDFT - Inverse Discrete Fourier Transform

LIZ - Limite Inferior da taxa de passagem por Zero

LSES - Limite Superior da Energia do Sinal

PCM - Pulse Code Modulation

R(z) - Função de transferência para o efeito de radiação nos lábios

RDP - Rádio Difusão Portuguesa

TPZ ou tpz - Taxa de Passagem por Zero

V(z) - Função de transferência do trato vocal

z - Número de elementos acima de LIZ

CAPÍTULO 1

INTRODUÇÃO

1. INTRODUÇÃO

1.1 Objectivos e Enquadramento Deste Trabalho

O objectivo deste trabalho é o estudo e desenvolvimento de sistemas que de uma forma automática realizem a extracção de um grupo de parâmetros, definido por um modelo, a partir de um sinal de fala com a menor intervenção possível do utilizador.

Do funcionamento deste sistema pretende-se que o mais facilmente possível se possa pronunciar um som, palavra ou frase para um microfone e que este sinal seja armazenado electronicamente no computador para que seja posteriormente analisado pelo sistema desenvolvido, extraindo uma sequência de um grupo de parâmetros estabelecido por um modelo para os diferentes tipos de sinais de fala de forma a permitir que um sintetizador de formantes, baseado no mesmo modelo definido para os sinais de fala, alimentado por essa sequência de parâmetros produza novamente o sinal de fala original com o mínimo de distorção possível. Um sistema desta natureza é habitualmente designado por "VOCODER" e permite armazenar e transmitir sinais de fala ocupando, quer no armazenamento quer na transmissão, um número de bits muito menor que o sinal de fala original.

Um conversor texto-fala desenvolvido para o português durante este trabalho, faz uso de um conjunto de regras de concatenação de fonemas onde são implementados os sons da parte estável do fonema e as próprias transições entre fonemas fazendo variar de uma forma precisa os parâmetros dos modelos para os sinais de fala. O estabelecimento destas regras baseia-se num método denominado análise por síntese, em que se vai iterativamente chegando aos valores dos parâmetros por um processo de alteração de um valor destes, verificar o resultado e prosseguir as alterações no sentido de aproximar os sons sintetizados dos sons pretendidos. Este é um método de tentativa e erro um pouco penoso, moroso e com resultados de qualidade não óptima. Torna-se então clara a necessidade de um sistema que realize a análise de sinais de fala indicando assim quais as variações que devem seguir os parâmetros para produzir o som pretendido.

As ferramentas de análise aqui desenvolvidas têm aplicações em sistemas de análise de fala, em conversores texto-fala e sistemas de reconhecimento quer da fala quer do falante.

Estas matérias têm hoje um interesse crescente para aplicação nas áreas de: reabilitação, onde por exemplo um conversor texto-fala permite a um invisual trabalhar com recurso informáticos como o computador, que pronunciará o que se passa no écran, a um incapacitado, temporária ou permanentemente, da sua voz, recorrer a um programa adequado para fazer ouvir as suas mensagens. Por outro lado, os programas de análise permitem a um terapeuta da fala diagnosticar eventuais danos no aparelho fonador do paciente; na área das comunicações estes sistemas têm o seu interesse na medida em que realizam uma compressão da informação dos sinais de fala permitindo reduzir a taxa de transmissão necessária para estes sinais (por ex. sistema GSM); na área da indústria existem já hoje imensos brinquedos "falantes" e por outro lado vão sendo cada vez mais comuns as mensagens faladas emitidas por máquinas; no ensino, estes sistemas podem ajudar as crianças a aprender como se pronunciam as palavras, ou o contrário, a partir da locução de uma palavras aprender como esta se escreve, estes sistemas de ensino têm também utilidade para o ensino da língua a estrangeiros; finalmente, a "interface" com

as máquinas caminha no sentido de que estas possam ser comandadas através da fala humana como referiu Bill Gates no seu discurso de lançamento do Windows 95.

As técnicas de análise abordadas nesta dissertação foram a análise cepstral, predição linear pelas matrizes autocorrelação e covariância e análise temporal síncrona com o período fundamental, tendo-se conseguido atingir bons resultados com todos os métodos, que são apresentados e comparados.

Os modelos usados para os sinais de fala são baseados em formantes tendo sido considerado para a fala vocalizada o conjunto de parâmetros constituído por 4 formantes, as 4 respectivas larguras de banda, a frequência fundamental e a amplitude, para a fala não vocalizada foi considerado um modelo com um pólo e um zero. Contudo, conclui-se que os sistemas de análise da fala vocalizada podem modelizar razoavelmente também os sons não vocalizados.

Um conjunto adequado de sinais base seleccionados e, pronunciados por dois locutores, para estudo neste trabalho foi recolhido em sala insonorizada para fita magnética tendo posteriormente sido convertidos para digital a uma frequência de amostragem de 22.05 Khz e armazenados em memória de computador recorrendo a uma placa de som existente no mercado. Estes sinais são depois decimados resultando numa frequência de 11.025 Khz. Procurou-se recolher estes sinais com o mínimo de distorção possível tomando medidas de precaução relativamente ao ruído ambiente e à aparelhagem usada.

1.2 Aspectos de Organização da Dissertação

Procurou-se dar uma ordem aos temas desenvolvidos neste trabalho que permita ao leitor um acompanhamento e percepção das questões envolvidas em cada assunto tratado.

O capítulo 2 é iniciado com uma descrição do aparelho fonador humano e do processo de produção da fala. É ainda realizada uma caracterização do ponto de vista acústico e por vezes articulatório dos diversos sons produzidos pela língua portuguesa de forma a permitir uma compreensão dos modelos para diferentes tipos de sons. Este capítulo termina com a apresentação de uma lista de fonemas básicos usados na língua portuguesa.

No terceiro capítulo relativo aos processos envolvidos na recolha e aquisição do sinal de fala, discutem-se os requisitos mínimos necessários a este processo e descrevem-se as condições em que decorreu a recolha do sinal neste trabalho.

No quarto capítulo são desenvolvidos os modelos de produção da fala e a parametrização destes sinais. Discutem-se as funções de transferencia que simulam o trato vocal V(z), o efeito de radiação nos lábios R(z) e a fonte de excitação glotal G(z). São estudados os conjuntos de parâmetros usados nos modelos dos sinais de fala e apresentadas as técnicas para extracção destes baseadas no modelo de formantes. Estas técnicas são análise por síntese, análise síncrona com o período fundamental, análise cepstral e análise por predição linear também conhecida em outros ambientes por métodos de identificação. Este capítulo termina com uma breve referências às técnicas usadas para a codificação dos sinais de voz.

O quinto capítulo faz uma descrição genérica dos sistemas de conversão texto-fala referindo-se às partes de processamento linguístico e de processamento acústico. Este termina com uma descrição do sistema de conversão texto-fala MULTIVOX e do desenvolvimento da língua portuguesa neste conversor, especificando as regras de

conversão grafema-código de fonema, grupo básico de fonemas usado, regras de concatenação de fonemas e as regras de prosódia.

O sexto capítulo descreve a aplicação e desenvolvimento de uma conjunto de ferramentas criadas e outras apenas usadas, devido à sua existência prévia no programa Matlab, para aplicações nos sistemas finais de extracção automática dos parâmetros dos sinais de fala. As ferramentas aqui abordadas foram a amplitude média deslizante, a energia média deslizante, a taxa de passagem por zero tendo-se também discutido métodos de classificação/segmentação dos sinais de fala em que foi proposto um método original de classificar/segmentar os sinais baseado nas ferramentas anteriores e num domínio de decisão. São também apresentadas as ferramentas, CZT (Chirp Z Transform) que permite neste âmbito a apreciação expandida de uma zona de frequências do espectro , o cepstro nas variantes de cepstro de potências e cepstro complexo, uma função criada para representação dos espectrogramas apresentados neste trabalho, um sintetizador usado para criar sinais com variações dos parâmetros dos modelos conhecidas com a finalidade de testar se os parâmetros extraídos pelos métodos de análise seguem essas variações, e, finalmente são discutidos os processos de elaboração das matrizes autocorrelação e covariância para o método de predição linear.

O capítulo 7 apresenta os sistemas criados para a extracção automática dos parâmetros dos sinais de fala fazendo usos de todo o conhecimento desenvolvido nos capítulos anteriores. Inicia com a descrição de alguns métodos para determinação automática da frequência fundamental. O método do cepstro, a determinação deste parâmetro no erro residual de predição linear e um método com processamento exclusivamente no domínio temporal com grandes facilidades de implementação em "hardware". Este capítulo continua com a descrição de uma estrutura de análise para os sinais vocalizados criada, que usa segmentos de comprimento fixo com sobreposição de 50%. Nesta estrutura podem-se seleccionar os métodos de análise pretendidos (análise cepstral, predição linear pela matriz autocorrelação ou covariância). Estes métodos são aqui desenvolvidos iniciando com o método de análise cepstral em que se determina o espectro alisado por uma lifteragem nas quefrências, é determinada a frequência fundamental e depois segue-se um processo de detecção dos picos para determinação dos formantes a partir do espectro alisado. Seguem-se os dois métodos de predição linear sendo apresentados os algoritmos de determinação das frequências formantes a partir dos coeficientes de predição linear. Ainda para a fala vocalizada é apresentado um método de análise síncrona com o período fundamental. Seguindo-se um processo de alisamento não linear para corrigir alguns parâmetros fora de uma sequência ao longo do tempo definida pelos valores dos mesmos parâmetros em instantes anteriores e posteriores ("outliers"). Para a fala não vocalizada é apresentado um método de determinação das frequências de um pólo e um zero segundo o modelo descrito no quarto capítulo.

No capítulo 8, destinado à avaliação dos resultados obtidos, é descrito o estado de desenvolvimento do conversor texto-fala MULTIVOX para o português e são realizados testes e apresentados os resultados de extracção automática dos parâmetros pelos métodos desenvolvidos, para alguns sinais. Os sinais testados são a vogal [i] pronunciada continuamente, a palavra "ama" e um sinal sintetizado. São ainda comparados os resultados dos vários métodos.

Esta dissertação é encerrada com o nono capítulo para apresentação de conclusões e desenvolvimentos futuros.

CAPÍTULO 2

CARACTERIZAÇÃO FONÉTICA DA LÍNGUA PORTUGUESA

2. CARACTERIZAÇÃO FONÉTICA DA LÍNGUA PORTUGUESA

2.1 Introdução

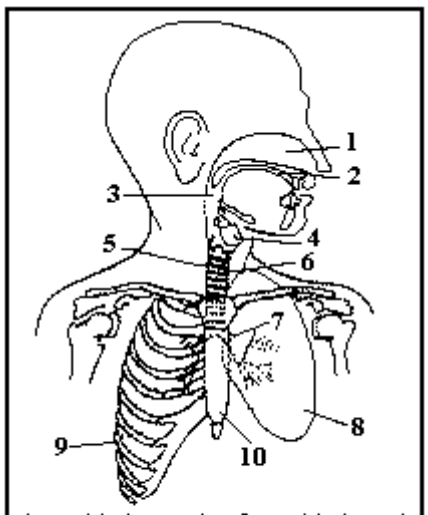
O objectivo deste capitulo centra-se numa introdução ao sistema fonador humano tendo em vista a compreensão de alguns termos importantes usados durante esta dissertação e um conhecimento do processo de produção da fala para se poder estabelecer uma analogia com os modelos de produção de fala desenvolvidos no capitulo 4. Por outro lado é feita uma caracterização fonética do ponto de vista acústico da língua portuguesa de forma a permitir uma consciencialização de alguns detalhes requeridos nos modelos de produção de fala. O desenvolvimento deste capitulo é ainda importante do ponto de vista de implementação de um conversor texto fala para o português, nomeadamente nas regras de conversão grafema-fonema deste conversor.

2.2 Produção da Fala

O processo de produção da fala está intrinsecamente ligado aos órgãos e sistema de respiração.

No processo de expiração são permitidas maiores variações de pressão do que no processo de inspiração, tornando-se audível, pela produção de ondas sonoras que, modeladas pela laringe e as cavidades superiores orais e nasais, darão as características da voz.

No acto da expiração o fluxo de ar segue um trajecto inverso do trajecto seguido pelo ar no processo de inspiração, saindo dos alvéolos, passando pelos brônquios, traqueia, laringe, faringe e finalmente cavidade nasal e/ou oral.



1. cavidade nasal; 2. cavidade oral; 3. faringe; 4. laringe; 5. esófago; 6. traqueia; 7. brônquios; 8. pulmões; 9. caixa torácica; 10. esterno.

Figura 2.1 - Aparelho fonador humano.

Depois do fluxo de ar sair dos pulmões (alvéolos e brônquios) e da passagem pela traqueia, passa pela laringe que desempenha um papel fundamental no processo de produção da fala. É no interior da laringe que se situam as chamadas "cordas vocais", compostas por ligamentos e músculos. Na parte superior situam-se as cordas ventriculares - que são glandulares -, também chamadas "falsas cordas vocais" por não terem função na fonação. Na parte inferior situam-se as cordas vocais, que são musculares. No espaço entre estes dois tipos de cordas situa-se o ventrículo. O espaço entre as duas cordas vocais, direita e esquerda, é a glote.

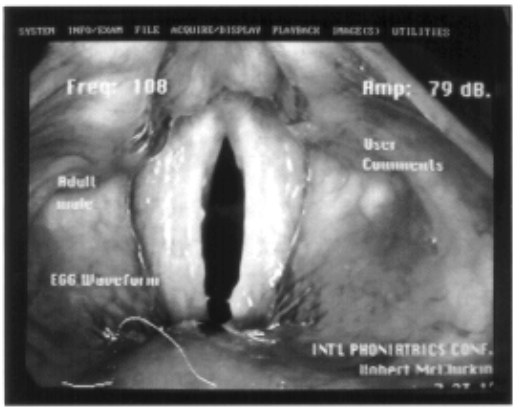


Figura 2.2 - Glote.

Durante a fonação, a função das cordas é actuar como um gerador de som, abrindo e fechando rapidamente a passagem ao fluxo de ar vindo dos pulmões. A junção das cordas vocais cria pressão subglotal que vai aumentando até ser suficiente para afastar as cordas vocais uma da outra. Quando as cordas se afastam, o ar sai, a pressão decresce, e as cordas voltam a aproximar-se. A cada ciclo destes dá-se o nome de impulso glotal ou período fundamental. Este ciclo repete-se durante a fonação.

Prosseguindo com o trajecto percorrido pelo fluxo de ar no processo de expiração segue-se a faringe que é um "tubo" que se situa entre a laringe e as cavidades oral e nasais. Tem três partes distintas: a faringe laríngea, a faringe oral e a faringe nasal. Entre a faringe oral e a nasal existe o véu palatino que separa estas duas partes. É um músculo ligado por um dos lados à parte posterior do palato (palato duro) e às paredes laterais da parte posterior da cavidade oral, ficando o outro lado solto. À extremidade solta chama-se úvula. Esta pode-se deslocar separando as cavidades nasais da cavidade oral.

As cavidades nasais são duas e estão separadas pelo septrum nasal. A base ou "chão" das cavidades nasais é constituído pelo palato duro.

A cavidade oral situa-se entre a faringe, posterior, e os dentes, anterior e lateral. A parte inferior é constituída pelo "chão" da boca e pela língua e na parte superior pelo palato duro. A zona interna da implantação dos dentes é a zona dos alvéolos ou alveolar. Os dentes do maxilar superior estão fixos enquanto os do maxilar inferior acompanham os movimentos deste. A parte oral do palato estende-se entre os alvéolos dos dentes e a úvula. Depois dos alvéolos encontra-se uma zona de "rugas" - o palato duro - com um silhão palatino, depois o véu palatino - ou palato mole - e a úvula. A língua é um órgão muscular coberto por uma membrana mucosa. Os bordos anteriores e laterais estão soltos; a parte inferior está ligada ao "chão" da boca e ao osso hióide.

Na parte posterior-inferior está ligada a epiglote. A ponta anterior chama-se apex (a "ponta da língua") ou parte apical e a sua superfície o dorso. O apex e o dorso constituem o corpo da língua, enquanto a parte posterior é a raiz da língua. A massa principal da língua é constituída por uma série de quatro músculos: o superior longitudinal, o inferior longitudinal, o vertical e o transversal que permitem, durante a elocução, o alongamento ou constrição da língua no seu todo, ou só em segmentos específicos necessários à articulação dos sons da fala.

A faringe e cavidades nasais e oral formam o conjunto das cavidades supraglotais, que desempenham um papel fundamental na fonação de diferentes sons. Alterando a forma e as dimensões das cavidades supraglotais quando sujeitas a uma fonte sonora pode produzir-se uma grande variedade de sons. As cavidades supraglotais formam como que um conjunto de ressoadores que favorecem a passagem de algumas frequências e a atenuação de outras consoante as suas formas e dimensões. Às frequências favorecidas pelas cavidades supraglotais dá-se o nome de frequências formantes ou simplesmente formantes, e ao conjunto das formas tomadas pelas cavidades supraglotais chama-se trato vocal. Por vezes as cavidades nasais impõem anti-ressonâncias, denominadas por anti-formantes que se opõem ou desfavorecem a passagem de certas frequências sonoras.

A este modelo de produção da fala em que se separa a fonte sonora (excitação acústica que pode ser causada pela vibração das cordas vocais ou pelo simples fluxo de ar) da operação de filtragem realizada pelo trato vocal dá-se normalmente o nome de modelo de engenharia e pode ser representado como na figura 2.3.

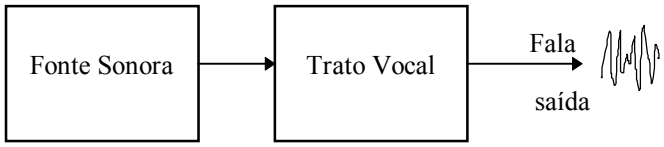


Figura 2.3 - Modelo de Engenharia.

2.2.1 Fonte Sonora

A fonte sonora no modelo de engenharia é um sinal excitador que pode ser periódico ou um sinal de ruído aleatório. O caso do sinal excitador ser periódico acontece nos sons vocalizados pelas cordas vocais com o abrir e fechar da glote. A frequência deste sinal é o inverso do tempo de duração do impulso glotal e é denominada por frequência fundamental (F0). O ruído aleatório, como sinal excitador, é produzido com a passagem de ar pela glote completamente aberta.

2.2.2 Frequência Fundamental

O valor da referida frequência fundamental varia com as pessoas, sendo também diferentes as gamas de valores de F0 para falantes masculinos, femininos e crianças. A gama típica de valores desta frequência para os homens é dos 80 aos 200 Hz, para as mulheres entre os 200 e 300 Hz e para as crianças dos 400 aos 500 Hz. Os valores de F0 para cada falante são ainda função de outras variáveis como sejam: o período

do dia (manhã, tarde e noite); estado nervoso, etc. Cada pessoa faz ainda variar esta frequência ao longo do seu discurso para dar a entoação pretendida, sendo neste aspecto conhecidas algumas variações típicas para frases do tipo interrogativa, declarativa, exclamativa como será referido no capítulo V.

Outro parâmetro importante do ponto de vista clínico relativamente à frequência fundamental é o "JITTER", como sendo a variação alternada de termo rápido, isto é inter-períodos do impulso glotal. Este parâmetro está associado à instabilidade física da fonte sonora e às suas patologias.

2.2.3 Formantes

No modelo de parametrização usado no decurso desta dissertação, a que é feita alusão no capítulo 4, denominam-se formantes como sendo zonas ou gamas de frequências mais favorecidas pelas cavidades supraglotais, as frequências de ressonância do trato vocal. As formantes correspondem às zonas mais escuras do tradicional espectrograma de banda larga de um sinal de fala.

O termo formante é um adjectivo atribuído às frequências que "formam" a fala.

A frequência da formante será assim a frequência central da região de amplitudes mais pronunciadas e a largura de banda será a sua banda a -3 dB da mesma zona no espectro quando analisado um intervalo de tempo curto do sinal de fala.

Cada som é caracterizado por um conjunto de formantes. Uma sequência especifica de sons dá origem a um fonema que é percebido pelo cérebro humano. Assim um fonema terá sempre ao longo do tempo uma variação idêntica de uma série de formantes.

Normalmente e durante esta dissertação denomina-se F1 como sendo a formante de frequência mais baixa ou a primeira formante, F2 a segunda formante, F3 a terceira formante e F4 a quarta formante.

A figura 2.4 apresenta o espectro alisado típico de um som da vogal [i] onde são visíveis os 4 formantes.

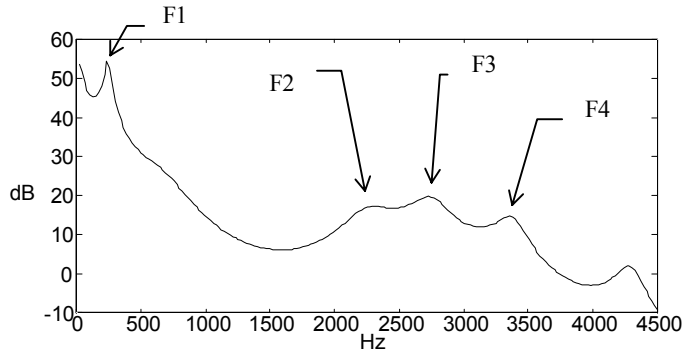


Figura 2.4 - Espectro alisado de um segmento de som da vogal [i].

Assim como há ressonâncias (formantes) também há anti-ressonâncias. Estas estão presentes sempre que haja um ressoador suplementar, como no caso das nasais, ou que a fonte sonora não esteja localizada na glote, como acontece na locução de algumas consoantes. Quando uma ressonância e um anti-ressonância estão próximas,

a segunda tende a cancelar a primeira. A presença de anti-ressonâncias reflecte-se, assim, sobretudo na amplitude dos formantes.

2.3 Fonemas Do Português

Segue-se uma caracterização dos traços acústicos e dos parâmetros das consoantes de modo a conseguir uma caracterização classificativa dos fonemas do português bem como conhecer o seu processo de articulação.

2.3.1 Caracterização de Alguns Traços Fonéticos

Apesar de os traços distintivos acústicos, propostos por [Jakobson 63], terem sido preteridos em relação aos traços distintivos articulatórios, propostos por [Chomsky 68], por considerarem [Mateus 90] este último tipo de traços mais adequado do ponto de vista da linguística e menos artificial no que diz respeito à fonética, passarei a caracterizar cada um dos traços distintivos acústicos do ponto de vista acústico e articulatório, bem como a caracterizar os fonemas ou classes de fonemas com estes traços, em virtude de serem definidos pelas suas características espectrais e a sua relação com as características de produção, permitindo-nos uma melhor compreensão e verificação de adequação dos modelos de produção de fala discutidos no capítulo 4.

2.3.1.1 Traços de Sonoridade

Tabela 2.1 - Traços de sonoridade. [Mateus 90], [Martins 92], [Silveira 86]

Traços de Sonoridade	Caracterização Acústica	Caracterização Articulatória	Exemplos
Vocálico/não Vocálico	- Presença/Ausência de uma estrutura formântica. - As anti-ressonâncias não são significativas. - Pequeno amortecimento dos formantes, o que origina larguras de banda relativamente estreitas. - Espectro de energia decrescente ao longo das frequências. - Globalmente os sons vocálicos apresentam uma	Excitação da glote associada a uma passagem livre do ar através do trato vocálico.	Vocálico: vogais e líquidas. Não vocálico: deslizantes e obstruintes.

	energia superior aos sons não vocálicos.		
Consonântico/não Consonântico	- Presença de uma energia globalmente baixa em oposição a energia globalmente alta. - Presença de anti-ressonâncias que afectam todo o espectro. - Larguras de banda significativamente maiores.	São produzidos por uma obstrução significativa do trato vocálico.	Consonântico: consoantes obstruintes e líquidas. Não consonântico: vogais e deslizantes.
Compacto /Difuso	- Predominância de formantes na zona central do espectro em oposição à predominância de formantes fora da zona central do espectro.	São produzidos com uma configuração do trato vocal em que os volume das cavidades anterior e posterior à constrição são próximos.	Compacto: vogais abertas (exemplo [a]), consoantes velares e palatais. Difuso: vogais fechadas (exemplo [i][u]), consoantes alveolares, dentais e labiais
Tenso/Relaxado	- Um som tenso tem maior duração e energia global.	Maior afastamento da configuração do trato vocálico da sua posição de repouso a que corresponde uma maior tensão muscular e um maior aumento da pressão da massa de ar atrás da constrição.	Tenso: vogais e consoantes longas e consoantes aspiradas. Relaxado: vogais e consoantes breves e consoantes não aspiradas.
Vozeado (ou vocalizado)/não Vozeado (ou não vocalizado)	- Presença/Ausência de uma excitação periódica de baixa frequência. - Estrutura formântica nítida/ou não.	Vibração periódica das cordas vocais.	Vozeado: vogais, liquidas, deslizantes, nasais e consoantes [b, d, g, v, z, j]. Não vozeado: consoantes [p, t, k, f, s].
Nasal/Oral	- Introdução de	Produzido pela	Nasal: consoantes

	anti-ressonâncias e frequências de ressonância adicionais. - Aumento da largura de banda. - Diminuição da energia dos formantes, particularmente de F1.	ressonância da cavidade nasal associada à ressonância da cavidade oral.	[m, n, nh] e vogais nasais. Oral: todas as que não são nasais.
Contínuo/não Contínuo	- Nos sons não contínuos ou interrompidos existe um ataque abrupto associado a alterações bruscas nas características espectrais em oposição aos sons contínuos em que o ataque é gradual.	Os sons não contínuos são caracterizados pelo surgimento ou desaparecimento rápido da fonte, devido a uma oclusão ou a uma abertura do trato vocálico.	Contínuo: fricativas e laterais. Não contínuo: oclusivas e vibrantes.
Estridente/não Estridente	- Maior intensidade de ruído. - Apresentam formas de onda extremamente irregulares, apresentando uma distribuição aleatória das energias no espectro.	O fluxo de ar é dirigido contra um obstáculo na vizinhança da constrição dando origem a um forte ruído de turbulência.	Estridente: fricativas [f, s, x, v, z, j].
Bloqueado/não Bloqueado	- Terminação abrupta que é no entanto menos proeminente que um ataque abrupto.	Produzido por uma compressão ou oclusão total da glote.	Bloqueado: implosivas, cliques e ejectivas.

Caberá aqui uma pequena explicação do que são ejectivas, implosivas, cliques e obstruintes.

Fechando completamente a glote e formando uma constrição no trato vocal, constitui-se um corpo de ar no interior da cavidade supralaríngea. A laringe pode ser movida na direcção vertical para cima ou para baixo.

Quando a laringe é movida para cima, resulta uma redução do volume de ar supraglotal e consequentemente a pressão atrás da constrição aumenta. As consoantes

produzidas com este movimento articulatório são conhecidas como ejectivas e podem ser oclusivas ou fricativas, conforme o grau de constrição imposto acima da glote pelos articuladores móveis (ex. [p], [v]).

Quando a laringe é movida para baixo, o volume orofaríngeo expande-se e a pressão correspondente baixará, ficando inferior à pressão atmosférica. Com a libertação da oclusão o ar exterior entra para a boca devido ao diferencial de pressão (ar glotal ingressivo). Os sons produzidos com este mecanismo de ar glotal ingressivo são geralmente vozeados e denominam-se consoantes implosivas. Estes sons podem ocorrer com as oclusivas vozeadas (ex. [b], [d], [g]).

Os cliques, vulgarmente usados como um som de desaprovação ou um som de incentivo aos cavalos, baseiam-se na criação de um volume de ar independente na região bucal, encostando o dorso da língua à região velar e formando uma constrição anterior. Uma vez estabelecido o volume de ar velar, o corpo da língua desce de modo a fazer baixar a pressão nessa região da boca. Com a libertação da oclusão anterior à oclusão velar, o ar exterior entra para a boca (mais uma vez mecanismo de ar ingressivo) produzindo assim um clique.

Os sons obstruintes ou não soantes são produzidos com uma configuração do trato vocálico que não permite o vozeamento espontâneo.

2.3.1.2 Traços de Tonalidade

Tabela 2.2 - Traços de tonalidade.

Traços de Tonalidade	Caracterização Acústica	Caracterização Articulatória	Exemplos
Grave/Agudo	Concentração da energia numa zona baixa das frequências no espectro (F2 próximo de F1) em oposição ao traço agudo a que corresponde uma concentração de energia numa zona alta de frequências no espectro (F3 próximo de F4).	Para o traço grave a cavidade de ressonância é mais ampla e menos compartimentada, pelo que este traço se produz em zonas periféricas em oposição ao traço agudo que se produz em zonas centrais.	Grave: vogais graves (exemplo [u]) e consoantes labiais e velares. Agudo: vogais agudas (exemplo [i]) e consoantes dentais e palatais.
Bemolizado/não Bemolizado	- Redução da energia de alguns ou todos os formantes.	Produz-se por um arredondamento dos lábios e um aumento da cavidade anterior à constrição.	Bemolizado: consoantes labializadas e as vogais [o, u].
Diesado/não Diesado	- Elevação das frequências nas componentes de	Elevação do corpo da língua simultaneamente	Permite distinguir consoantes palatizadas de não

	frequências mais altas.	com uma dilatação da cavidade faríngea.	palatizadas.

2.3.2 Parâmetros Classificatórios das Consoantes

Distinguem-se, tradicionalmente, dois grandes parâmetros classificatórios específicos das consoantes: o modo de articulação e o ponto de articulação. Correspondendo ao primeiro o modo da passagem do ar pelo trato vocal e o segundo à região do trato vocal em que se situa a maior constrição imposta pelos articuladores no canal bucal.

2.3.2.1 Modo de Articulação

A classificação do modo de articulação das consoantes é função da aproximação dos articuladores, da duração dessa aproximação ou da modificação da configuração do trato vocal devido à aproximação dos articuladores superiores e inferiores.

2.3.2.1.1 Consoantes Oclusivas

A articulação das consoantes oclusivas é realizada pelo impedimento da passagem de ar pelo canal bucal por um fechamento completo dos articuladores. Se o véu palatino estiver levantado e encostado à parede da faringe, impedindo a passagem do ar pelos canais nasais, o fluxo de ar sofre uma obstrução completa, formando-se uma oclusiva oral cujo som se produz logo que os articuladores se afastam. As oclusivas podem ser vozeadas, com vibração das cordas vocais ou não vozeadas, sem vibração das cordas vocais.

Tabela 2.3 - Oclusivas orais para o português.

Oclusivas		Causa da Oclusão
Vozeada	Não Vozeada	
[b]	[p]	Fechamento dos lábios.
[d]	[t]	Coroa da língua encostada aos incisivos superiores.
[g]	[k]	Dorso da língua encostada ao véu palatino.

Se o fluxo de ar puder passar pelas cavidades nasais, devido ao véu palatino estar descido produz-se uma oclusiva nasal. Estas consoantes são sempre vozeadas.

Tabela 2.4 - Oclusivas nasais para o português.

Oclusiva nasal	Causa da oclusão
[m]	fechamento dos lábios.
[n]	coroa da língua encostada aos incisivos superiores.

[nh]	lâmina da língua encostada ao palato.

2.3.2.1.2 Consoantes Fricativas

Na articulação destas consoantes os articuladores aproximam-se provocando uma obstrução parcial à passagem de ar, produzindo assim ruído. Os sons fricativos mais agudos são tradicionalmente denominados por sibilantes ([s] e [z]). As fricativas também podem ser vozeadas ou não vozeadas.

Tabela 2.5 - Fricativas para o Português.

Fricativa		Modo de obstrução à passagem de ar
Vozeada	Não Vozeada	
[v]	[f]	O lábio superior aproxima-se dos incisivos inferiores.
[z]	[s]	A coroa da língua aproxima-se da região dento-alveolar
[j]	[x]	A coroa da língua aproxima-se da região palato-alveolar.

2.3.2.1.3 Consoantes Laterais

São pronunciadas com uma obstrução do fluxo de ar provocado pela língua junto de um ponto da cavidade bucal, mantendo-se um canal para a saída do ar entre os lados da língua e o palato. Devido à fluidez que estas consoantes provocam, são muitas vezes denominadas de líquidas. Em português as consoantes laterais são sempre vozeadas.

Tabela 2.6 - Laterais para o português.

Consoante lateral	Obstrução
[l]	Formada pela ponta da língua junto dos alvéolos.
[lh]	Formada pela lâmina da língua junto do palato.

2.3.2.1.4 Consoantes Vibrantes

A designação destas consoantes advém do facto de o órgão articulador móvel utilizado na sua produção vibrar, ou tocar repetidamente no outro articulador.

Tabela 2.7 - Vibrantes para o português.

Consoante vibrante	Denominação	Articulação
[r]	Vibrante alveolar	Uma única obstrução provocada pela ponta da língua junto dos alvéolos (ex. caro).
[$\tilde{r}$]	Vibrante alveolar múltipla	A ponta da língua toca várias vezes os alvéolos (ex. carro).

[R]	Vibrante velar	Vibração da parte de trás da língua junto do velo.

2.3.2.1.5 Consoantes Africadas

Iniciam-se por uma oclusão completa e terminam com uma constrição própria das fricativas. Em Portugal a única consoante africada é não vozeada e existe apenas em alguns dialectos, representa-se por [t•] (ex. "tchau"). No Brasil as africadas podem ser vozeadas, [dz], ou não, [t•].

2.3.2.2 Ponto de Articulação

Tabela 2.8 - Ponto de articulação.

Denominação quanto ao ponto de articulação	Articuladores	Exemplos
Bilabiais ou labiais	Os dois lábios.	[b], [p], [m]
Labiodentais	O lábio inferior e os incisivos.	[v] e [f]
Dentais	Ponta da língua e os incisivos.	[d], [t], [z], [s] por vezes [e] e [n]
Alveolares	Ponta da língua e os incisivos superiores.	[l], [n] [r]
Pré-palatais	A lâmina da língua e o pré-palato.	[z] [x]
Palatais	A lâmina da língua e o palato.	[lh] [nh]
Velares	A parte de trás da língua e o véu palatino.	[g], [k], [R]

2.3.3 Lista de Fonemas do Português

É chegada a altura de apresentar o alfabeto fonético para o português usado neste trabalho [Martins 92], bem como a sua transcrição fonética [ ] e um exemplo da utilização de cada fonema numa palavra:

Vogais Orais		**Vogais Nasais**		**Semi Vogais**	
[i]	l***i***vro	[ĩ]	p***in***to	[j]	pa***i***
[e]	P***e***dro	[ẽ]	d***en***te	[w]	pa***u***
[ɐ]	t***e***rra	[α̃]	c***an***to		
[a]	p***a***to	[õ]	p***on***te		
[α]	m***a***no	[ũ]	f***un***do		
[ɔ]	g***o***la				
[u]	p***u***lar				

[ɛ] s***e***car

[o] p***o***ço

Consoantes Oclusivas Orais

[p] ***p***ara

[b] ***b***ata

[t] ***t***arde

[d] ***d***a***d***o

[k] ***c***ão

[g] ***g***ato

Consoantes Oclusivas Nasais

[m] a***m***a

[n] ***n***ada

[ɲ] pi***nh***o

Consoantes Fricativas

[f] ***f***ado

[s] ***s***ábado

[ʃ] ***ch***ão

[v] ***v***aca

[ʒ] ca***s***a

[z] ***j***ardim

Consoantes Laterais

[l] ***l***ado

[ʎ] fi***lh***o

Consoantes Vibrantes

[ɾ̃] po***r***ta

[r] ca***rr***o

Quando uma vogal aparece junta com uma semi vogal chama-se ditongo ao conjunto das duas.

CAPÍTULO 3

RECOLHA DO SINAL DE FALA

3. RECOLHA DO SINAL DE FALA

Neste capítulo discutem-se as condições em que deve decorrer a recolha dos sinais de fala e caracteriza-se o sistema de aquisição e armazenamento em disco de computador de uma base de sinais usada neste trabalho.

3.1 Introdução

Existem diversos modelos para a análise da fala, sendo por vezes distintos os sinais analisados e os processos de recolha desses sinais.

Assim, alguns processos exigem a captação independente do sinal na boca, para estudo do trato vocal, no nariz, para detecção de nasalização ou não e na glote, para total conhecimento do sinal excitador (detecção de vocalização e no caso de sinais vocalizados conhecimento do período fundamental e o seu sincronismo). Estes sinais são armazenados por um gravador, em pistas independentes de uma fita magnética, podendo depois ser representados num osciloscópio e gravados em câmara de vídeo. A este processo dá-se o nome de oscilografia articulatória.

Outro processo denominado cine-radiografia consiste em filmar a imagem do movimento de articulação da fala. Um sujeito, enquanto fala, está colocado entre um emissor e um captador de raios X. A cabeça é mantida fixa por um cranióstato. Os raios X são transmitidos a um amplificador de brilho. Nesse amplificador a câmara de filmar recolhe as imagens. Simultaneamente o sinal acústico é recolhido por um gravador magnético, o que permite uma análise da produção articulatória e uma análise acústica sobre o mesmo segmento de fala.

Um dispositivo mais actual e com uma finalidade diferente é o electroglotógrafo que permite analisar com uma maior especificidade o impulso glotal. Consiste num emissor e um transdutor de corrente colocados dos lados direito e esquerdo do pescoço envolvendo a glote. A corrente do emissor chega ao transdutor através da glote. A impedância da glote varia ao longo do impulso glotal, sendo mínima quando a glote está fechada, máxima quando a glote está aberta e tem valores intermédios entre a abertura e o fecho da glote. Assim a forma de onda da corrente no transdutor permite ter uma "imagem" da posição da glote. Este sinal, muito próximo do impulso glotal, conjuntamente com o sinal de fala adquirido com um microfone, é muito útil para estudar problemas associados com as cordas vocais e permite um conhecimento preciso do início e fim do impulso glotal, necessário para a análise síncrona com o período fundamental como será explicado no capítulo 7.

O processo mais usado, na recolha de sinais de fala para análise, consiste na aquisição, com um microfone, do sinal acústico produzido por um falante, a sua amplificação por um pré-amplificador de sinal e o armazenamento em fita magnética por um gravador para posterior conversão para digital ou imediata conversão para digital e respectivo armazenamento em formato digital (ver figura 3.1). Contudo este processo exige determinadas condições e meios para se obter uma boa recolha.

O sinal deve ser recolhido numa câmara insonorizada ou anecóica com um microfone unidireccional e linear nas frequências, amplificado com um pré-amplificador de sinal linear e recolhido por um gravador com boa qualidade para uma fita magnética também com boa qualidade. Para a conversão analógico/digital deve ser usado um

filtro anti-aliasing, a frequência de amostragem deve ser criteriosamente escolhida bem como o número de bits usados na conversão. O locutor que pronunciará o texto deve ser bem identificado e o texto a pronunciar deverá ser escolhido de acordo com o tipo de análise que se pretende realizar.

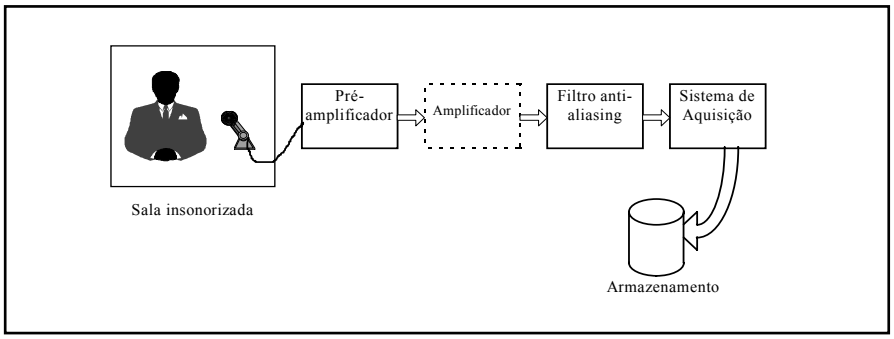


Figura 3.1 - Sistema de Recolha.

3.2 Elementos Importantes na Aquisição do Sinal de Fala

3.2.1 Câmara Insonorizada

Para evitar fenómenos de outras fontes de ruído e de reverberação o sinal deve ser recolhido numa câmara insonorizada. Esta câmara deve ter o menor número possível de pontos de contacto com o exterior, uma só porta isolada acusticamente que se deve manter fechada durante a aquisição. A câmara deve ser revestida de materiais que absorvem as ondas sonoras evitando assim a reflexão das ondas acústicas, que provocaria modificação das formas de onda recolhidas.

3.2.2 Microfone

Apesar dos cuidados a ter com a câmara insonorizada para que não apareçam reverberações, estas podem estar presentes por alguma deficiência da câmara por algum elemento no interior desta que reflicta as ondas sonoras ou ainda pelo facto de as ondas sonoras não serem completamente absorvidas pelo material que reveste a câmara. Assim será de grande importância que o microfone usado seja tanto quanto possível unidireccional, isto é, que a atenuação das ondas sonoras vindas da frente do microfone seja mínima e a atenuação das ondas que aparecem pelos lados e por trás do microfone seja grande. Neste caso acentua-se a importância de o falante se colocar numa posição estável em frente do microfone. É ainda essencial que o microfone tenha características de linearidade na gama de frequências desde sensivelmente os 20 Hz até aos 20 Khz para que não haja distorção de nenhuma componente espectral do sinal.

3.2.3 Pré-amplificador

Este elemento também necessário no sistema de recolha deverá ter uma característica também linear na mesma gama de frequências do microfone pelas mesmas razões. Será desejável também uma relação sinal ruído não inferior a 55 dB.

3.2.4 Filtro Anti-Aliasing

Na conversão do sinal analógico com valores contínuos no tempo para sinal digital com valores discretos no tempo por processo de amostragem é necessário evitar fenómenos de aliasing.

O teorema da amostragem diz-nos que se deve amostrar o sinal a uma frequência (Fs) superior a duas vezes a maior componente espectral do sinal para evitar o fenómeno de aliasing que se caracteriza pela distorção do sinal original pela dobragem das componentes espectrais do sinal superiores a metade da frequência de amostragem (frequência de Nyquist Fn) em componentes inferiores a esta frequência.

Por exemplo um sinal amostrado com Fs=10 Khz sendo Fn=5 Khz deve ter componentes espectrais nulas acima de Fn, porque por exemplo, uma componente de 6 Khz aparece sobreposta à componente de 4 Khz, a componente de 7 Khz sobrepõem-se à componente de 3 Khz acontecendo o mesmo para todas as componentes acima de Fn.

Este fenómeno pode-se evitar por dois procedimentos diferentes:

- Aumentando a frequência de amostragem para o dobro da maior componente espectral do sinal a amostrar.

- Passando o sinal por um filtro anti-aliasing. Este é um filtro passa baixo com uma frequência de corte igual à frequência de Nyquist. Esta deve ser a máxima frequência do sinal que interessa preservar.

O primeiro processo é inconveniente por não ser conhecido à partida a máxima componente espectral do sinal. Por outro lado pode levar a frequências de amostragem muito elevadas obrigando a um número exagerado de amostras tornando demasiado extenso o seu armazenamento e preservando componentes espectrais por vezes não desejadas. O segundo processo é normalmente o escolhido obrigando a que o filtro possua características de linearidade na banda passante, um pequeno "overshoot", uma pendente de corte acentuada e uma atenuação elevada na banda de rejeição.

3.2.5 Frequência de Amostragem

Como acaba de ser dito a frequência de amostragem, usando o filtro anti-aliasing, deve ser superior ao dobro da máxima componente espectral com interesse, sendo ajustada a frequência de corte do filtro para a máxima componente espectral.

Para os diversos modelos de sinais de fala, as frequências de interesse vão desde os 60 Hz até aos 4 a 10 Khz, sendo então utilizada uma frequência de amostragem com valores entre os 8 e 20 Khz. Valores abaixo dos 8 Khz cortam componentes espectrais com interesse. Valores acima dos 20 Khz aumentam desnecessariamente a área de armazenamento do sinal amostrado.

3.2.6 Número de Bits

Quanto maior o número de bits usados melhor a resolução do sinal. O número de níveis diferentes disponíveis com n bits é de 2^n. Para este tipo de análise é bom ter a melhor resolução possível no entanto não interessa ter uma resolução tão pequena (melhor resolução, n maior) de modo que o nível de ruído exceda esta resolução. A desvantagem de ter um grande número de bits é o aumento da área de armazenamento. Ter um número baixo de bits pode levar a uma resolução tão má que não permita distinguir diferentes níveis importantes do sinal.

Para se tirar o melhor partido do número de bits usado é importante ajustar o nível máximo do sinal para o valor fim de escala do conversor A/D.

Para os sinais de fala são usados habitualmente entre 10 a 16 bits.

3.2.7 Falante

O sujeito falante, também chamado o informante, que produz a fala a ser recolhida deve ser bem identificado quanto aos elementos que podem ser determinantes na análise da fala [Martins 92]. Assim, deve-se especificar o seu local de nascimento e locais onde viveu, é importante quanto à sua variação dialectal, a idade, o sexo, habilitações académicas, nível sociocultural e o seu eventual conhecimento prévio dos objectivos da gravação que realiza.

Devem-se criar condições para que o informante esteja confortável e não se sinta num ambiente estranho.

Quando se pretende analisar a fala e não o falante, como é o caso, o texto deve ser reproduzido por mais que um falante.

3.2.8 Texto

O texto a ser recolhido deve ser criteriosamente seleccionado de modo a que haja nesse texto uma grande riqueza das situações que se pretendem estudar. Assim, quando se pretende analisar uma vogal esta deve ser reproduzida continuamente no tempo. Se a prosódia é o objecto de estudo o texto deve ser composto por frases de diversos tamanhos do tipo que se pretenda: declarativo, interrogativo ou exclamativo. No caso interrogativo é ainda diferenciada a existência ou não de palavra interrogativa (exp.: quanto, onde, quem, etc.). Quando se realiza uma análise com objectivo de síntese por difones, em que é importante a junção de todos os fonemas com todos, o texto pode ser escolhido sob a forma de vocábulos, palavras ou frases de modo que se reunam todas as junções de fonemas pretendidas. Pretendendo-se a análise da fala vocalizada deve-se escolher uma vogal contínua, um conjunto de vogais, uma palavra com todos os sons vocalizados ou mesmo uma frase apenas com esses sons. Muitas outras situações se poderiam referir no entanto torna-se desnecessário já que estes exemplos esclarecem eficientemente a escolha criteriosa do texto a reproduzir.

Uma medida metódica importante na recolha do sinal de fala é catalogar/etiquetar convenientemente cada trecho de fala gravada para facilitar a consulta desses elementos, caso contrário esta pode-se tornar complicada quando o número de trechos gravados é grande.

O processo de catalogação/etiquetagem é diferente conforme os sistemas em que se armazena o sinal. No armazenamento em fita magnética é desejável que o início de cada trecho seja colocada uma marca numerada e criar uma lista com essas marcas e o correspondente texto do trecho. Assim com um gravador que permita o rápido acesso a essas marcas será fácil aceder ao texto desejado. Quando o sinal é armazenado digitalmente em ficheiros, o nome destes deve ser indiciador do trecho ou do texto que contém.

3.3 Condições em que Decorreu a Recolha/Aquisição dos Sinais de Fala Neste Trabalho

Conhecidas as condições ideais para a recolha do sinal, nem sempre é possível recriar paulatinamente todas essas condições, ora por falta de material, ora por o material disponível não se adaptar perfeitamente ao exigido.

Contudo neste trabalho procurou-se, com sucesso, seguir com bastante rigor as condições explicadas anteriormente, com a variante do sinal ser inicialmente armazenado em fita magnética num estúdio de rádio e depois reproduzido por um gravador para ser então amostrado por uma placa de som e armazenado digitalmente em ficheiros por um computador como mostram as figuras 3.2 e 3.3 com as condições a seguir explanadas.

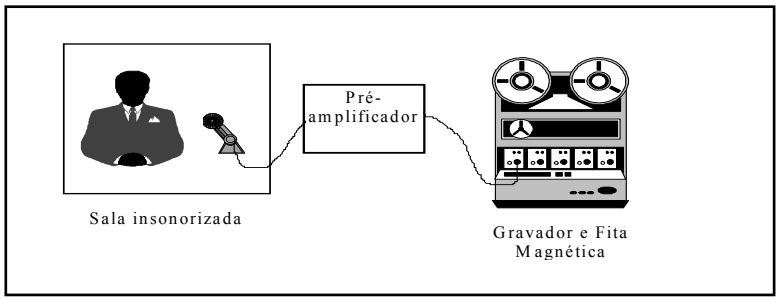


Figura 3.2 - Recolha do Sinal.

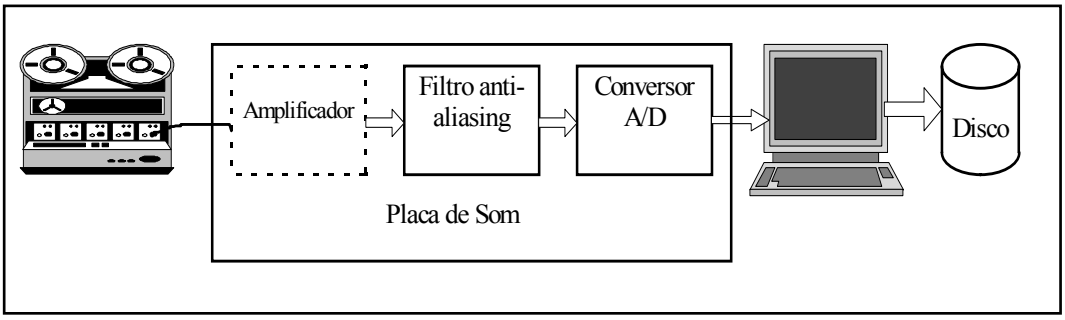


Figura 3.3 - Amostragem do Sinal.

3.3.1 Câmara Insonorizada

Este é um dos elementos menos fáceis de conseguir já que se trata de um compartimento exclusivamente dedicado a este fim. Assim o local escolhido para a recolha dos sinais de fala foi o estúdio 1 da RDP - Rádio Difusão Portuguesa Porto que gentilmente nos facilitou a utilização dos seus meios técnicos tais como a referida sala insonorizada, o microfone, gravador e um técnico de som.

Este estúdio possui todas as características exigidas a uma boa sala insonorizada como seria de esperar a um estúdio de emissão de uma instituição como a RDP.

3.3.2 Armazenamento do Sinal em Fita Magnética

O processo de recolha e armazenamento em fita magnética do sinal referente à figura 3.2 decorreu no referido estúdio tendo sido usado um microfone cardióide , um gravador STUDER de estúdio 1/4 de polegada 19 cm/seg., e a fita magnética foi BASF LGR 50.

Foi registado numa tabela para cada trecho gravado o texto correspondente e a posição na fita magnética do inicio do referido trecho.

3.3.3 Reprodução e Conversão Analógico/Digital com Recurso a Uma Placa de Som

Como ilustrado na figura 3.3 na passagem dos sinais de fala da fita magnética para o computador foi usado um gravador portátil NAGRA IV-D na reprodução dos sinais e estes foram adquiridos e armazenados com recurso à placa de som SOUND BLASTER 16 ASP. O computador usado neste processo foi um 486 DX2 a 66 Mhz, com 4MB de memória de dados e 540 MB de capacidade de armazenamento em disco.

3.3.3.1 Facilidades Oferecidas Pela Placa de Som

A utilização de uma placa de som (entenda-se "hardware" e "software") para adquirir o sinal impõe algumas restrições como sejam a frequência de amostragem, o número de bits usados e o facto de os sinais terem de ser armazenados em ficheiros tipo MIDI ou WAV, no entanto os valores permitidos para estes parâmetros estão dentro do considerado razoável para esta aplicação.

A placa de som tem ainda a vantagem de ser de fácil utilização, o que permite a qualquer momento adquirir novas amostras da fita magnética, não obrigando assim à reserva de uma grande área de memória do disco. Permite ainda, a reprodução de fala directamente para um microfone ligado à placa de som, devendo-se no entanto, tomar o devido cuidado relativamente à qualidade do microfone usado e às condições acústicas da sala onde este processo decorre. Outra facilidade oferecida pela placa é a possibilidade de reproduzir em qualquer momento quer o sinal adquirido quer outro sinal sintetizado por software desde que guardado em ficheiro do tipo MIDI ou WAV.

A placa permite amplificar o sinal a adquirir para ajustar ao valor fim de escala aproveitando assim toda a resolução oferecida pelo número de bits usado, o "software" permite cortar partes do sinal sem interesse não ocupando assim

desnecessariamente área de armazenamento e permite ainda monitorizar o sinal adquirido com ou sem amplificação e depois de cortado ou não, para verificação de que o sinal está nas condições pretendidas, antes de ser armazenado em disco.

Estas são as vantagens da utilização da placa de som do ponto de vista da realização de um trabalho desta natureza, recolha e reprodução de sinais.

3.3.3.2 Frequência de Amostragem

As frequências de amostragem permitidas pela placa de som são de 44.1 Khz, 22.05 Khz, 11.025 Khz e 5512.5 Hz.

Os 44.1 Khz são muito altos assim como os 5512.5 Hz são muito baixos. Optou-se então pela frequência de 22.05 Khz, deixando a possibilidade, que acabaria por ser aproveitada, de realizar a decimação de 2:1 para se obter a frequência de amostragem de 11.025 Khz.

3.3.3.3 Resolução

As facilidades oferecidas pela placa são de 8 ou 16 bits, optou-se pelos 16 bits já que os 8 não oferecem uma resolução pretendida para este trabalho. Assim resultam 2^{16}=65536 níveis diferentes de quantificação de amplitudes dos sinais..

3.3.3.4 Filtro Anti-Aliasing

Apesar das características técnicas da placa não incluírem qualquer informação sobre o filtro anti-aliasing foi decidido efectuar medidas para avaliar algumas características deste filtro para eventual compensação.

Com um gerador de sinal foram geradas ondas sinusoidais de amplitude constante e a várias frequências, sendo estas adquiridas pela placa de som com o mesmo número de bits e frequência de amostragem usados para a aquisição dos sinais de fala. Posteriormente foi determinada a energia dos sinais adquiridos às diferentes frequências e obteve-se a resposta do filtro representada na figura 3.4 .

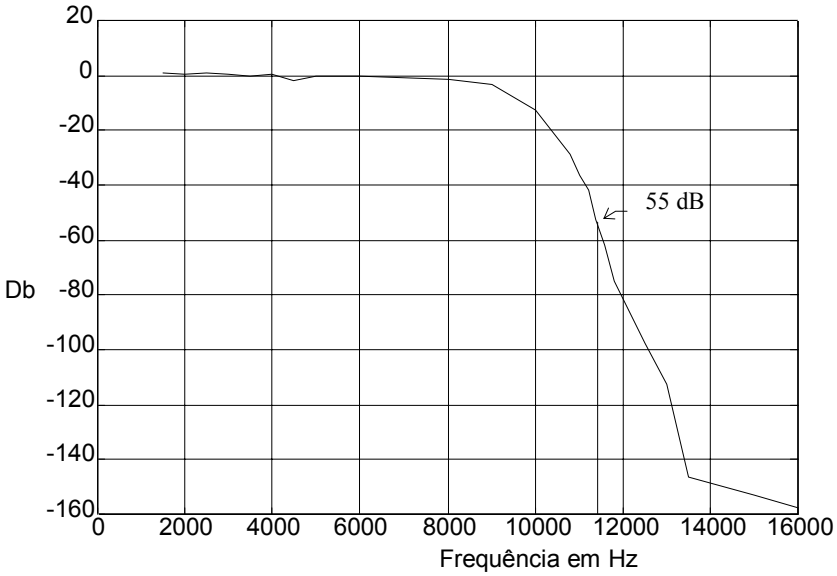


Figura 3.4 -Função de Transferência medida do filtro anti-aliasing da placa de som SOUND BLASTER.

De notar que a frequência de amostragem foi de 22.05 Khz, sendo a frequência de Nyquist de 11.025 Khz, pelo que se admite como bom, para a esta aplicação, o filtro anti-aliasing usado. De facto, a atenuação introduzida nesta frequência é da ordem de grandeza da relação sinal/ruído desejada na recolha dos sinais (55 dB).

Já se referiu que foi aproveitada a possibilidade de realizar uma decimação de 2:1 do sinal.

A decimação realiza uma reamostragem do sinal com uma taxa de amostragem de 1/R vezes a taxa de amostragem original. O resultado é um sinal R vezes mais curto com uma frequência de amostragem R vezes menor. Neste caso o valor de R utilizado foi 2.

A operação decimação filtra o sinal com um filtro passa baixo (filtro anti-aliasing) de Chebyshev do tipo I de oitava ordem com uma frequência de corte 0.8*(Fs/2)/R, antes de reamostrar e com um "ripple" na banda de passagem com um valor de 0.5 dB.

A figura 3.5 mostra a função de transferência deste filtro com a frequência de amostragem original de 22.05 Khz e a nova frequência de amostragem de 11.025 Khz.

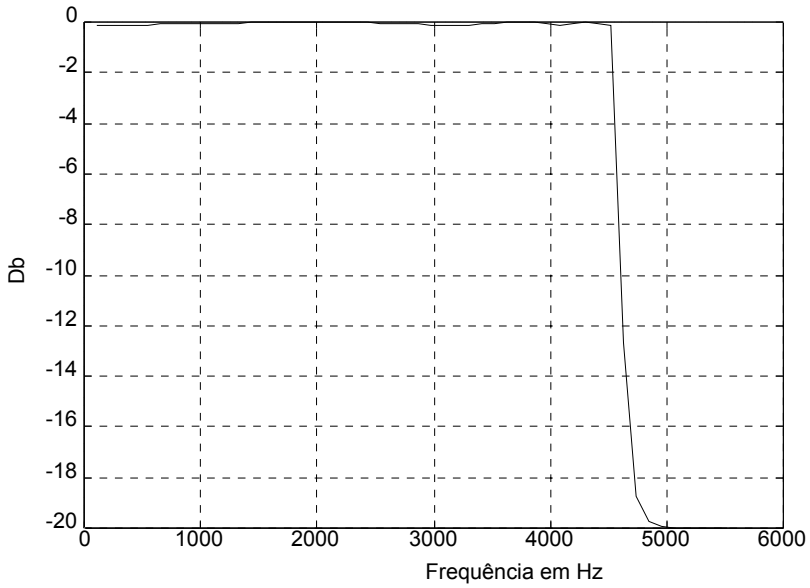


Figura 3.5 - Função de Transferência do filtro passa-baixo usado na decimação de 2:1.

Chama-se a atenção para a nova frequência de corte que é de 0.8 vezes a frequência de Nyquist, tendo um valor de sensivelmente 4.4 Khz.

3.3.3.5 Armazenamento do Sinal

Os sinais foram armazenados em ficheiros do tipo WAV sendo o nome de cada ficheiro identificativo da palavra, tipo de frase ou vogal contida. No nome de cada ficheiro consta ainda uma numeração 1 ou 2 que identifica o informante 1 ou 2.

Não foram armazenados em disco todos os sinais contidos na fita magnética, apenas os mais frequentemente usados, já que a leitura com a placa de som de outros trechos desejados pode ser realizada em qualquer momento, não ocupando assim desnecessariamente espaço em disco.

3.3.4 Texto Escolhido

O texto escolhido para recolha na sala insonorizada da RDP está partido em trechos curtos como sejam: junção de dois fonemas, uma vogal pronunciada continuamente, palavras isoladas, algumas frases declarativas e interrogativas curtas e finalmente um parágrafo de texto em leitura fluente com sensivelmente 5 linhas.

O texto é composto por:

- Todas as vogais pronunciadas continuamente.

- As semi-vogais juntas com todas as vogais, formando assim todos os ditongos possíveis.

- Todas as consoantes em posição inicial, final e intermédia de palavra, salvo as consoantes que não aparecem nunca em posição inicial (exemplo: [r], [□], […]) ou em posição final (exemplo: [□], […], consoantes oclusivas orais).

- Todas as consoantes juntas com todas as vogais nas duas sequências possíveis, VC (vogal-consoante) e CV (consoante-vogal).

Este texto foi estudado com o objectivo de cobrir todas as situações de ocorrência de todos os fonemas, nas diferentes posições da palavra e junto com qualquer outro fonema tendo em vista a sua parametrização para uso em sistema de conversão texto-fala para o português baseado em difones.

3.3.5 Falantes

Foram escolhidos dois sujeitos com fala relativamente harmoniosa do sexo masculino com idades de 32 e 40 anos, ambos com nascimento e vivência na cidade do Porto e com um nível sociocultural elevado (ambos docentes universitários) e com um prévio conhecimento total dos objectivos da gravação. O primeiro, identificado como informante 1 tem o nome de Rui Azevedo Guedes, o segundo identificado como informante 2 é o orientador desta dissertação Prof. Diamantino Rui da Silva Freitas.

CAPÍTULO 4

MODELIZAÇÃO DOS SINAIS DE FALA / CODIFICAÇÃO PARAMÉTRICA

4. MODELIZAÇÃO DOS SINAIS DE FALA / CODIFICAÇÃO PARAMÉTRICA

4.1 Introdução

O objectivo para todos os métodos de codificação paramétrica é usar a redundância inerente aos sinais de fala para reduzir a quantidade de informação usada na transmissão ou armazenamento. A redundância advém de dois aspectos: o mais óbvio é a repetição periódica das formas de onda, e, o segundo, é a presença de componentes de ruído em alguns sons de fala, para os quais a reconstrução exacta da forma de onda não é perceptualmente importante.

Podemos examinar e redundância considerando quatro tipos diferentes de sinal: vocalizado, não vocalizado, misto (em que há uma sobreposição dos dois primeiros tipos) e silêncio. Primeiro, para a fala vocalizada, a forma de onda do sinal é quase-periódica e a uma cadência correspondente à frequência do impulso glotal. Esta razão de periodicidade pode-se alterar ao longo da duração de segmentos de fala e a forma de onda periódica normalmente varia gradualmente de período para período. Segundo, para a fala não vocalizada, o sinal é algo parecido com ruído aleatório produzido pela turbulência do fluxo de ar nas restrições do trato vocal. Este ruído tem um espectro de energias parecido com o espectro do ruído branco ou uniforme, sendo depois moldado pelas ressonâncias do trato vocal. Terceiro, os dois tipos de sons ocorrem juntos nas fricativas vocalizadas (para o português: [v], [j] e [z]) e nas transições entre sons. Finalmente, nada está presente durante os períodos de silêncio entre segmentos de fala, as pausas, geralmente observadas entre palavras e antes das consoantes oclusivas. A figura 4.1 apresenta um sinal de fala correspondente à locução de "leitura de um parágrafo" pelo locutor 2.

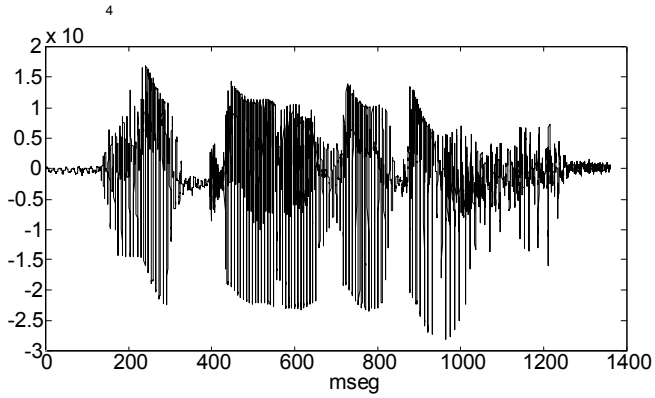


Figura 4.1 - Sinal de fala correspondente à locução de "leitura de um parágrafo".

Em qualquer destes casos é necessário ter mais informação do que a estritamente requerida para identificar o tipo de segmento. Essa informação é representada de forma paramétrica existindo uma considerável extensão de métodos de codificação, que reduzem a redundância, recorrendo a modelos que separam a componente de excitação dos sons de fala das componentes do envelope espectral.

A excitação é caracterizada como um trem de impulsos para os sons vocalizados ou ruído aleatório para os não vocalizados.

O envelope espectral que é produzido pelas ressonâncias/anti-ressonâncias do trato vocal pode ser caracterizado pelos parâmetros de um filtro frequencial selectivo que tenha a mesma característica de transferencia que o trato vocal.

A vantagem da codificação advém da lentidão relativa a que os parâmetros se alteram, tanto do sinal de excitação como da característica do filtro. É difícil determinar exactamente a razão temporal a que os parâmetros se alteram, isso depende do falante e do seu discurso, contudo, é raro que os parâmetros tenham valores significativamente diferentes dos observados 5 ms antes e mesmo que estes tenham valores similares aos de 30 ms antes. Por isso, a razão segmental, 1/ts, a que se definem os segmentos é habitualmente seleccionada de forma que a duração dos segmentos, ts, esteja na gama 5 ms< ts < 30 ms [Rowden 92].

4.2 Modelos de Produção da Fala

Surgem naturalmente diferentes modelos de produção da fala sendo, porém, comum a todos, a separação da excitação do trato vocal.

4.2.1 Modelos de Tubos Sem Perdas

Um modelo largamente usado na produção da fala é baseado na suposição de que o trato vocal pode ser representado como a concatenação de tubos acústicos sem perdas como na figura 4.2.

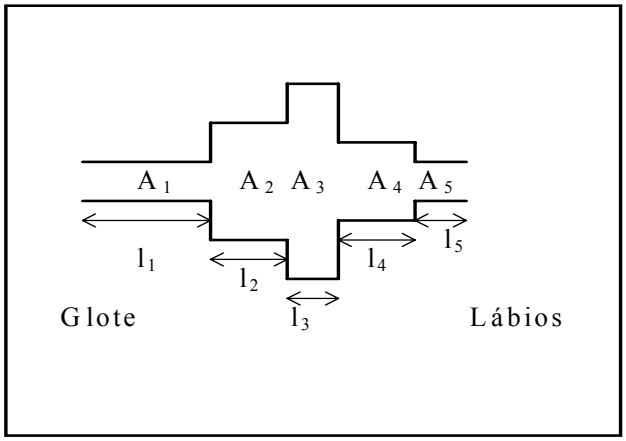


Figura 4.2 - Concatenação de 5 tubos acústicos sem perdas.

As áreas transversais A_k dos tubos são escolhidas para aproximar a função[i], A(x), da área do trato vocal. Se for usado um grande número de tubos com um comprimento pequeno, pode-se esperar razoavelmente que as frequências de ressonância da concatenação dos tubos sejam próximas das de um tubo com uma função de área continuamente variável. Contudo, apesar de esta aproximação desprezar as perdas

[i] A função de área referida indica para cada instante as dimensões transversais do trato vocal em função do seu comprimento, x, a partir da glote e até aos lábios.

devido à fricção, à variação de condutibilidade com o calor e à vibração das paredes também se pode esperar razoavelmente que as larguras de banda das ressonâncias difiram das de um modelo detalhado que inclua essas perdas. Daqui advém a denominação "sem perdas" para este modelo. No entanto as perdas podem ser tomadas em consideração na modelização da glote e dos lábios de forma a representarem fielmente as propriedades de ressonância dos sinais de fala, como mostram [Rabiner 78], [Rowden 92], [Schafer 70] e [Flanagan 64].

Uma razão importante para a apresentação deste modelo é o facto de o modelo de tubos sem perdas ser uma transição conveniente entre os modelos contínuos e os modelos discretos no tempo.

4.2.2 Modelo de Engenharia

Uma aproximação válida e a mais usada para representar os sinais de fala é o modelo de terminais análogos como mostra a figura 4.3.

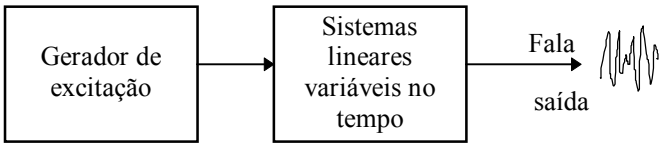


Figura 4.3 - Modelo de engenharia.

Trata-se de um modelo linear em que a saída tem as propriedades desejadas da fala quando este é controlado por um grupo de parâmetros relacionados com o processo de produção da fala.

O modelo é algo parecido com o modelo físico de produção de fala humana mas a sua estrutura interna não imita fisicamente a produção da fala.

Para produzir sinais parecidos com a fala o modo de excitação e as propriedades de ressonância do sistema linear devem variar com o tempo. A natureza desta variação já foi comentada na introdução deste capítulo e recorda-se que, embora sendo fortemente não estacionário, o sinal de fala mantém as características por um período de 5 a 30 ms. Então, o modelo de engenharia envolve uma variação lenta no tempo de um sistema linear excitado por um sinal cuja natureza básica pode ser um trem de impulsos quase periódicos para a fala vocalizada, um ruído aleatório para a fala não vocalizada ou um sinal misto para fricativas vocalizadas e zonas de transição.

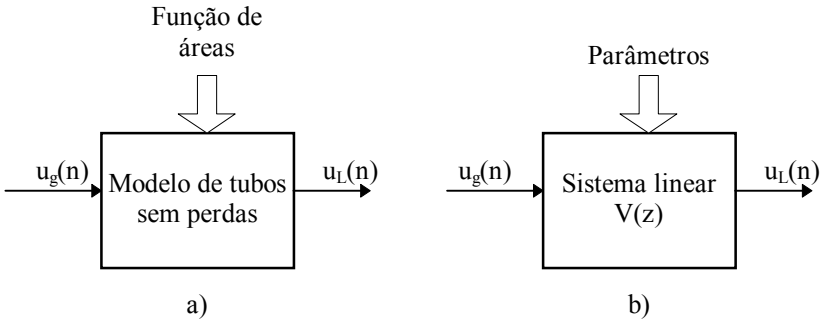


Figura 4.4 - Diagrama de blocos representando: a) modelo de tubos sem perdas; b) modelo de engenharia.

O modelo de tubos sem perdas apresentado atrás serve como uma introdução ao presente modelo, de que se apresentam as funcionalidades essenciais na figura 4.4.

O sistema de trato vocal da figura 4.4 a) é caracterizado por um conjunto de áreas e o da figura 4.4 b) por um outro conjunto de parâmetros.

A relação entre a entrada e a saída do sistema linear pode ser representada pela função de transferencia V(z) da forma

$$V(z) = \frac{G}{1 - \sum_{k=1}^{N} \alpha_k z^{-1}} \tag{4.1}$$

onde G e $\{\alpha_\kappa\}$ dependem muito de perto da função de área.

Conjuntamente com a resposta do trato vocal, um modelo completo de engenharia, inclui a variação da função de excitação e dos efeitos da radiação do fluxo de ar nos lábios.

De seguida será examinado separadamente cada componente do sistema de produção de fala e então combinado num modelo completo.

4.2.2.1 Trato Vocal

As ressonâncias da fala, modelizadas por formantes, correspondem aos pólos da função de transferencia V(z). Um modelo só de pólos é uma muito boa representação dos efeitos do trato vocal para a maioria dos sons da fala. Todavia, sabemos da teoria acústica que os sons nasais e fricativos requerem ressonâncias e anti-ressonâncias (pólos e zeros) para a sua correcta representação. Nestes casos devemos incluir zeros na função de transferência ou aumentar o número de pólos que segundo [Rabiner 78] simula o efeito de um zero na função. Em muitos casos é preferível esta segunda aproximação.

Desde que os coeficientes do denominador de V(z) sejam reais, as raízes do polinómio denominador serão reais ou pares complexos conjugados. A típica frequência complexa de um ressoador do trato vocal é:

$$s_k, s_k^* = -\sigma_k \pm j2\pi F_k \tag{4.2}$$

e os correspondentes pólos complexos conjugados serão:

$$\begin{aligned} z_k, z_k^* &= e^{-\sigma_k T} e^{\pm j2\pi F_k T} \\ &= e^{-\sigma_k T}\cos(2\pi F_k T) \pm j e^{-\sigma_k T} sen(2\pi F_k T) \end{aligned} \tag{4.3}$$

em que T=1/Fs.

As larguras de banda das ressonâncias do trato vocal são aproximadamente $2\sigma_k$ e a frequência central é $2\pi F_k$ [Rabiner 78].

No plano Z, o raio desde a origem até ao polo determina a largura de banda, pois

$$|z_k| = e^{-\sigma_k T} \tag{4.4}$$

e o angulo nesse plano determina a frequência central

$$\theta_k = 2\pi F_k T \tag{4.5}$$

Assim se o denominador de V(z) for factorizado, as correspondentes frequências dos formantes e larguras de banda serão encontradas usando estas expressões.

A frequência natural complexa do trato vocal humano encontra-se no semi-plano esquerdo do plano s, pelo que estamos perante um sistema estável. Então, $\sigma_k > 0$ e $|z_k| < 1$, isto é, todos os pólos correspondentes ao modelo se devem situar dentro do circulo unitário do plano z, que é condição para garantir estabilidade do modelo.

Para a realização da função do trato vocal V(z) pode-se usar uma implementação de forma directa de V(z) como na figura 4.5.

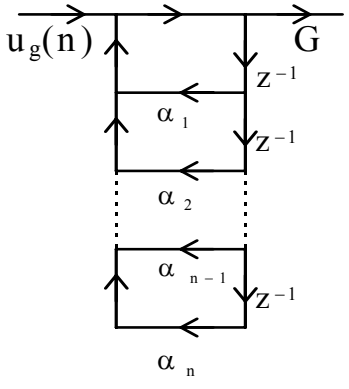


Figura 4.5 - Implementação directa da função de transferência do modelo só com pólos.

Alternativamente pode-se representar V(z) como uma cascata de sistemas ressoadores de segunda ordem. Neste caso

$$V(z) = \prod_{k=1}^{M} V_k(z) \tag{4.6}$$

onde M é o número de pares de pólos complexos conjugados, e

$$V_k(z) = \frac{(1 - 2|z_k|\cos(2\pi F_k T) + |z_k|^2)}{(1 - 2|z_k|\cos(2\pi F_k T)z^{-1} + |z_k|^2 z^{-2})} \tag{4.7}$$

o numerador de $V_k(z)$ é escolhido de forma que o produto tenha o mesmo ganho que o modelo de tubos sem perdas. Note-se que à frequência zero (z=1), $V_k(1)=1$.

$|z_k|$ é a distância do polo à origem e F_k a frequência do polo.

Apresenta-se o modelo em cascata na figura 4.6.

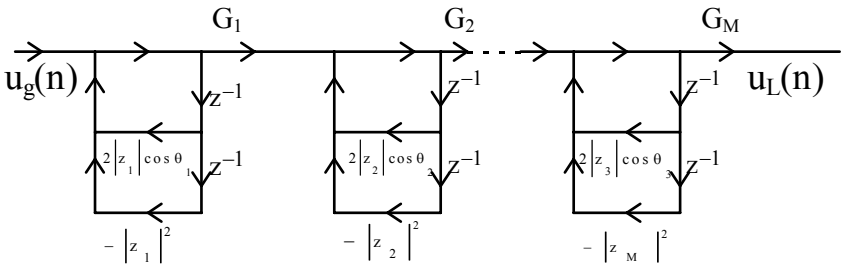


Figura 4.6 - Implementação em cascata de factores de 2ª ordem da função de transferencia do modelo só com pólos ($G_k = 1 - 2|z_k|\cos\theta_k + |z_k|^2$).

Este é um dos modelos para o trato vocal usado no decurso do trabalho de preparação da dissertação, no entanto existem outros modelos para o trato vocal como veremos mais adiante neste capitulo.

4.2.2.2 Efeito de Radiação

Até aqui foi considerada a função de transferência V(z) como a velocidade do volume de ar desde a fonte até aos lábios. Se se pretender um modelo para a pressão de ar nos lábios, então os efeitos de radiação devem ser incluídos [Rabiner 78].

O que desejamos é uma relação similar, no domínio z, da forma $P_L(z) = R(z)U_L(z)$, reportando à figura 4.7.

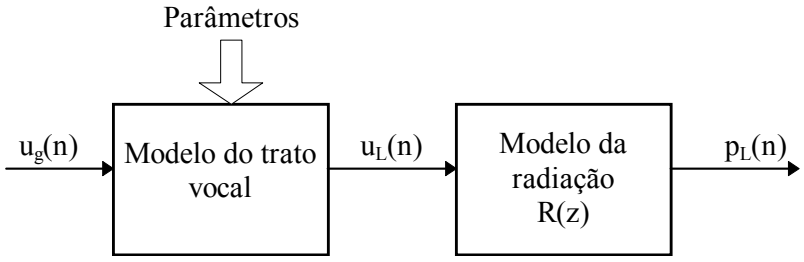


Figura 4.7 Modelo de engenharia incluindo os efeitos de radiação.

Segundo [Rabiner 78] a pressão está relacionada com a velocidade volumétrica do ar por uma operação de filtragem passa alto. De facto a baixas frequências pode-se dizer que a pressão é a derivada da velocidade volumétrica. Os mesmos autores chegam à razoável aproximação dos efeitos de radiação com a expressão

$$R(z) = R_0(1 - z^{-1}) \tag{4.8}$$

Este modelo de radiação pode ser considerado em cascata com o modelo do trato vocal como na figura 4.7.

V(z) pode ser implementado da forma mais conveniente e os parâmetros requeridos serão os apropriados para a configuração escolhida, isto é, funções de área para o modelo de tubos sem perdas ou frequência de formantes e larguras de banda para o modelo em cascata.

4.2.2.3 Excitação

Recapitulando que a maioria dos sons de fala podem ser classificados como vocalizados ou não vocalizados, já vimos que em termos gerais o que é necessário é uma fonte geradora de excitação capaz de produzir formas de onda de impulsos quase-periódicos e ruído aleatório. É ainda desejável a capacidade de combinar estes dois tipos de forma de onda para uma excitação mista. No caso da fala vocalizada a forma de onda de excitação deve ser algo parecida com a da figura 4.8.

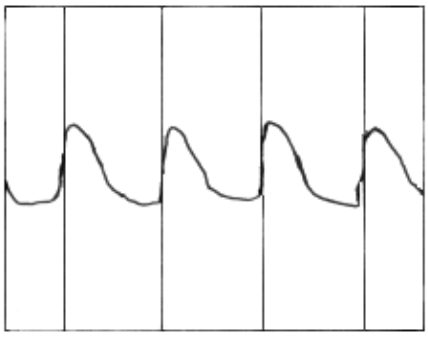


Figura 4.8 - Impulso glotal humano.

Uma forma conveniente de representar a geração da onda glotal é apresentada na figura 4.9.

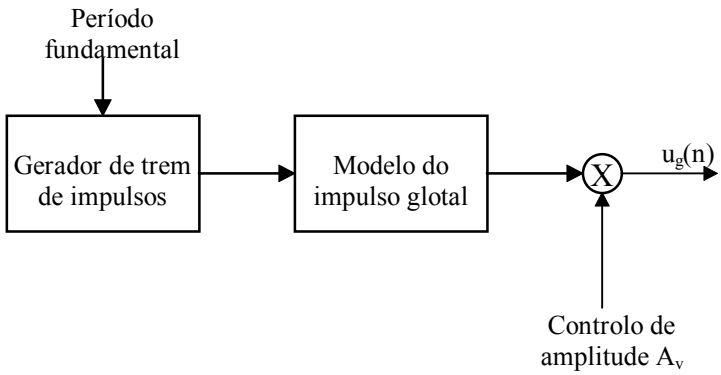


Figura 4.9 - Gerador do sinal de excitação para a fala vocalizada.

O gerador de trem de impulsos produz sequências de impulsos unitários espaçados pelo período fundamental desejado. Por sua vez esta sequência de impulsos excita um sistema linear com resposta impulsional g(n) desejada para a forma de onda glotal. O controlo do ganho Av, controla a amplitude de excitação.

Existem na literatura diversas funções de transferência para o impulso glotal das quais cito algumas:

$$G(z) = B\sum_{k=1}^{L_i}(1-\alpha_k z^{-1})\sum_{k=1}^{L_0}(1-\beta_k z^{-1}) \tag{4.9}$$

onde os zeros α_k e β_k podem estar ambos dentro ou fora do circulo unitário [Rabiner 78].

Para [Rosenberg 71] a resposta impulsional é da forma

$$\begin{aligned} g(n) &= \tfrac{1}{2}\left[1-\cos(\pi n/N_1)\right] && 0 \le n \le N_1 \\ &= \cos(\pi(n-N_1)/2N_2) && N_1 \le n \le N_1+N_2 \\ &= 0 && \text{outros valores de n} \end{aligned} \tag{4.10}$$

ainda por [Rabiner 78]

$$G(z) = \sum_{n=0}^{N_g} g(n)z^{-n} = B\prod_{n=1}^{N_g}(1-z_n z^{-1}) \tag{4.11}$$

e

$$G(z) = \frac{-ae\ln(a)z^{-1}}{(1-az^{-1})^2} \tag{4.12}$$

Por [Rowden 92].

Sendo contudo todos concordantes em que G(z) se comporta como um filtro de duração finita.

A figura 4.10 apresenta a forma de onda do impulso glotal sintético da expressão (4.12). A figura 4.11 apresenta o mesmo sinal no domínio das frequências, como esperado o efeito do impulso glotal no domínio das frequências é idêntico ao efeito de introdução de um filtro passa baixo.

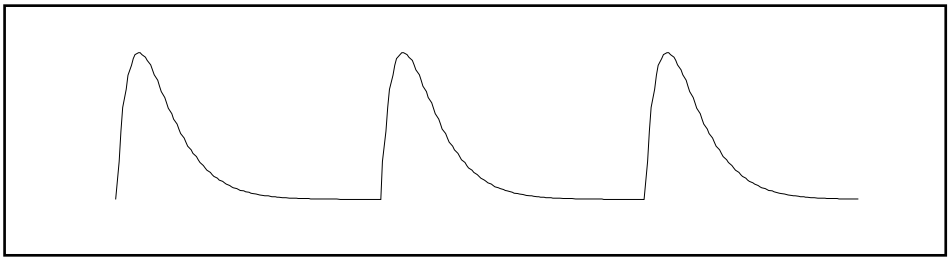


Figura 4.10 - Forma de onda do impulso glotal sintético $G(z) = \frac{-ae\ln(a)z^{-1}}{(1-az^{-1})^2}$ com a=0,90.

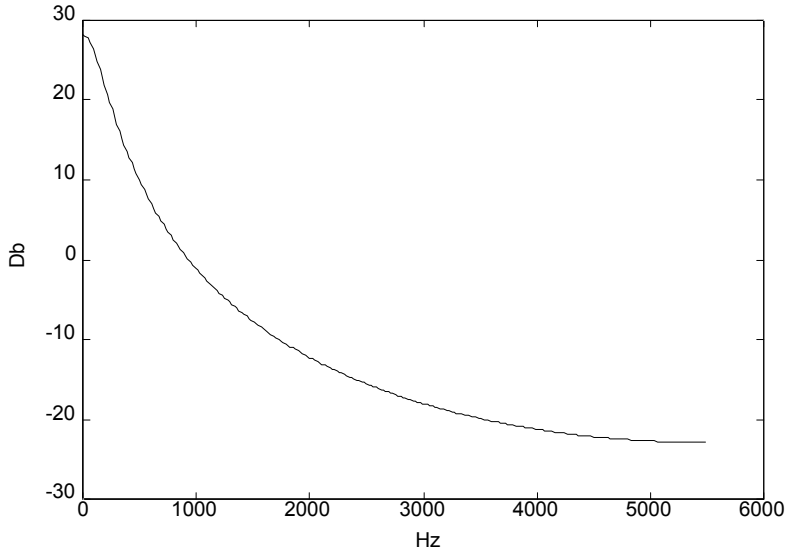


Figura 4.11 - Espectro do impulso glotal sintético $G(z) = \frac{-ae\ln(a)z^{-1}}{(1-az^{-1})^2}$ com a = 0,90.

Para sons não vocalizados o modelo de excitação é mais simples. Tudo o que é necessário é uma fonte de ruído aleatório e um parâmetro de ganho para controlar a intensidade da excitação não vocalizada. Para modelos discretos no tempo, um gerador de números aleatórios providencia uma fonte de ruído com espectro plano. A distribuição de probabilidade das amostras de ruído não parecem ser críticas segundo [Rabiner 78].

Para se obter uma excitação mista, basta um somador que some o ruído aleatório com o sinal periódico sendo o resultado a excitação mista.

4.2.2.4 O Modelo Completo

Juntando agora todas as partes obtemos o modelo da figura 4.12.

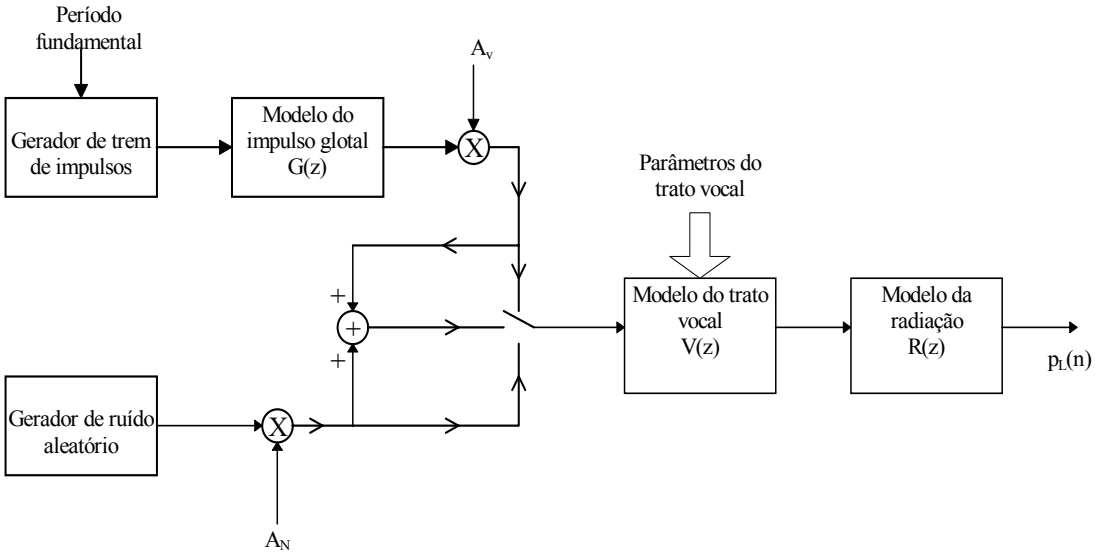


Figura 4.12 - Modelo genérico para a produção da fala.

Comutando entre o gerador de excitação vocalizada ou não vocalizada e excitação mista pode-se modelar a alteração do modo de excitação. O trato vocal pode ser modelado numa larga variedade de maneiras, algumas já foram apresentadas e outras ainda o serão. Em alguns casos torna-se conveniente combinar o modelo do impulso glotal com o modelo de radiação num único sistema. Veremos ainda que na análise por predição linear torna-se conveniente combinar as componentes do impulso glotal, radiação e trato vocal todas juntas e representa-las como uma única função de transferência só com pólos.

$$H(z)=G(z)V(z)R(z) \tag{4.13}$$

Com isto pretende-se dizer que o modelo da figura 4.12 é apenas uma representação genérica com muitas possibilidades de modificação.

Apresentam-se de seguida algumas deficiências do modelo adiantando que nenhuma delas limita seriamente a sua aplicabilidade.

Primeiro põe-se a questão da variação temporal dos parâmetros. Em sons contínuos como as vogais, os parâmetros variam lentamente e o modelo funciona perfeitamente bem. Com os sons de transição como as oclusivas, o modelo não é perfeito mas continua adequado. Chama-se a atenção para o facto das funções de transferência e das funções de resposta em frequência assumirem implicitamente a representação do sinal numa base de tempo curta. Isto é, assume-se que os parâmetros do modelo são constantes por um intervalo de tempo de 5 a 30 ms. A função de transferência V(z) serve realmente para definir a estrutura do modelo cujos parâmetros variam lentamente no tempo. A segunda limitação é a falta de zeros no modelo do trato vocal como requerido teoricamente para as nasais e fricativas. Isto torna-se definitivamente uma limitação para as nasais, contudo para as fricativas não é uma limitação severa.

Esta limitação pode ser ultrapassada impondo a dificuldade adicional no modelo de trato vocal de comportar um zero deixando assim de ser um modelo só com pólos.

Os modelos apresentados foram a base de trabalho na preparação desta dissertação. Devemos pensar neles de duas maneiras diferentes. Um ponto de vista é a análise da fala e o outro é a síntese da fala. Na análise da fala estamos interessados nas técnicas de estimação dos parâmetros do modelo, a partir dos sinais de fala natural, que se assumem como sendo a saída do modelo. Na síntese da fala o modelo é usado para criar sinais de fala sintética sendo este controlado por parâmetros estáveis. Estes dois pontos de vista aparecem muitas vezes interligados em problemas de substâncias diferentes.

4.2.3 Modelamento Sinusoidal da Fala

O modelamento sinusoidal da fala é uma técnica com raízes na análise de Fourier e na teoria da estimação. Tem vindo a desenvolver-se desde as décadas de 70 e 80 revestindo-se hoje de um interesse crescente. O objectivo destas técnicas é a representação do sinal de fala através da sobreposição de sinusóides com amplitudes , frequências e fases variáveis no tempo. O espectro de cada uma destas sinusóides está centrado em torno da frequência média da sinusóide. Então, em cada intervalo, o modelo sinusoidal pode ser entendido como uma representação do sinal de fala no domínio das frequências já que produz uma decomposição do sinal em bandas de frequência distintas.

Outra motivação para o uso do modelo sinusoidal é a existência de zonas extensas do sinal de fala produzidas com vozeamento, apresentando uma estrutura quase-periódica. Se o sinal de fala vozeada fosse periódico poderia ser representado, de uma forma exacta por um modelo sinusoidal estacionário, ou seja, pela sobreposição de sinusóides com amplitudes frequências e fases constantes ao longo do tempo.

Apesar do sinal de fala vozeada não ser exactamente periódico, a sua estrutura altera-se habitualmente de forma suficientemente lenta para que possa ainda ser bem aproximado pela soma de sinusóides, desde que as amplitudes frequências e fases destas variem lentamente ao longo do tempo.

O espectro do sinal de fala vozeada em intervalos de 20 a 40 ms é habitualmente composto por riscas espectrais aproximadamente centradas em múltiplos da frequência de vibração das cordas vocais. O objectivo do modelamento sinusoidal é a representação de cada risca espectral por meio de uma sinusóide.

Estas interpretações acerca do modelamento sinusoidal da fala vozeada estão na origem de duas classes de modelos sinusoidais: modelos no domínio da frequência e modelos no domínio do tempo [Marques 90], que não será aqui mais alongado por não estar no seguimento desta dissertação.

Por último, ainda segundo [Marques 90], os modelos sinusoidais estão especialmente vocacionados para o modelamento de sons produzidos por vibração das cordas vocais, estando menos adaptados ao modelamento de sons não vozeados e sons de transição.

4.3 Parâmetros Para os Modelos dos Sinais de Fala

Como vimos os modelos são usados no caso da análise para extrair um conjunto de parâmetros que são depois usados na síntese para reproduzir o mesmo sinal de fala.

Este conjunto de parâmetros pode ser diferente de acordo com o modelo preciso que é usado. No entanto em qualquer dos casos o número de bits usado para guardar ou enviar a informação contida nos parâmetros é certamente bastante menor que o número de bits usado para o sinal original.

Recorda-se que cada conjunto de parâmetros se repete em cada segmento de duração entre 5 a 30 ms.

Para o modelo genérico discutido anteriormente é indispensável que os parâmetros contenham informação à cerca do modo de excitação, a amplitude do sinal e no caso de ser um sinal vocalizado é necessária a informação da duração do período de "pitch". No modelo do trato vocal é onde a diversidade de implementação pode ser mais variada, levando também a uma maior diversidade do conjunto de parâmetros para o modelo.

A seguir apresentam-se alguns conjuntos de parâmetros que parametrizam o modelo do trato vocal.

4.3.1 Parametrização por Formantes

Quando se usam formantes para parametrizar o trato vocal tudo o que é necessário guardar são as frequência destes e as respectivas larguras de banda.

Uma questão importante que se coloca é o número de formantes a utilizar. Assim, sabe-se que o primeiro e segundo formantes são de extrema importância, pois os valores das suas frequências chegam para caracterizar uma vogal [Martins 92] [Mateus 90]. Contudo apenas dois formantes não chegam para modelizar o trato vocal, então há alguns autores que usam modelos de três formantes [Schafer 70], [Rabiner 78] com uma qualidade razoável, outros usam quatro formantes (MULTIVOX[i]) já com uma qualidade boa, e, outros ainda, usam cinco formantes (PHONOVOX[ii]) com uma qualidade também um pouco melhor. Há ainda alguns autores que referem o uso de um modelo com quatro formantes mais um polo e um zero para adequar melhor o modelo para os sons nasalizados e algumas fricativas.

Para os sons não vocalizados [Rabiner 78], propõe um modelo para o trato vocal com um pólo e um zero com a função de transferência

$$V(z) = \frac{(1 - 2e^{-\beta T}\cos(2\pi F_P T) + e^{-2\beta T})(1 - 2e^{-\beta T}\cos(2\pi F_Z T)z^{-1} + e^{-2\beta T}z^{-2})}{(1 - 2e^{-\beta T}\cos(2\pi F_P T)z^{-1} + e^{-2\beta T}z^{-2})(1 - 2e^{-\beta T}\cos(2\pi F_Z T) + e^{-2\beta T})} \tag{4.14}$$

em que a frequência do polo F_P é escolhida como a frequência a que ocorre o maior pico do espectro logarítmico alisado a baixo dos 1000 Hz, e a frequência do zero F_Z satisfazendo a formula empírica

$$F_Z = (0.0065F_P + 4.5 - \Delta)(0.014F_P + 28)$$
$$\text{onde} \qquad \Delta = 20\log_{10}\left|H(e^{j2\pi F_P T})\right| - 20\log_{10}\left|H(e^{j0})\right| \tag{4.15) e (4.16}$$

[i] MULTIVOX - Sistema de conversão texto-fala multilíngua, de origem Húngara, incorporando um sintetizador de fala.
[ii] PHONOVOX - Sistema de ajuda ao desenvolvimento de uma língua no conversor texto fala MULTIVOX, usando um sintetizador de fala de 5 formantes.

Quando é usado o modelo de predição linear por vezes os parâmetros usados não são os formantes mas sim os coeficientes do polinómio do denominador e do numerador, quando for caso disso, da função de transferência. A diferença reside apenas numa questão algébrica já que a partir dos coeficientes do polinómio se chega facilmente aos valores das frequências formantes e respectivas larguras de banda e vice versa. Por outro lado a quantidade de informação para um e outro casos é idêntica já que para cada formante e respectiva largura de banda corresponde um par de pólos e por sua vez dois coeficientes.

4.4 Técnicas de Análise Para Obtenção dos Parâmetros dos Sinais de Fala Baseadas no Modelo de Formantes

Baseado no modelo genérico de produção de fala e na parametrização por formantes do trato vocal, que habitualmente se denomina por modelo de formantes, apresentam-se agora diferentes técnicas usadas na análise, extracção, dos parâmetros do sinal de fala.

4.4.1 Análise por Síntese

Esta técnica tem vindo a ser muito usada desde a década de 60 para e estimação da frequência dos formantes.

A ideia básica da análise por síntese é a seguinte: Assume-se que inicialmente se começa com uma forma de onda de fala ou qualquer outra representação do sinal de fala. Depois é assumido um modelo de produção de fala. Esse modelo (modelo de trato vocal ou de engenharia) comporta um grupo de parâmetros que podem ser ajustados para produzir diferentes sons de fala. A partir do modelo pode-se derivar uma representação do modelo com a mesma forma de representação do sinal de fala. Então variando os parâmetros do modelo de uma forma sistemática pode-se por exemplo, procurar encontrar os valores para o grupo de parâmetros que aproximam o modelo com um erro mínimo, do sinal de fala. Quando essa aproximação é encontrada, assumem-se os valores dos parâmetros do modelo como os valores dos parâmetros do sinal de fala.

Caberá aqui referir que as regras de concatenação de fonemas para o português do conversor texto-fala MULTIVOX foram desenvolvidas durante o trabalho de preparação desta dissertação e como será discutido no capitulo seguinte o processo de criação/modificação de um grupo de ABU's[i] com os parâmetros do modelo utilizado e a sua concatenação para produzir sons pretendidos de fonemas e conjunto de dois fonemas (difones) foi baseado na técnica de análise por síntese. Procedeu-se a um ajuste sistemático dos parâmetros das ABU's por forma a minimizar a diferença ouvida entre os sons produzidos pelo modelo e os sons conhecidos de um fonema. Ainda neste mesmo trabalho, com o mesmo modelo, procedeu-se a um ajuste dos parâmetros com recurso a um espectrógrafo minimizando a diferença do espectrograma dos sons produzidos pelo modelo para uma determinada sequência de fonemas ou palavra com o espectrograma dos sons da mesma sequência de fonemas ou palavra produzidas por uma pessoa.

[i] ABU - Acoustic Building Unit - é uma unidade acústica usada no conversor texto-fala MULTIVOX como um segmento de sinal de fala com uma duração variável entre 10 e 50 ms. No conversor MULTIVOX existe um grupo de 255 ABU's formando como que uma base de sons mínimos . Todos os sons de fala são produzidos pela concatenação criteriosa de algumas destas ABU's.

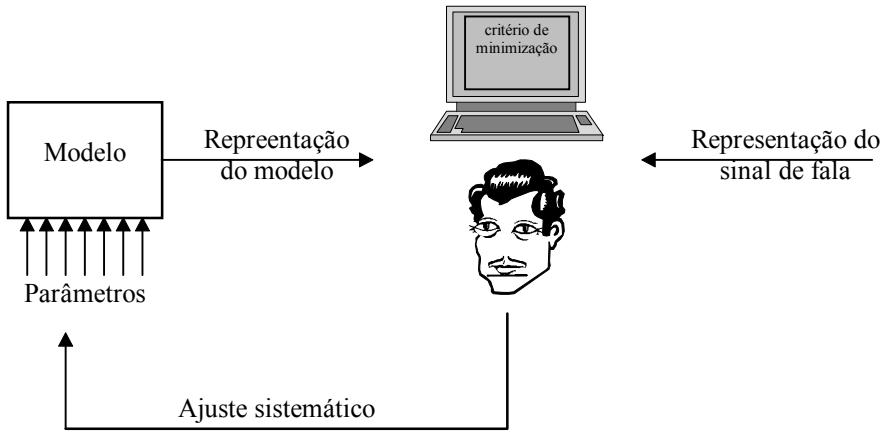


Figura 4.13 - Sistema de análise por síntese.

Nestes dois exemplos a comparação das representações do modelo e do sinal é realizada por um ser humano, assim como o ajuste sistemático também é feito manualmente e obedecendo a critérios muito pessoais. No entanto outros sistemas de análise por síntese têm critérios analíticos precisos das distancias entre as representações do modelo e do sinal bem como sequências conhecidas para proceder a um ajuste sistemático. Então o ciclo, comparação das representações do modelo e do sinal, ajuste dos parâmetros, é feito automaticamente através de um computador com pouca ou nenhuma intervenção humana, resultando assim num processo muito mais rápido que o anterior com a total intervenção humana.

Um sistema de análise por síntese pode ser representado pela figura 4.13.

4.4.2 Análise Síncrona com o Período Fundamental

A função de transferência do modelo de engenharia estudado anteriormente para a produção de fala vozeada é o seguinte:

$$H(z) = R(z)V(z)G(z) \tag{4.17}$$

Onde R(z) representa o efeito de radiação, que basicamente aparece como uma diferenciação às baixas frequências, e que adequadamente modelado para este propósito por uma simples equação às diferenças de 1ª ordem, sendo a sua representação na transformada Z

$$R(z) = 1 - z^{-1} \tag{4.18}$$

V(z) é a função de transferencia do trato vocal na forma

$$V(z) = \frac{A}{\prod_{k=1}^{M} (1 - 2e^{-\sigma_k T}\cos(2\pi F_k T)z^{-1} + e^{-2\sigma_k T}z^{-2})} \tag{4.19}$$

e G(z) a função de transferência do impulso glotal g(n) que como já se viu pode assumir várias formas.

O método de análise síncrono com o período fundamental ("Pitch Sincronous Analysis") usado para a fala vozeada é baseado na detecção do início e fim do impulso glotal para separar convenientemente cada período do sinal de fala e analisa-los separadamente reduzindo ou eliminando assim o efeito do impulso glotal.

Elliot Pinson [Pinson 63]] desenvolveu uma técnica de análise síncrona com o "pitch" com recurso à análise por síntese para estimação dos formantes e larguras de banda que se baseia na aproximação da função

$$f(t) = \sum_{i=1}^{N} e^{-\mu B_i t}(a_i \cos(2\pi F_i t) + b_i \mathrm{sen}(2\pi F_i t)) + a_0 \qquad (4.20)$$

à função p(t) que é o sinal adquirido de fala vocalizada durante o período em que a glote se encontra fechada, não havendo portanto fonte excitadora, apenas a ressonância do trato vocal, por minimização da função erro

$$E = \sum_{M=0}^{k-1} W_M^2 (P_M - F_M)^2 \qquad (4.21)$$

Este método permite a obtenção dos N primeiros formantes e bandas passantes com uma maior precisão que outros métodos devido ao seu modelo se basear no estudo do sinal de fala vocalizada no momento em que a glote se encontra fechada (não há passagem de fluxo de ar).

As dificuldades de implementação deste método são a sua difícil obtenção dos parâmetros, conhecimento exacto do instante de abertura e fecho da glote (podendo obrigar ao uso de um electroglotógrafo), a exigência do falante ter um "pitch" baixo, <140 Hz, para que haja pontos amostrados em quantidade suficiente para realizar a aproximação da função f(t) a p(t), e o conhecimento prévio de alguns valores iniciais de alguns parâmetros de convergência que resultem num processo iterativo convergente.

4.4.3 Análise Cepstral

A análise cepstral é uma das maneiras de separar as características do filtro do trato vocal da sequência de excitação no modelo até agora usado em que se assume que o sinal de fala, s(t) é composto por um sinal de excitação e(t) aplicado ao filtro do trato vocal, com uma resposta impulsional v(t), de um ponto de vista no domínio temporal como apresentado na figura 4.14.

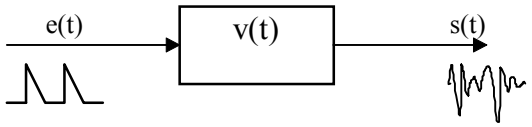


Figura 4.14 - Modelo simples de produção da fala no domínio temporal.

s(t) é a convolução de e(t) com v(t)

$$s(t) = e(t) \otimes v(t) \tag{4.22}$$

que no domínio da frequência fica simplesmente

$$S(w)=E(w)V(w) \tag{4.23}$$

em que S(w), E(w) e V(w) são as transformadas de Fourier das funções contínuas no tempo s(t), e(t) e v(t) ou as transformadas discretas de Fourier das sequências de amostras temporais s(n), e(n) e v(n).

O sinal de excitação E(w) e o filtro do trato vocal aparecem combinados multiplicativamente, o que torna difícil a sua separação. Contudo, tomando o logaritmo de S(w) a excitação e a função do trato vocal tornam-se aditivas.

$$\log[S(w)] = \log[E(w)] + \log[V(w)] \tag{4.24}$$

A propriedade aditiva do espectro logarítmico continua-se a verificar quando lhe for aplicada a Transformada de Fourier, sendo o resultado dessa operação chamado função cepstral ou cepstro. Sumariamente, o processo de estimação do cepstro apresenta-se no diagrama de fluxo da figura 4.15. As suas propriedades e as variantes cepstro complexo e cepstro de potência são explicadas mais detalhadamente no capitulo 6.

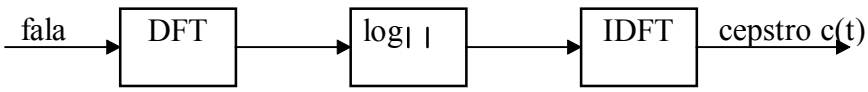


Figura 4.15 - Diagrama de fluxo da estimação do cepstro.

A forte componente periódica no espectro logarítmico de um sinal de fala vocalizada num intervalo de frequências equivalente ao inverso do período fundamental T, aparece no cepstro como um pico.

A variável independente, no eixo horizontal, da função cepstral tem dimensões temporais e o nome de quefrências. Para a fala vocalizada pode ser feita uma clara distinção entre a componente de excitação e a contribuição do trato vocal que aparece como um amaranhado de componentes aos baixos valores de quefrência afastado da componente do período fundamental que aparece a valores mais altos de quefrência.

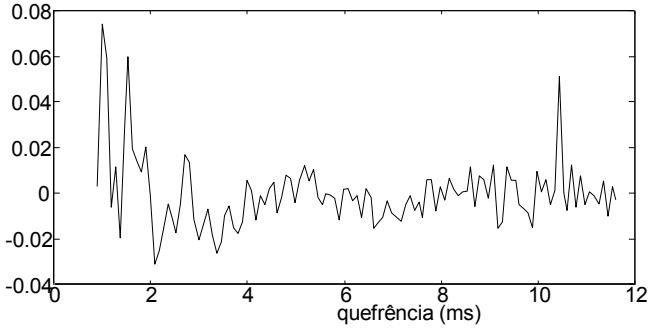


Figura 4.16 - Cepstro de um segmento de fala da vogal [a].

Na figura 4.16 apresenta-se o cepstro de um segmento de fala vocalizada. É visível o pico correspondente ao período fundamental próximo da quefrência de 10 ms, correspondendo à frequência fundamental de 100 Hz, separado das componentes do trato vocal às baixas quefrências. Nesta figura são apresentadas apenas as componentes do cepstro superiores a sensivelmente 1 ms, pois as componentes de mais baixas quefrências têm valores comparativamente muito superiores aos restantes e a sua apresentação não deixaria claro o pico correspondente à frequência fundamental.

A função de transferência do trato vocal e a função de excitação da fala aparecem em partes separadas da escala de quefrências, pelo que podem ser separadas as duas funções, ou removida uma delas por um processo de "lifteragem". O cepstro é constituído por um conjunto de valores cepstrais discretos, que são o conjunto de valores de saída do processo final de IDFT (Inverse Discrete Fourier Transform). Então podemos aplicar um "lifter" com uma função de janela rectangular. Contudo, a transformada de Fourier da janela rectangular indica que tal processo gera lobos laterais indesejados às componentes requeridas no domínio das frequências logarítmicas [Rowden 92], pelo que é requerida uma função de uma janela mais gradual. Assim aplicando esta função de "lifteragem" ao sinal do cepstro e tomando a transformada de Fourier Discreta Inversa (IDFT) do sinal resultante, obtém-se uma versão alisada do espectro logarítmico do filtro do trato vocal.

Este alisamento cepstral pode ser usado para estimar a curva do envelope espectral como na figura 4.17.

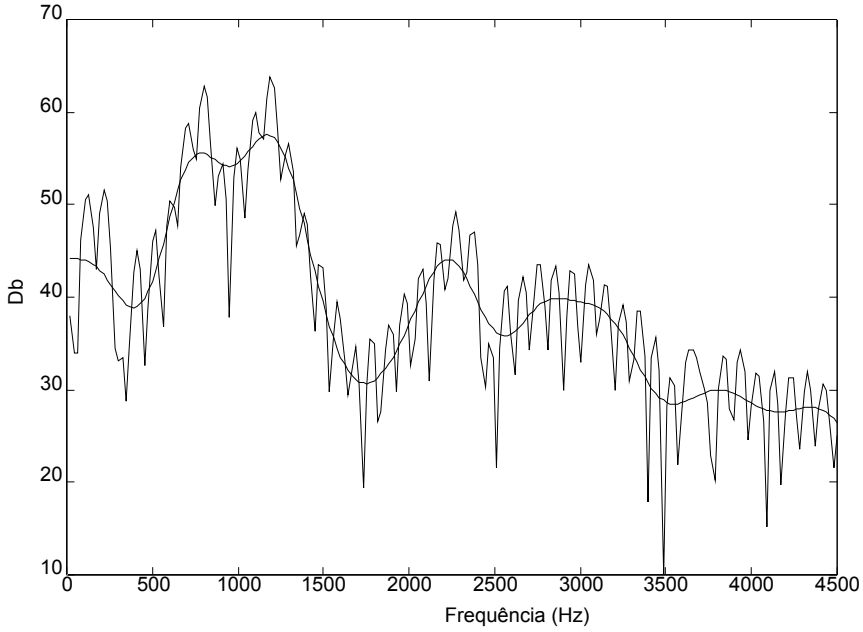


Figura 4.17 - Espectro e envelope espectral (pelo método do cepstro) de um segmento de fala da vogal [a].

Uma vez obtido o envelope espectral, com alguns cuidados como veremos no capítulo 7, extrai-se as frequências, larguras de banda e amplitude das formantes.

4.4.4 Análise por LPC's

Uma das mais poderosas técnicas de análise da fala é o método de análise por predição linear. Este método começou por ser a técnica predominante na estimação dos parâmetros básicos da fala, isto é, frequência fundamental, formantes, larguras de banda, funções de área do trato vocal e para a representação da fala a uma baixa taxa de transmissão ou armazenamento. A importância deste método está ligada tanto à habilidade para estimar os parâmetros da fala com uma extrema precisão como à sua computação relativamente rápida.

A ideia básica para a análise por predição linear é que qualquer amostra do sinal de fala pode ser aproximado por uma combinação linear das amostras anteriores. A minimização da soma das diferenças quadradas (num intervalo finito) entre a amostra de fala actual e a predita linearmente, leva a que um único conjunto de coeficientes de predição possam ser determinados. Os coeficientes de predição são o peso dos coeficientes usados na combinação linear.

A filosofia da predição linear está intimamente relacionada com o modelo de fala básico já discutido em que a fala pode ser modelada como a saída de um sistema linear variante no tempo, excitado por impulsos quase periódicos, durante a fala vocalizada, ou ruído aleatório na fala não vocalizada. O método de predição linear é robusto, fiável e preciso para a estimação dos parâmetros que caracterizam o sistema linear variante no tempo.

Neste sub-capítulo será apresentado um ponto de vista genérico da predição linear de um modelo só com pólos e o modo como a ideia básica da predição linear lida com

um grupo de técnicas de análise que podem ser usadas na estimação dos parâmetros do modelo de fala.

A ideia de predição linear tem sido usada nas áreas de controlo, teoria da informação com os nomes de sistemas de estimação e sistemas de identificação.

A importância da predição linear está ligada à precisão com que o modelo básico se aplica à fala.

4.4.4.1 Princípios Básicos da Análise por Predição Linear

A forma particular do modelo de terminais análogos apropriada para a discussão da análise por predição linear é apresentado na figura 4.18.

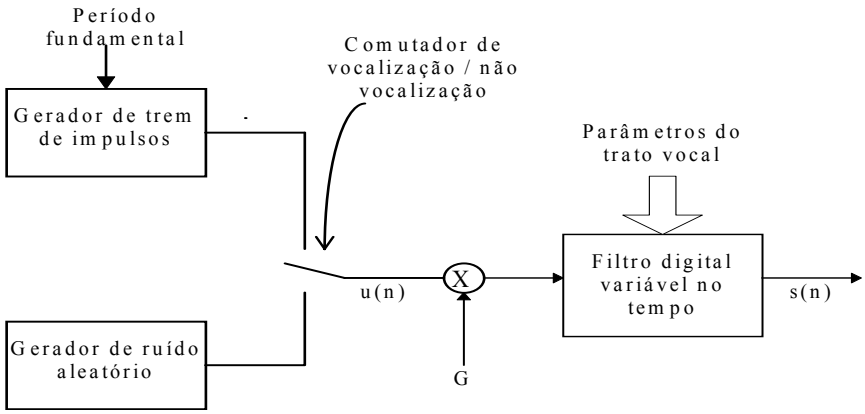


Figura 4.18 - Diagrama de blocos do modelo simplificado de produção da fala.

Neste caso a combinação dos efeitos de radiação, trato vocal e excitação glotal estão representados por um filtro digital variante no tempo cuja função do sistema em estado estacionário é da forma

$$H(z) = \frac{S(z)}{U(z)} = \frac{G}{1 - \sum_{k=1}^{p} a_k z^{-k}} \tag{4.25}$$

Este sistema é excitado por um trem de impulsos para a fala vocalizada ou por uma sequência de ruído aleatório para a fala não vocalizada. Então, os parâmetros deste modelo são: classificação do modo de excitação, período fundamental para a fala vocalizada, parâmetro ganho G, e os coeficientes $\{a_k\}$ do filtro digital. Todos estes parâmetros variam lentamente com o tempo.

O modo de excitação e o período fundamental podem ser estimados usando métodos discutidos nos capítulos 6 e 7 ou por métodos mais elaborados de análise por predição linear.

Este modelo simplificado é um modelo só com pólos que como se viu é uma representação natural para sons não nasais e vocalizados, mas para sons nasalizados e fricativos a teoria acústica refere modelos para o trato vocal com pólos e zeros. Contudo [Rabiner 78] prova que se a ordem p do modelo for suficientemente elevada, o modelo só com pólos provê uma boa representação para a maioria dos sons de fala.

A grande vantagem deste modelo, só com pólos, é que o parâmetro de ganho G, e os coeficientes do filtro $\{a_k\}$ podem ser estimados de uma maneira directa e computacionalmente eficiente pelo método de análise por predição linear.

Para o sistema da figura 4.18 as amostras de fala s(n) estão relacionadas com a excitação u(n) pela simples equação às diferenças.

$$s(n) = \sum_{k=1}^{p} a_k s(n-k) + Gu(n) \tag{4.26}$$

Um sistema de predição linear com os coeficientes de predição $\{\alpha_k\}$ tem a saída definida por

$$\tilde{s}(n) = \sum_{k=1}^{p} \alpha_k s(n-k) \tag{4.27}$$

A função de predição linear de ordem p do sistema é caracterizada pelo polinómio

$$P(z) = \sum_{k=1}^{p} \alpha_k z^{-k} \tag{4.28}$$

O erro de predição e(n) é definido como

$$e(n) = s(n) - \tilde{s}(n) = s(n) - \sum_{k=1}^{p} \alpha_k s(n-k) \tag{4.29}$$

Da equação anterior mostra-se que a sequência do erro de predição é a saída de um sistema cuja função de transferencia é

$$A(z) = 1 - \sum_{k=1}^{p} \alpha_k z^{-k} \tag{4.30}$$

É visível por comparação das equações (4.26) e (4.29) que se o sinal de fala obedece exactamente à equação do modelo (4.26), e se $\alpha_k = a_k$, então e(n) = Gu(n). Então o filtro de predição linear, A(z), será um filtro inverso para o sistema, H(z) da equação (4.25), isto é

$$H(z) = \frac{G}{A(z)} \tag{4.31}$$

O problema básico da análise por predição linear é a determinação do conjunto de coeficientes de predição $\{\alpha_k\}$ directamente do sinal de fala de forma a obter uma boa estimativa das propriedades espectrais do sinal de fala com recurso ao uso da equação (4.31). Devido à natureza de variação temporal dos sinais de fala os coeficientes de predição devem ser estimados em segmentos curtos do sinal de fala. A aproximação

básica é encontrar o conjunto de coeficientes de predição que minimizem o erro quadrático médio de predição num curto segmento da forma de onda da fala. Os parâmetros resultantes são assumidos como os parâmetros da função do sistema, H(z), no modelo de produção da fala.

Que estas aproximações levem a resultados bons não é imediatamente óbvio, mas podem ser justificados de diferentes maneiras. Primeiro, relembrando que se $\alpha_k = a_k$, então e(n) = Gu(n). Para a fala vocalizada isto quer dizer que e(n) consistirá num trem de impulsos, isto é, e(n) será pequeno a maior parte do tempo. Então trata-se de encontrar os α_k's que minimizem o erro de predição respeitando a consistência desta observação. A segunda motivação para esta aproximação resulta do facto de que se o sinal é gerado pela equação (4.26) com coeficientes não variáveis no tempo e excitado quer por um único impulso quer por ruído aleatório estacionário na entrada, então pode-se provar que os coeficientes de predição resultantes da minimização do erro quadrático médio de predição são idênticos aos da equação (4.26). Em terceiro lugar, a justificação muito pragmática para o uso do erro quadrático médio de predição como base para estimação dos parâmetros do modelo resulta desta aproximação conduzir a um conjunto de equações lineares que podem ser resolvidas de uma forma eficiente para obter os parâmetros de predição. Por último e mais importante é que os parâmetros resultantes fazem uma compressão muito útil e precisa da representação do sinal de fala como já foi discutido.

O erro de predição médio num intervalo curto é definido como

$$E_n = \sum_m e_n^2(m) \tag{4.32}$$

$$= \sum_m \left(s_n(m) - \tilde{s}_n(m)\right)^2 \tag{4.33}$$

$$= \sum_m \left[s_n(m) - \sum_{k=1}^{p} \alpha_k s_n(m-k) \right]^2 \tag{4.34}$$

em que $s_n(m)$ é um segmento de fala seleccionado na vizinhança da amostra n, isto é

$$s_n(m) = s(m+n) \tag{4.35}$$

Os limites dos somatórios das equações (4.32) a (4.34) não são temporariamente especificados, mas desde que se deseje desenvolver a técnica de análise num período de tempo curto o somatório será sempre limitado a um intervalo finito. De notar também que para se determinar a média se deveria dividir pelo intervalo de tempo do segmento do sinal de fala. Contudo como se trata de uma constante, torna-se irrelevante para o conjunto de equações lineares que se obtêm, sendo portanto omitido.

Podem-se encontrar os valores para α_k que minimizam E_n na equação (4.34) impondo $\partial E_n / \partial \alpha_i = 0$, i=1,2...,p, resultando as equações

$$\sum_m s_n(m-i)s_n(m) = \sum_{k=1}^{p} \hat{\alpha}_k \sum_m s_n(m-i)s_n(m-k) \qquad 1 \le i \le p \qquad (4.36)$$

em que $\hat{\alpha}_k$ são os valores de α_k que minimizam E_n [Makhoul 75].

Se se definir

$$\phi_n(i,k) = \sum_m s_n(m-i)s_n(m-k) \qquad (4.37)$$

Então pode-se escrever mais compactamente o sistema de equações (4.36)

$$\sum_{k=1}^{p} \alpha_k \phi_n(i,k) = \phi_n(i,0) \qquad i = 1,2,...,p \qquad (4.38)$$

O conjunto de p equações com p incógnitas pode ser resolvido de uma forma eficiente para os coeficientes de predição desconhecidos $\{\alpha_k\}$ que minimizam o erro quadrático médio de predição para o segmento $s_n(m)$[i]. Usando as equações (4.34) e (4.36) o erro quadrático médio de predição mínimo resulta

$$E_n = \sum_m s_n^2(m) - \sum_{k=1}^{p} \alpha_k \sum_m s_n(m)s_n(m-k) \qquad (4.39)$$

e usando a equação (4.38) pode-se expressar E_n como

$$E_n = \phi_n(0,0) - \sum_{k=1}^{p} \alpha_k \phi_n(0,k) \qquad (4.40)$$

Então o erro mínimo total consiste numa componente fixa e numa componente que depende dos coeficientes de predição.

Para determinar os coeficientes de predição óptimos, deve-se determinar primeiro os valores de $\phi_n(i,k)$ para $1 \le i \le p$ e $0 \le k \le p$. Depois é necessário resolver as equações (4.38) para obter os α_k's. Assim, em principio a análise por predição linear é directa. Contudo, os detalhes de computação de $\phi_n(i,k)$ e da subsequente solução de equações pode parecer algo intricada pelo que se apresentam alguns métodos de resolução.

Até aqui não se indicou explicitamente os limites dos somatórios das equações (4.32) a (4.34) e (4.36); contudo deve-se realçar que os limites dos somatórios nas equações (4.36) são idênticos aos limites assumidos para o erro quadrático médio de predição nas equações (4.32) a (4.34). Como se constatou, se se pretender desenvolver um procedimento de análise em intervalos curtos de tempo, os limites devem estar dentro

[i] Está claro que os α_k's são função de n (o índex temporal correspondente ao instante a que são estimados) contudo esta dependência não aparece explicita. É vantajoso deixar o índex n em E_n, s_n e $\phi_n(i,k)$ quando não é passível de qualquer confusão.

de um intervalo finito. Existem duas aproximações básicas para esta questão, emergindo dois métodos de análise de predição linear que se apresentam de seguida.

4.4.4.2 Método da Autocorrelação

Uma tentativa de determinação dos limites das equações (4.32) a (4.34) e (4.36) é assumir que a forma de onda do segmento, $s_n(m)$, é zero fora do intervalo $0 \le m \le N-1$. Isto pode ser convenientemente expresso como

$$s_n(m) = s(m+n)w(m) \tag{4.41}$$

onde w(m) é uma janela de comprimento finito (exp.: janela de Hamming) com zeros fora do intervalo $0 \le m \le N-1$.

O efeito desta assunção na questão dos limites do somatório da expressão de E_n pode ser considerado atendendo à equação (4.29). Claramente, se $s_n(m)$ for diferente de zero apenas para $0 \le m \le N-1$, então o correspondente erro de predição $e_n(m)$, para um preditor de ordem p será diferente de zero dentro do intervalo $0 \le m \le N-1+p$. Então, neste caso E_n é devidamente expresso como

$$E_n = \sum_{m=0}^{N+p-1} e_n^2(m) \tag{4.42}$$

Os limites da expressão para $\phi_n(i,k)$ na equação (4.37) são idênticos aos da equação (4.42). Mas, como $s_n(m)$ é nulo fora do intervalo $0 \le m \le N-1$ é fácil mostrar que

$$\phi_n(i,k) = \sum_{m=0}^{N+p-1} s_n(m-i)s_n(m-k) \qquad \begin{matrix} 1 \le i \le p \\ 0 \le k \le p \end{matrix} \tag{4.43}$$

pode ser expresso como

$$\phi_n(i,k) = \sum_{m=0}^{N-1-(i-k)} s_n(m)s_n(m+i-k) \qquad \begin{matrix} 1 \le i \le p \\ 0 \le k \le p \end{matrix} \tag{4.44}$$

Além disso pode-se mostrar que neste caso $\phi_n(i,k)$ é idêntico à função de autocorrelação de curta duração avaliada para (i-k), Isto é

$$\phi_n(i,k) = R_n(i-k) \tag{4.45}$$

onde

$$R_n(k) = \sum_{m=0}^{N-1-k} s_n(m)s_n(m+k) \tag{4.46}$$

Como $R_n(k)$ é uma função par segue-se que

$$\phi_n(i,k) = R_n(|i-k|) \qquad \begin{matrix} i = 1,2,...,p \\ k = 0,1,...,p \end{matrix} \tag{4.47}$$

Então as equações (4.38) podem ser expressas por

$$\sum_{k=1}^{p} \alpha_k R_n(|i-k|) = R_n(i) \qquad 1 \le i \le p \tag{4.48}$$

Da mesma forma o erro quadrático médio predito mínimo da equação (4.40) toma a forma

$$E_n = R_n(0) - \sum_{k=1}^{p} \alpha_k R_n(k) \tag{4.49}$$

O conjunto de equações em (4.48) pode ser expresso de forma matricial como

$$\begin{bmatrix} R_n(0) & R_n(1) & R_n(2) & ... & R_n(p-1) \\ R_n(1) & R_n(0) & R_n(1) & ... & R_n(p-2) \\ R_n(2) & R_n(1) & R_n(0) & ... & R_n(p-3) \\ ... & ... & ... & ... & ... \\ ... & ... & ... & ... & ... \\ R_n(p-1) & R_n(p-2) & R_n(p-3) & ... & R_n(0) \end{bmatrix} \times \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ ... \\ ... \\ \alpha_p \end{bmatrix} = \begin{bmatrix} R_n(1) \\ R_n(2) \\ R_n(3) \\ ... \\ ... \\ R_n(p) \end{bmatrix} \tag{4.50}$$

A matriz de autocorrelação de dimensão $p \times p$ é uma matriz Toeplitz; isto é, trata-se de uma matriz simétrica e todos os elementos ao longo de qualquer diagonal são iguais. Esta propriedade especial pode ser explorada para obter algoritmos eficientes para a solução da equação (4.48).

A partir do sistema de equações e da matriz de Toeplitz há vários métodos para determinar os coeficientes α_k's. Assim, as soluções recursivas de Durbin [Rabiner 78], é dos mais eficientes e os algoritmos de Levinson e Robinson são os mais populares.

No capitulo 7 foram usados métodos directos para obter os coeficientes do sistema a partir da matriz de Toeplitz e do sistema de equações.

4.4.4.3 Método da Covariância

O segundo método para definir o segmento de fala $s_n(m)$ e os limites do somatório consiste em fixar o intervalo em que o erro quadrático médio é calculado e então considerar o efeito do calculo de $\phi_n(i,k)$. Isto é, se se definir

$$E_n = \sum_{m=0}^{N-1} e_n^2(m) \tag{4.51}$$

então $\phi_n(i,k)$ virá

$$\phi_n(i,k) = \sum_{m=0}^{N-1} s_n(m-i)s_n(m-k) \qquad \begin{matrix} 1 \le i \le p \\ 0 \le k \le p \end{matrix} \tag{4.52}$$

Neste caso se for trocado o índex do somatório também se pode expressar $\phi_n(i,k)$ como

$$\phi_n(i,k) = \sum_{m=-1}^{N-i-1} s_n(m)s_n(m+i-k) \qquad \begin{matrix} 1 \le i \le p \\ 0 \le k \le p \end{matrix} \tag{4.53}$$

ou

$$\phi_n(i,k) = \sum_{m=-k}^{N-k-1} s_n(m)s_n(m+k-i) \qquad \begin{matrix} 1 \le i \le p \\ 0 \le k \le p \end{matrix} \tag{4.54}$$

Apesar da equação parecer muito similar à equação (4.44), os limites do somatório não são os mesmos. A equação (4.54) usa valores de $s_n(m)$ fora do intervalo $0 \le m \le N-1$. Então para determinar $\phi_n(i,k)$ para todos os valores requeridos de i e k é necessário usar valores de $s_n(m)$ no intervalo $-p \le m \le N-1$. Se se pretender ser consistente com os limites em E_n na equação (4.51) tem que se dispor dos valores requeridos. Neste caso não faz sentido anular o segmento de fala nos extremos como no método da autocorrelação pois os valores necessários encontram-se fora do intervalo $0 \le m \le N-1$. Este tratamento leva a uma função que não é a verdadeira função de autocorrelação, mas sim, a correlação cruzada entre dois segmentos de comprimento finito muito similares, mas não idênticos do sinal de fala. Apesar das diferenças entre as equações (4.53) a (4.54) e as equações (4.44) parecerem ser de menor detalhe computacional, o conjunto de equações

$$\sum_{k=1}^{p} \alpha_k \phi_n(i,k) = \phi_n(i,0) \qquad i = 1,2,...,p \tag{4.55}$$

tem propriedades significativamente diferentes que afectam fortemente o método de solução e as propriedades do preditor óptimo resultante. Estas equações aparecem na forma matricial

$$\begin{bmatrix} \phi_n(1,1) & \phi_n(1,2) & \phi_n(1,3) & \dots & \phi_n(1,p) \\ \phi_n(2,1) & \phi_n(2,2) & \phi_n(2,3) & \dots & \phi_n(2,p) \\ \phi_n(3,1) & \phi_n(3,2) & \phi_n(3,3) & \dots & \phi_n(3,p) \\ \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots \\ \phi_n(p,1) & \phi_n(p,2) & \phi_n(p,3) & \dots & \phi_n(p,p) \end{bmatrix} \times \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \dots \\ \dots \\ \alpha_p \end{bmatrix} = \begin{bmatrix} \phi_n(1,0) \\ \phi_n(2,0) \\ \phi_n(3,0) \\ \dots \\ \dots \\ \phi_n(p,0) \end{bmatrix} \quad (4.56)$$

Neste caso, apesar de $\phi_n(i,k) = \phi_n(k,i)$, a matriz de dimensão $p \times p$ é simétrica mas não é Toeplitz.

O método de análise baseado nesta técnica de determinação da matriz $\phi_n(i,k)$ começou a ser conhecido como o método da covariância porque a matriz tem as propriedades de uma matriz covariância.

4.5 Codificação Digital Com Qualidade de Telefonia da Forma de Onda da Voz

A representação digital mais simples da voz, está ligada a uma representação directa usando técnicas de codificação tais como: PCM ("Pulse Code Modulation"), APCM ("Adaptive Pulse Code Modulation"), DPCM ("Differential Pulse Code Modulation") e ADPCM ("Adaptative Differential Pulse Code Modulation"), DM ("Delta Modulation") e ADM ("Adaptative Delta Modulation"). Estes tipos de codificação realizada no domínio dos tempos não consideram as questões da reprodução da voz em função da descrição das excitações, ressonâncias do trato vocal ou parâmetros articulatórios, apresentando vantagem do ponto de vista da complexidade do codificador. São aplicados em larga escala na codificação da voz em sistemas que exigem uma reprodução com qualidade de telefonia, e do ponto de vista de quantificação da onda.

A técnica de codificação PCM envolve processos de amostragem, quantificação, codificação, transmissão, descodificação e reconstrução. Um sinal após ter sido amostrado é quantificado digitalmente evitando assim que durante a transmissão lhe seja adicionado ruído. A quantificação pode ser uniforme ou logarítmica em que normalmente o sinal de saída é constituído pela soma do sinal informação e por ruído que é função do sinal de entrada. Se o sinal de ruído for independente do sinal de entrada, a operação de quantificação tem perdas de distorção mínimas. Esta condição pode ser alcançada usando um sinal "dither" apropriado, que é um sinal adicionado à entrada do quantificador e depois é subtraído na saída do quantificador.

Quando o quantificador usado é adaptativo a técnica toma o nome de APCM ("Adaptive Pulse Code Modulation"). Neste caso o degrau do quantificador varia no tempo: quando o nível do sinal é baixo, utiliza-se um degrau pequeno e quando a amplitude do sinal é maior usa-se um tamanho do degrau apropriado. O ajustamento do degrau é feito por operações lógicas sobre a sequência de amostras.

A codificação DPCM, permite eliminar a redundância existente nos sinais de voz permitindo assim reduzir a taxa de transmissão necessária à representação. O esquema de representação diferencial está baseado no facto de que a relação entre amostras sucessivas é elevada e, à medida que a frequência de amostragem aumenta a correlação de amostra para amostra aumenta aproximando-se da unidade para taxas de

amostragem muito elevadas. Portanto a melhor forma de explorar a redundância de amostra para amostra consiste na codificação da diferença de amostras adjacentes. Como a gama de diferenças de amostras é menor que a gama de amostras individuais são necessários um menor número de bits para codificar a diferença de amostras. Um ponto de vista mais geral, considera o codificador DPCM um caso especial de um preditor linear que codifica e transmite o erro de predição.

Dentro da quantificação diferencial pode-se considerar o uso de adaptação surgindo ADPCM.

A modelação delta (DM) é uma técnica preditiva simples, que está baseada num sistema DPCM de 1 bit. De facto, o único bit utilizado especifica a polaridade da diferença das amostras, indicando se o sinal aumentou ou diminuiu desde a última amostra. Esta técnica, explora o facto de que, se a taxa de amostragem for muito elevada, isto é, muito superior à taxa de Nyquist, a correlação entre amostras adjacentes é próxima da unidade e desta forma, ao sobreamostrar o sinal, consegue-se obter uma estratégia de codificação mais simples em que se usa um codificador de 1 bit.

Da mesma forma que em PCM e DPCM é possível melhorar a gama dinâmica do codificador usando técnicas adaptativas, também isso é conseguido em DM resultando ADM.

CAPÍTULO 5

CONVERSORES TEXTO-FALA

5. CONVERSORES TEXTO-FALA

5.1 Introdução

Nos sistemas de conversão texto-fala pretende-se que um texto escrito de forma electrónica seja lido em voz alta através de um sintetizador de fala.

Porque um Chinês não lê bem o português, tal como um português não lê correctamente russo, também estes sistemas devem ser dedicados a uma língua, pois o conjunto base de fonemas, a prosódia e a escrita são diferentes em línguas diferentes.

Este tipo de sistemas tem hoje um crescente interesse para as mais diversas áreas.

Uma primeira motivação para o estudo e desenvolvimento destes sistemas é a ajuda preciosa que podem dar a pessoas com determinadas incapacidades. As únicas formas que um invisual tem de comunicar com um computador são o terminal "braille" com custos muitas vezes inacessíveis e o conversor texto-fala que lhe pode dar uma "imagem auditiva" do que acontece no écran do computador permitindo-lhe assim ter acesso a livros escritos em forma electrónica, mensagens recebidas via correio electrónico, jornais que estejam disponíveis na rede publica de dados, aproveitar as suas excelentes capacidades para desenvolver programação, etc.. Para deficientes temporária ou definitivamente impedidos ou com grandes dificuldades de falar, permite que estes se façam ouvir com recurso a um programa dedicado de onde podem seleccionar e compor rapidamente e com facilidade um grande número de mensagens pré-gravadas.

Não se pode esquecer o esforço que tem vindo a ser desenvolvido para facilitar cada vez mais a "interface" da pessoa com o computador onde se caminha a passos largos para que esta comunicação possa ser realizada através de mensagens faladas.

Depois há toda uma série de campos para estes sistemas como sejam: serviço de atendimento a clientes em bancos e estações de meteorologia, dicionários multilíngua para turistas, correio de voz, instruções em simuladores, etc..

Não restando agora dúvidas do interesse destes sistemas é fácil compreender que para serem bem aceites pelo público devem ter uma grande qualidade de voz sintética. Esta qualidade deve hoje ir mais além do que a simples inteligibilidade, trata-se de conseguir a naturalidade de um falante humano.

Um leitor humano perante um texto introduz uma informação denominada prosódia, que normalmente não aparece explicita no texto. A prosódia relaciona os diferentes sons de uma mensagem falada e permite reflectir tanto elementos linguísticos, imprescindíveis para o sentido da oração, como elementos não linguísticos, sejam: características do leitor, o seu estado de ânimo, etc..

5.2 Sistemas de Conversão Texto-Fala

Para se atingir uma boa qualidade da fala com síntese artificial é indispensável juntar à mensagem textual toda a riqueza de informação que nos proporciona a fala natural.

Um sistema de conversão texto-fala é composto por dois módulos claramente distintos, que requerem para a sua realização uma metodologia e conhecimentos de

base radicalmente distintos: o processamento linguístico-prosódico e o processamento acústico (figura 5.1).

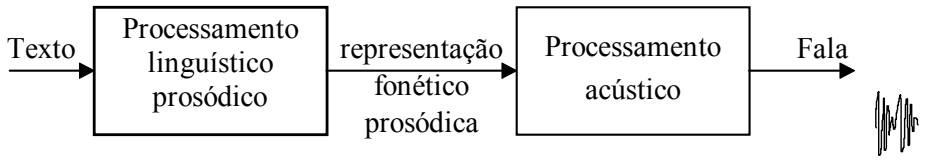


Figura 5.1 - Diagrama de blocos genérico de um sistema de conversão texto-fala.

5.2.1 Processamento Linguístico

É objectivo do processamento linguístico determinar, a partir de um texto, dois tipos de informação necessários para proporcionar ao processamento acústico dados que lhe permitam gerar fala natural. Estes dois tipos de informação são conhecidos como informação segmental e informação suprasegmental.

1) A informação segmental está associada à cadeia de sons que compõem a mensagem . Para cada língua existe um conjunto limitado de sons base ideais que permitem produzir, quando correctamente combinados, todas as particularidades da fala nessa língua. Criam-se assim uma série de representações abstractas denominadas por fonemas cujo número depende da língua em causa.

2) A informação suprasegmental está associada à prosódia. Reflecte tanto elementos linguísticos, tais como tipos de frase, pausas, acentuação e agrupação de elementos de significado como elementos não linguísticos. Esta informação segundo [López 93] e [Gósy 91], é a chave para conseguir uma elevada naturalidade em sistemas de síntese de fala. Esta informação vem geralmente codificada através de três parâmetros acústicos do sinal de fala:

a) A evolução temporal da frequência fundamental, que é o aspecto mais importante do ponto de vista preceptivo.

b) Duração dos segmentos de som que compõem a frase.

c) Curva de energia do sinal acústico.

Nos conversores texto-fala actuais estes dois tipos de informação são extraídos por uma sequência de tarefas que genericamente se representam pela figura 5.2.

Pré-processamento - A primeira tarefa a realizar no processamento linguístico é a formatação do texto representando adequadamente na sua forma textual, números, abreviaturas, etc.. Esta é uma tarefa bastante dependente da aplicação em que se desenvolverá o conversor.

Análise linguística - Depois do pré-processamento realiza-se a análise linguística que abarca tanto uma análise sintáctica como uma análise semântica, no intuito de encontrar o foco (segmento com maior conteúdo semântico) da oração e tentar modelar aspectos como a ênfase. Esta tarefa é bastante complexa e muito dependente do idioma. A análise gramatical é normalmente realizada sobre um dicionário com um

léxico relevante (formas verbais, expressões comuns) e uma tabela de prefixos e sufixos. Normalmente, podem incluir-se regras de conteúdo gramatical para determinar as categorias gramaticais das palavras que não tenham sido encontradas no dicionário.

Análise morfosintáctico-prosódica - Nesta tarefa pretende-se, a partir da análise anterior, marcar, por um lado, fronteiras sintáctico-prosódicas, e por outro lado, os acentos. As fronteiras sintáctico-prosódicas ficam definidas pela sua natureza (relação lógica entre duas estruturas consecutivas) e força relativa (pausas, alargamento de sílabas). Os acentos obedecem melhor a aspectos rítmicos e de ênfase principal.

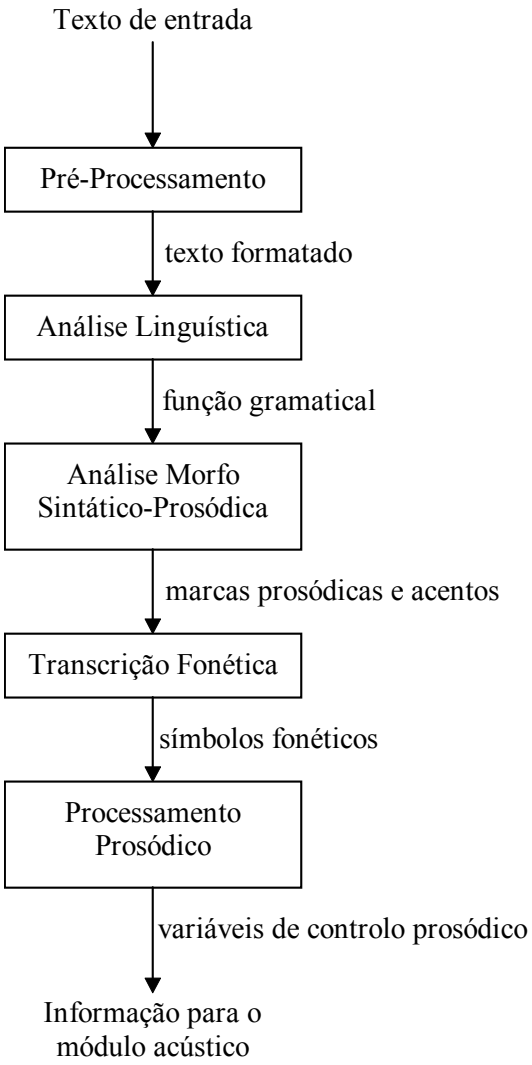


Figura 5.2 - Diferentes tarefas do processamento linguístico.

Transcrição fonética - A transcrição fonética automática do texto cuja saída é uma sequência de códigos fonéticos em vez de um texto com eventuais marcas impostas pelas tarefas anteriores, é realizada, geralmente, mediante regras dependentes do contexto que devem ter em conta, além do contexto, a existência de pausas no contexto e acentuação, por efeito de coarticulação. Para a língua portuguesa estas regras são particularmente complexas no que diz respeito à transcrição das vogais já

que para a mesma vogal do alfabeto natural correspondem várias vogais do alfabeto fonético, consoante a sua posição na palavra, acentuação e fonemas adjacentes.

Processamento prosódico - A última tarefa a realizar denomina-se processamento prosódico e recolhe a informação suprasegmental e segmental extraída dos últimos passos (marcas prosódicas e transcrição fonética) para traduzi-las em variações de duração segmental (ritmo), frequência fundamental (entoação) e inserção de pausas que existam com uma duração adequada.

5.2.2 Processamento Acústico

O objectivo do processamento acústico é converter a sequência fonética e as variáveis de controlo prosódico na forma de onda associada à voz sintetizada.

Um diagrama de blocos típico para o processamento acústico é o representado na figura 5.3.

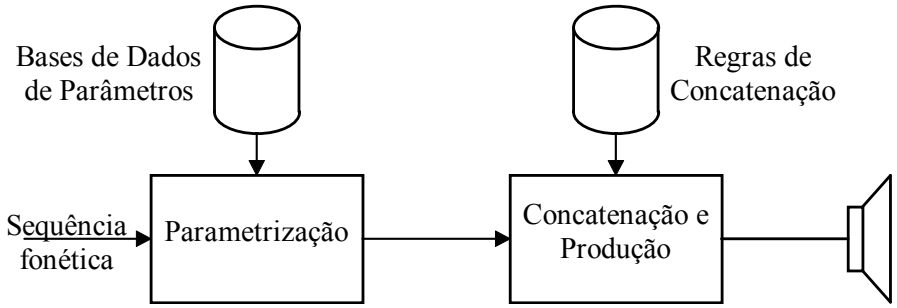


Figura 5.3 - Diagrama de blocos do processamento acústico.

Como na figura 5.3, existem dois blocos envolvidos no processamento acústico. O primeiro aspecto a realçar é a existência de um compromisso entre, por um lado, o número de regras de parametrização e concatenação, destinadas a evitar transições bruscas desagradáveis ao ouvido e, por outro lado, o tamanho da base de dados de parâmetros. Assim, do ponto de vista do processamento acústico pode-se estabelecer uma ampla gama de sistemas de conversão texto-fala que abarca desde os sistemas comandados por regras aos sistemas comandados por dados. De uma maneira concisa num sistema "puro" comandado por regras, estas geram a representação paramétrica que alimentará um sintetizador de fala e num sistema comandado por dados, estes representam directamente segmentos de fala. Entre estes dois casos podem-se encontrar sistemas intermédios.

Em qualquer caso o modelo de produção de fala deve ser flexível para o controlo prosódico e deve ter uma alta qualidade na geração de fala sintética. Assim segundo [López 93] utilizam-se actualmente três grupos principais de modelos de sintetizador.

> 1) **Sintetizadores de formantes:** nestes sintetizadores as sequências fonética e prosódica controlam as ressonâncias e a excitação do sintetizador de formantes. O sintetizador de formantes consiste numa composição de filtros que modelam as ressonâncias e anti-ressonâncias das cavidades vocal e nasal. A configuração mais genérica para o modelo destes filtros é a sua ligação em

série e em paralelo. Trata-se de um procedimento, com enorme flexibilidade, que sintetiza a fala com elevada qualidade, mediante ajuste manual dos parâmetros do sintetizador. Sem dúvida que é necessário um número enorme de regras para a síntese automática, o que requer compiladores cada vez sofisticados capazes de integrar todo o conhecimento adquirido com a experiência de trabalho com sintetizadores.

2) **Sintetizadores mediante modelos articulatórios:** trata-se de simular a propagação das ondas acústicas no trato vocal. Os segmentos e as variáveis prosódicas traduzem-se em parâmetros de um modelo simplificado do aparelho fonador humano, que explicitamente restringem a dinâmica do sistema podendo produzir voz da mais alta qualidade. Surgiram para fazer corresponder explicitamente os sintetizadores de formantes a um modelo mais explicito do trato vocal. O seu interesse centra-se no facto de as restrições implícitas neste modelo permitirem ver a fala como um contínuo acústico pelo que se evitam os problemas de concatenação de segmentos. A dificuldade principal destes tipos de sistemas é que ainda não se conhece totalmente o processo de produção da fala humana.

3) **Sintetizadores baseados em concatenação de unidades:** nestes sistemas realiza-se a concatenação de um conjunto de unidades extraídas da produção humana. Neste tipo de sintetizadores deve estar presente um algoritmo que permita, além da concatenação de unidades, modificar prosodicamente os segmentos a concatenar. Adicionalmente, nestes sintetizadores podem-se usar técnicas de codificação de voz para reduzir a necessidade de armazenamento da base de unidades acústicas. Também existe a possibilidade de incluir no modelo de codificação de voz as tarefas de concatenação e modificação prosódica, sempre que o codificador parametrize o sinal de fala com uma flexibilidade suficiente para a modificação prosódica das unidades.

5.3 O Conversor Texto-Fala MULTIVOX

No decurso deste trabalho foi desenvolvida uma primeira fase de implementação da língua portuguesa no conversor texto-fala MULTIVOX. Trata-se de um conversor multilíngua de origem Húngara que conta já com a implementação de nove idiomas entre os quais o Húngaro, Alemão, Italiano e Espanhol.

Após esta primeira fase de desenvolvimento do português (cerca de 6 meses), o conversor possui uma qualidade inteligível, denotando evidentemente a necessidade de mais desenvolvimento.

As ferramentas usadas nesta implementação foram:

- O próprio sistema de desenvolvimento que, ao nível do processamento acústico, com alguma facilidade, permite ouvir o som sintetizado de uma lista de parâmetros, alterar a base de sons e as regras de concatenação.

- O Phonovox que permite sintetizar ficheiros com sequências de parâmetros de fala natural previamente analisada, ou converter um texto escrito, já que este sistema faz uso da estrutura e bases de dados do conversor MULTIVOX, permitindo assim por comparação auditiva ou inspecção visual analítica ou gráfica da sequência de valores dos formantes, larguras de banda, frequência fundamental e amplitude, aproximar os parâmetros da fala sintetizada dos

parâmetros da fala natural. Este sistema é extraordinariamente útil no estudo de uma sequência de parâmetros, já que permite a audição continuada repetidamente ou esporádica de partes desta sequência de segmentos de parâmetros, desde a sequência completa passando por apenas uma parte seleccionada dessa sequência até à audição de apenas um segmento, ao mesmo tempo que se podem alterar valores dos parâmetros desta sequência, e ouvir os seus efeitos. É também útil no estudo de regras de entoação de frases já que permite acompanhar a evolução da frequência fundamental ao longo de uma oração.

- Um conjunto de programas em linguagem C já desenvolvidos para outros idiomas, realizando alguns módulos do processamento linguístico que por simples edição e adequação do conjunto de regras para a língua portuguesa realiza, parte do processamento linguístico para a língua portuguesa.

A seguir apresenta-se um diagrama de blocos do funcionamento do conversor.

5.3.1 Blocos Constituintes

Do ponto de vista do desenvolvimento fonético este sistema difere dos restantes por possuir módulos dependentes do idioma representados na figura 5.4 a traço descontinuo. As setas bidireccionais da figura referem-se à interdependência de certos blocos que deve ser considerada durante o desenvolvimento do sistema.

A entrada do sistema é o texto numa sequência de códigos ASCII. Esta sequência de códigos é convertida por um "filtro de pré-processamento" numa sequência de caracteres pertencentes à representação interna de caracteres do MULTIVOX. Por a língua portuguesa compreender caracteres não habituais em outras línguas, foi criada uma representação interna adicional de caracteres do MULTIVOX para o português. No Anexo A1 é apresentada a tabela de conversão dos códigos ASCII 850 e 860, por não serem coincidentes e por ambos serem usados em Portugal, na representação interna de caracteres do MULTIVOX.

Neste módulo foi desenvolvido também para o português uma rotina de conversão de números em códigos fonéticos desses mesmos números até 1 bilião.

Ainda neste módulo e especialmente para o português, foi desenvolvida uma rotina de conversão de abreviaturas e acrónimos mediante uma tabela de regras de conversão facilmente editável e passível de alteração, acrescento ou remoção das regras. Nesta tabela é também possível editar regras para a correcta transcrição fonética de palavras homógrafas, que embora escritas da mesma maneira têm formas diferentes de pronúncia consoante a sua categoria gramatical e portanto a sua posição na oração, ou fonemas adjacentes, já que este sistema não faz a distinção gramatical das palavras.(exemplo: "frango no espeto" - sendo "espeto" pertencente à categoria gramatical substantivo masculino com transcrição fonética [•petu] em oposição a "espeto a agulha" - sendo neste caso "espeto" presente do indicativo do verbo espetar conjugado na primeira pessoa do singular com transcrição fonética [•pɐtu]).

Esta sequência de caracteres internos do MULTIVOX é a entrada do módulo "Nível de conversão grafema-fonema", que realiza a transcrição fonética regido por uma tabela de regras "Regras de conversão grafema-código de fonema na forma tabular", em formato tabular para a conversão de grafemas em códigos de fonemas dependente

da língua e com base num grupo de fonemas básico "Grupo de sons de fonemas base", para a língua em causa.

A saída deste sistema é o nível 1 de representação em códigos fonéticos. Este nível de representação consiste em códigos de fonemas e algumas marcas prosódicas inerentes já dos módulos anteriores.

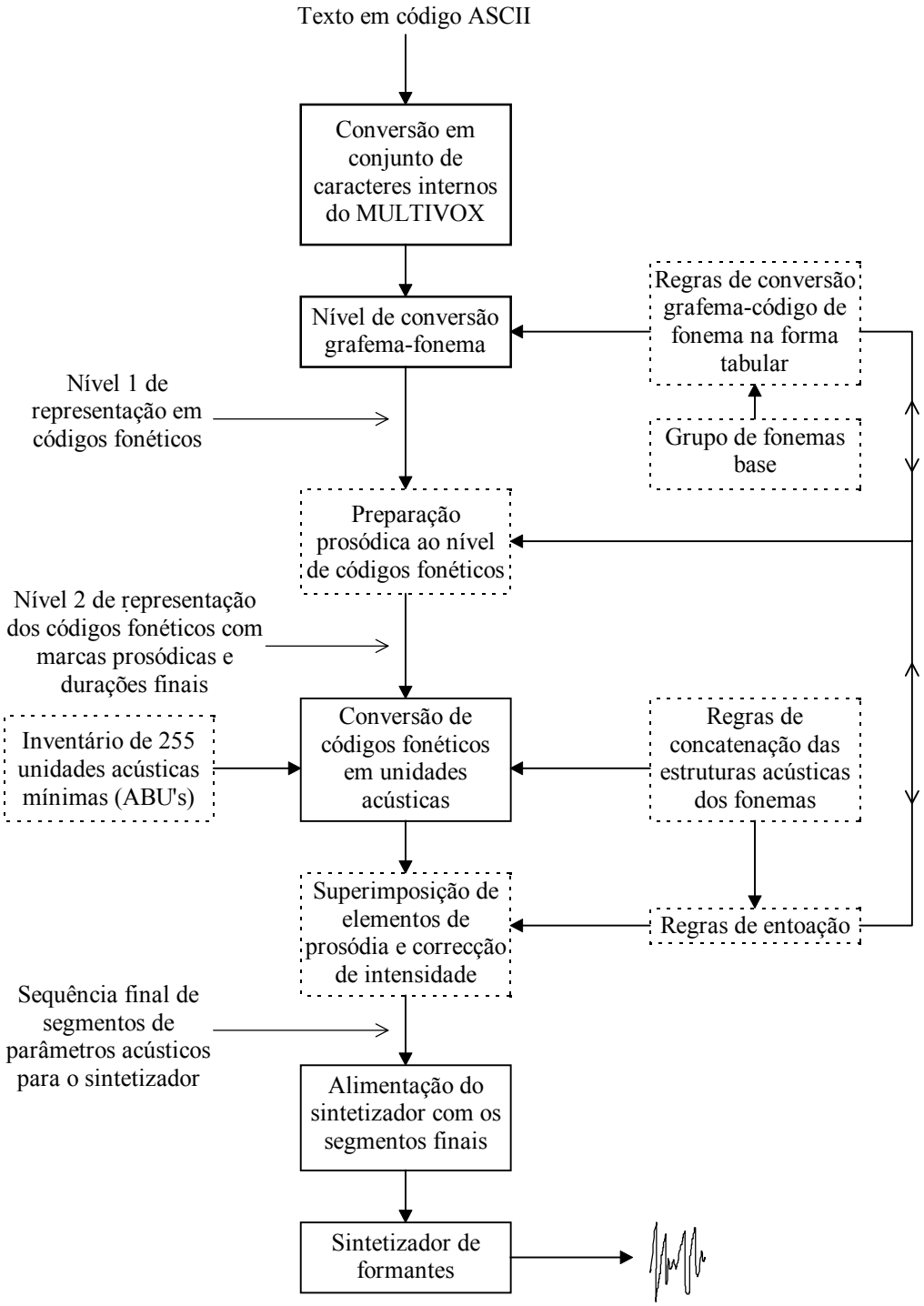


Figura 5.4 - Módulos constituintes do conversor MULTIVOX.

O módulo seguinte, "Preparação prosódica ao nível do som", baseado nos códigos fonéticos e algumas marcas prosódicas realiza, recorrendo a algoritmos desenvolvidos em linguagem C, um ajuste de códigos fonéticos para algumas regras de excepção não possíveis de implementar de forma tabular (exemplo: códigos fonéticos de [e] [m] seguidos de código fonético de uma consoante, troca a sequência de códigos fonéticos de [e] [m] pelo código fonético de [ẽ]). Este conjunto de regras para o português é apresentado no anexo A2. Ainda neste módulo é desenvolvida uma preparação prosódica ao nível de marcação da sílaba tónica, de entoação das orações e programação rítmica.

A saída deste módulo é o nível 2 de representação dos códigos fonéticos com marcas prosódicas e durações finais.

A partir deste nível 2 de representação os códigos fonéticos são convertidos no módulo "Conversão de códigos fonéticos em unidades acústicas" baseado nas regras de concatenação das estruturas acústicas dos fonemas, dependentes do idioma, em unidades acústicas de um inventário de 255 elementos de unidades acústicas, também dependente da língua

Este inventário de unidades acústicas (ABU's - Acoustic Building Units) é próprio para cada língua tendo sido desenvolvido para o português a partir do inventário para o espanhol. Cada unidade acústica corresponde a um segmento de fala com duração entre 10 e 50 ms e composto pelos parâmetros do modelo do sintetizador (4 formantes, 4 larguras de banda, amplitude, duração e indicação de vocalização ou não). A frequência fundamental não é um parâmetro destas unidades acústicas porque é imposta por elementos de prosódia consoante a posição na palavra e na oração de cada segmento.

Esta superimposição é realizada no módulo seguinte "Superimposição de elementos de prosódia e correcção de intensidade" baseada nas regras de entoação bem como nas marcas prosódicas existentes no nível 2 de representação dos códigos fonéticos. Este módulo não só impõe a variação da frequência fundamental, mas também faz correcções de amplitude e duração dos segmentos acústicos.

A saída deste bloco é a sequência final de segmentos de parâmetros acústicos que são enviados para o sintetizador de formantes.

O sintetizador de formantes implementado em "hardware" tem um modelo cujos parâmetros já foram referidos, 4 ou 5 formantes e respectivas larguras de banda, duração, amplitude e indicação da frequência fundamental no caso de segmentos vocalizados ou indicação de ruído em segmentos não vocalizados.

5.3.2 A Conversão Grafema Fonema Para o Português

A conversão aqui denominada por grafema-fonema que visa realizar automaticamente a transcrição fonética de um texto em fonemas, ou neste caso preciso em códigos de fonemas, para o português foi desenvolvida a partir do zero visto não se encontrar na literatura qualquer referência a regras que permitam realizar esta conversão.

A abordagem deste tema pretende não só explicar a forma de implementação desta conversão no MULTIVOX, mas também deixar documentada uma primeira abordagem ao conjunto de regras que realizam automaticamente a transcrição fonética de um texto em português.

Para a realização desta conversão é necessário estabelecer previamente um conjunto de fonemas base que caracterize a língua em causa, neste caso o português. Este conjunto foi estabelecido tendo por base a matriz fonológica para o português segundo [Martins 92] e [Mateus 90], já apresentada no capitulo 2, acrescentando alguns sons básicos que teoricamente são realizados pela junção de dois fonemas mas que na prática, no sintetizador, resulta desastrosa. Assim foi criado um conjunto de 38 sons básicos. A maior parte deles já foram caracterizados no capitulo 2, mais os sons [ão], [qu] e [t•], este último normalmente não usado na língua portuguesa mas que pode ser necessário para pronunciar nomes estrangeiros. A lista de fonemas e respectivos códigos usados é apresentada na tabela 5.1.

Tabela 5.1 - Símbolos fonéticos e respectivos códigos usados no conversor texto-fala MULTIVOX (versão portuguesa).

Símbolo	**Fonema**	**Código do fonema**	**Exemplo**
	(pausa)	1	(pausa)
a	[a]	2	p**a**to
oo	[ɔ]	3	g**o**la
o	[o]	4	p**o**ço
u	[u]	5	p**u**la
@	[ɐ]	6	s**e**car
i	[i]	7	l**i**vro
e^	[e]	8	P**e**dro
a-	[ɑ]	9	bal**a**
e	[ɛ]	10	t**e**rra
b	[b]	11	**b**ata
p	[p]	12	**p**ara
d	[d]	13	**d**a**d**o
t	[t]	14	**t**erra
g	[g]	15	**g**ato
k, c	[k]	16	**c**asa
ão	[ɐ̃u][i]	17	c**ão**
lh	[ʎ]	18	fi**lh**o
m	[m]	19	a**m**a
n	[n]	20	**n**ada
nh	[ɲ]	21	pi**nh**o
ii	[j]	22	pa**i**
livre	livre	23	
v	[v]	24	**v**aca
f	[f]	25	**f**ilho
z	[ʒ]	26	ca**s**a
s	[s]	27	**s**ábado
j	[ʒ]	28	**j**ardim
uu	[w]	29	pa**u**

[i] Teoricamente trata-se de um conjunto de dois fonemas [$\tilde{\alpha}$] e [u]. No MULTIVOX foi implementado como um som base, não havendo para este som uma representação no alfabeto fonético internacional.

ch	[•]	30	**ch**ama
tch	[t•]	31	**tch**eco
l	[l]	32	a**l**a
rr	[r]	33	ca**rr**o
r	[r̃]	34	ca**r**o
qu	[k] e [w]	35	**qu**ando
an	[α̃]	36	c**an**to
on	[õ]	37	p**on**te
in	[ĩ]	38	p**in**to
un	[ũ]	39	f**un**do
en	[ẽ]	40	d**en**te

De referir que o texto de entrada para esta conversão está na forma de representação interna dos caracteres para o MULTIVOX cujos caracteres que diferem dos habituais são apresentados no anexo A1.

A conversão grafema fonema no MULTIVOX é realizada em duas etapas. Uma primeira na forma tabular que comporta as regras elementares e de fácil especificação bem como algumas excepções à regra (normalmente vocábulos, palavras e pequenos agrupamentos habituais de palavras). A segunda etapa comporta regras, que pela sua complexidade não são possíveis de implementar na primeira etapa, programadas em linguagem C e actuam sobre os códigos fonéticos transcritos na primeira etapa.

As regras da primeira etapa na forma tabular são apresentadas no anexo A3 e o seu funcionamento é o seguinte:

- Armazenamento das regras: as regras estão organizadas por ordem alfabética da sua primeira letra. Cada letra do alfabeto representa um bloco separado de regras na tabela. A organização de cada um destes blocos tem contornos similares ao triângulo "saw-tooth" desenvolvido por [Olaszy 90].], isto é, no topo do bloco ficam as regras mais compridas (com maior número de caracteres) e no pico do triângulo, ao fundo, localiza-se a última regra que consiste em apenas uma letra (a letra inicial do próprio bloco de regras). A posição da regra no bloco é conducente com a sua prioridade. Assim o topo do bloco contém as regras de mais alta prioridade, sendo esta decrescente ao longo do bloco. Esta prioridade advém do modo como é feita a pesquisa das regras na tabela. Esta pesquisa é realizada da seguinte forma:

- Algoritmo de conversão:

1) O algoritmo de conversão toma a primeira letra do texto.

2) Seguidamente, coloca-se no bloco de regras iniciado por essa letra.

3) Começa por comparar a primeira coluna (lado esquerdo das regras) do bloco com a sequência de letras do texto, enquanto incrementa o apontador passo a passo.

4) Se todas as letras da primeira coluna são as mesmas que as do texto, o conversor coloca os códigos do lado direito da regra num "buffer" e

posiciona o apontador na letra seguinte do texto da acordo com a última letra da regra. A pesquisa seguinte inicia-se na nova posição do apontador.

5) Se a comparação, da esquerda para a direita, falha numa letra o conversor coloca o apontador do texto na posição anterior ao início da pesquisa e o procedimento de comparação inicia-se no próximo bloco de regras correspondente à nova letra inicial.

- Edição das regras: cada regra contém um sinal de igualdade separando o lado esquerdo da regra que contém a sequência de caracteres a converter, do lado direito que contém os códigos fonéticos correspondentes aos caracteres do lado esquerdo. São usados dois símbolos especiais para a formatação das regras. O "backslash" (\) indica um passo atrás no apontador da sequência de caracteres na regra. O til (~) indica um caracter qualquer, isto é o algoritmo de pesquisa aceitará qualquer caracter no lugar do (~) quando usar a regra. Estes dois símbolos são usados apenas no lado esquerdo da regra. O zero (0) do lado direito serve como uma marca de fim da sequência de códigos de fonemas. Estes símbolos do lado esquerdo e o zero (0) do lado direito podem ser usados para aumentar o tamanho da regra dando-lhe assim maior prioridade. São também usados, sem qualquer efeito a nível de regra, para igualar o número de caracteres do lado esquerdo da regra com o número de códigos fonéticos do lado direito, já que esta igualdade é obrigatória por imposição de software. Chama-se a atenção para o facto de as regras iniciadas, no lado esquerdo, com um espaço ou terminadas com um espaço e "backslash", são relativas respectivamente ao início e fim de palavras.

Quer nas regras de conversão na forma tabular quer na segunda etapa desta conversão são impostas algumas marcas de prosódia. Estas marcas têm códigos, que aparecem na tabela de regras do anexo A3, com numeração superior aos códigos dos fonemas (máx. 40). Assim caberá aqui uma explicação de qual a função prosódica relativa a cada uma dessas marcas (tabela 5.2).

Tabela 5.2 - Função prosódica relativo a cada código de marca no nível 1 de representação do conversor texto-fala MULTIVOX.

Código da marca prosódica	**Função Prosódica**
45	marca um elemento não acentuado
49	palavra interrogativa
52	aumenta a frequência fundamental (tom) durante a próxima palavra
60	aumenta intensidade da próxima palavra
61	reduz intensidade da próxima palavra
64	aumenta a frequência fundamental (tom) no texto a partir desta marca

65	reduz a frequência fundamental (tom) no texto a partir desta marca
80	palavra mais rápida
81	palavra mais lenta
85	Impede que os códigos fonéticos da palavra sejam alterados
88	Junção à próxima palavra (elimina pausa posterior)
89	Junção à palavra anterior (elimina pausa anterior)
202 a 240	Sílaba tónica. O fonema (vogal) acentuado(a) fica com o código 200 + código do fonema.
254	Virgula

A segunda etapa da conversão grafema fonema é implementada em linguagem C e as suas regras estão descritas no anexo A2.

Em forma de finalizar o assunto da transcrição fonética para o português refere-se que o sistema de regras descrito não está completo, tendo necessidade de mais desenvolvimento, contudo a experiência do trabalho com este sistema deixa antever que este caminho leva a uma transcrição fonética perfeita. Neste momento os resultados são já bons pois para texto corrido a transcrição fonética é realizada com um sucesso elevado, sendo poucos e pouco graves os erros cometidos.

5.3.3 Regras de Concatenação de Fonemas

As regras de concatenação de fonemas são dependentes da língua, bem como o inventário de 255 unidades de construção acústica.

O inventário de ABU's para o português foi criado a partir do inventário de ABU's para o espanhol.

O número de regras de concatenação iguala o número de combinações dos fonemas base mais um, dois a dois nas duas posições possíveis[i].

O intento destas regras é implementar sons, com base no inventário de ABU's desde a zona estável do primeiro fonema até à parte estável do segundo fonema, realizando a transição entre fonemas.

O estabelecimento destas regras foi realizado através de um processo de análise por síntese. Procurou-se ajustar a concatenação de unidades acústicas e as próprias unidades acústicas de forma a que o som sintetizado fosse o mais próximo possível do som pretendido, bem conhecido pela produção humana.

O reconhecimento das distâncias entre os sons sintetizados e os sons humanos foi realizado por três processos diferentes. O primeiro, e mais usado, foi a comparação auditiva dos sons sintetizados com os sons produzidos naturalmente pelas pessoas. Este processo apesar de ser o mais conclusivo em termos de resultados finais, não permite, antes de uma larga experiência com o sistema avaliar em que sentido se deve

[i] Para o caso da implementação da versão portuguesa no MULTIVOX este número ascende a 39*39=1521 regras.

proceder ao ajuste dos parâmetros nem que parâmetros devem ser ajustados. Contudo, este processo é sempre usado simultaneamente com os próximos dois, já que permite uma avaliação final e conclusiva da proximidade dos sons sintetizados dos sons pretendidos. O segundo processo recorre ao uso da ferramenta Phonovox, já apresentada neste capitulo. Neste processo a medida da distância dos parâmetros dos sons sintetizados e dos parâmetros previamente obtidos por um processo de análise da fala natural, é mensurável e permite saber que parâmetros devem ser ajustados e em que sentido. Este processo está obviamente limitado à qualidade com que foi realizada a análise e ao conjunto de sons analisados (recorda-se que o número de regras de concatenação é de 1521). Esta foi concerteza uma grande motivação para o desenvolvimento de sistemas de análise durante este trabalho de preparação da dissertação, apresentados nos capítulos seguintes. O terceiro processo, permite avaliar as distâncias dos espectrogramas de sons sintetizados dos espectrogramas de sons produzidos naturalmente por voz humana. Este processo permite visualizar o contorno dos formantes dos sons produzidos naturalmente e ajustar esses parâmetros dos sons sintetizados. É portanto mais uma medida objectiva das distâncias dos sons. Este processo foi usado durante pouco tempo, numa estadia em Budapeste, já que faz uso de um espectrógrafo não disponível no desenvolvimento decorrente no Porto. Realça-se mais uma vez o uso indispensável do primeiro processo já que, embora não permitindo um conhecimento de que parâmetros ajustar, é sem dúvida uma avaliação final e conclusiva do estado de aproximação dos sons.

O facto das regras de concatenação usarem apenas ABU's de um inventário de 255 unidades reduz muito a quantidade de memória usada pelo sistema de conversão texto-fala, no entanto ao tentar estabelecer as regras de concatenação nem sempre se podem usar os sons que mais se adequariam à concatenação de certos fonemas, estando-se limitado à escolha do segmento do inventário que mais se aproxima do pretendido.

5.3.4 Regras de Prosódia

Um grupo de regras para imposição da prosódia foi estudado e implementado. Haverá concerteza melhoramentos a realizar especialmente nos aspectos que dizem respeito ao ritmo e à entoação das orações neste conversor.

Passam-se a enumerar as mais importantes regras que foram implementadas ao nível da acentuação das palavras, da entoação das frases e da programação rítmica.

a) Regras de acentuação das palavras:

1. - O texto escrito que contenha uma das seguintes letras será convertido com a marca de acentuação da palavra. {á, é, í, ó, ú, à, è, ì, ò, ù, ã, õ, â, ê, ô}. As marcas de acentuação correspondentes a estas regras são inseridas nas regras de conversão da forma tabular.

2. - Quando não há nenhuma marca de acentuação nas palavras terminadas por uma das seguintes sequências, é colocada uma marca de acentuação na última sílaba. {al, el, ol, ar, er, ir, az, ez, iz, oz, uz}.

3. - Quando não existe ainda na palavra nenhuma marca de acentuação, é seguida a regra geral de acentuação na penúltima sílaba.

As 2ª e 3ª regras são implementadas ao nível da preparação prosódica.

Elementos não acentuados: os artigos, conjunções e prefixos (exp.: que, se, um) são marcados com uma marca de não acentuação (código 45) nas regras de conversão da forma tabular.

O algoritmo de realização da acentuação ajusta a duração da vogal da sílaba acentuada para o dobro (duração de 2 vogais), sobe ligeiramente a frequência fundamental e a amplitude sendo depois decrescentes até ao final da palavra.

b) Entoação da frase:

Foram estudadas e programadas regras de entoação para frases do tipo declarativo, interrogativo sem palavra interrogativa e interrogativo com palavra interrogativa.

Para as frases declarativas foi programado um decaimento da frequência fundamental antes do fim da frase. Depois da acentuação da penúltima palavra foi implementada uma tendência decrescente da frequência fundamental. Na última palavra da frase não é realizada nenhuma variação de frequência fundamental associada à marca de acentuação.

Para frases interrogativas com palavra interrogativa (exp.: como, onde, que, qual, quando, quantos) reconhece-se um contorno da frequência fundamental diferente das mesmas frases sem palavra interrogativa. Assim foram implementados os seguintes contornos da frequência fundamental para estas frases: uma subida no início da palavra interrogativa e uma descida brusca no início da palavra seguinte. No final da frase uma nova subida da frequência fundamental é imposta na última palavra com uma eventual descida caso a última palavra não seja acentuada na última sílaba.

Nas frases interrogativas sem palavra interrogativa não acontece a subida e descida da frequência fundamental no início da frase como no caso anterior e apenas a última palavra é afectada pela variação da frequência fundamental como no caso anterior.

É reconhecida uma entoação diferente para frases com tamanhos relativamente diferentes. Foi já iniciado o estudo para o contorno da frequência fundamental em frases de diferentes tamanhos, não tendo sido ainda implementada nenhuma regra desta natureza.

Ainda ao nível de entoação, quando é detectada uma virgula insere-se uma pausa antes desta e processa-se um tratamento especial da frequência fundamental na palavra anterior, impondo uma descida na penúltima sílaba até ao final da palavra.

c) Programação rítmica:

A implementação de simples regras de programação rítmica tornam a fala sintética menos robótica e mais fluente. Por exemplo a junção de alguns artigos com a palavra vizinha (tanto a anterior como a próxima), alterando ou suprimindo alguns códigos de fonemas na reunião de algumas palavras, especialmente nos casos de algumas vogais e fricativas. Noutros casos quando a partícula "e" é usada, produz-se um efeito idêntico ao efeito de uma virgula.

Muitas destas regras são programadas no nível de preparação prosódica, outras são realizadas pela inclusão de marcas da tabela 5.1 nas regras de conversão na forma tabular.

Exemplo: Junção à próxima palavra:

Marca 88 + códigos de fonemas da palavra. Esta marca pode ser usada para marcar as palavras que devem ser agrupadas por eliminação da pausa intermédia à próxima palavra.

CAPÍTULO 6

FERRAMENTAS USADAS/CRIADAS COMO SUPORTE À DETERMINAÇÃO AUTOMÁTICA DE PARÂMETROS

6. FERRAMENTAS USADAS/CRIADAS COMO SUPORTE À DETERMINAÇÃO AUTOMÁTICA DE PARÂMETROS

6.1 Introdução

Neste capitulo faz-se uma abordagem a um conjunto de ferramentas computacionais importantes que foram usadas na determinação automática dos parâmetros do modelo dos sinais de fala quer pelo método do cepstro quer pelo recurso à predição linear. Algumas destas ferramentas foram criadas com recurso à programação em Matlab sendo aqui descrito o seu algoritmo, outras já existem implementadas no programa Matlab tendo sido apenas utilizadas, mas que pela sua importância é apresentada uma breve fundamentação teórica.

As primeiras ferramentas descritas (amplitude média deslizante, energia média deslizante, taxa de passagem por zero e classificação/segmentação) desenvolvem todo o seu processamento exclusivamente no domínio temporal, enquanto as restantes ferramentas actuam no domínio das frequências.

É também apresentada a implementação de um simples sintetizador desenvolvido em Software baseado no modelo de formantes utilizado na análise. Este sintetizador foi usado para testar trechos de fala analisados.

6.2 Amplitude Média Deslizante

A amplitude média de uma sequência de amostras do sinal de fala é uma ferramenta simples que pode ajudar na decisão de vocalização ou não ao longo de um sinal de fala.

O seu cálculo é obtido pelo somatório do módulo da amplitude das N mais recente amostras dividido por N.

$$M(n) = \frac{1}{N} \sum_{m=n-N+1}^{n} |x(m)| \tag{6.1}$$

O efeito deste cálculo baseia-se em olhar para uma sequência infinitamente longa de amostras do sinal de fala x(m) através de uma janela de apenas N amostras e calcular a média da amplitude das amostras vistas pela janela.

Quando se fala em média deslizante não é mais do que fazer deslizar esta janela, dentro da qual se calcula a média, ao longo da sequência infinitamente longa de amostras do sinal de fala.

Uma formula alternativa da amplitude média torna o cálculo da média deslizante mais explicito quando se realiza a convolução da função da janela w(n-m) com a sequência de amostras do sinal de fala x(m):

$$M(n) = \frac{1}{N} \sum_{m=-\infty}^{\infty} |x(m)| \times w(n-m) \tag{6.2}$$

onde w() é a função da janela, que selecciona algumas amostras da sequência infinita x(m) para as quais n-m não é zero. É bem conhecido que a convolução no domínio do tempo é equivalente à multiplicação no domínio das frequências, pelo que a amplitude média de curta duração ("short-time"), M(n) pode também ser estimada filtrando a amplitude do sinal com um filtro linear de resposta impulsional w(). Este filtro tem as características de um filtro passa baixo.

Há duas razões para preferir a utilização da média deslizante ao filtro passa baixo. A primeira é que com a utilização do filtro, devido ao tempo de estabelecimento da sua resposta, as primeiras amostras do sinal filtrado, antes do tempo de estabelecimento, não tenham sentido por serem erradas. Em oposição, com a média deslizante todas as amostras, excepto nos extremos a uma curta distância de metade do comprimento da janela, menor que no caso do filtro, fazem sentido. Além disso, calculando a média com uma janela cada vez mais curta nos extremos do sinal, pode-se fazer com que todas as amostras tenham sentido pela sua correcção. Esta primeira razão é especialmente importante para sinais curtos. A segunda razão é que o alisamento provocado por estes dois processos é mais facilmente controlável com a média deslizante. Enquanto com o filtro o parâmetro de controlo é a frequência de corte, na média deslizante esse parâmetro é o comprimento da janela.

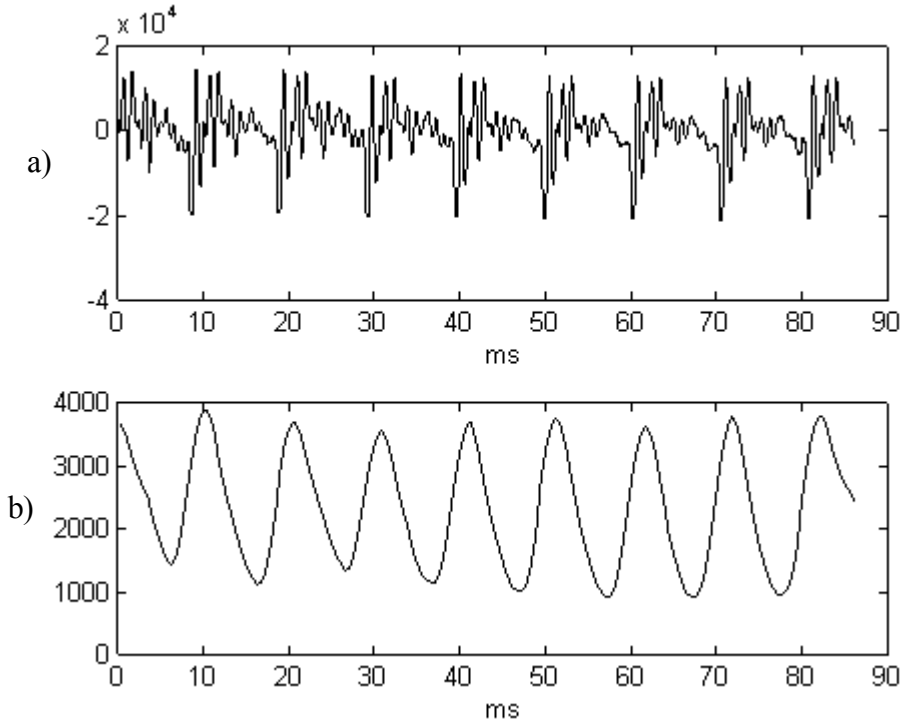


Figura 6.1 - Aplicação da amplitude média a um sinal de fala vocalizado. a) sinal de fala da vogal [a]; b) sinal alisado pela amplitude média deslizante com uma janela de comprimento 150 e espaçamentos de 10 amostras.

Assim o recurso à média deslizante no campo da análise da fala tem duas funções importantes. A primeira, usando uma janela relativamente curta (da ordem de um período fundamental), na fala vocalizada consegue-se um alisamento que praticamente elimina as componentes de alta frequência correspondentes ao trato

vocal, ficando um sinal que grosso modo corresponde ao período fundamental como mostra a figura 6.1. Neste sinal alisado é relativamente fácil a medição da frequência fundamental como veremos no próximo capitulo, bem como, com um pouco mais de cuidado, detectar o sincronismo com o período fundamental. A segunda função desta ferramenta é que usando uma janela com mais amostras consegue-se um maior alisamento do sinal resultando num elemento importante para detecção de zonas vocalizadas, não vocalizadas, mistas ou pausas ao nível da palavra ou frase (figura 6.2). Isto deve-se ao facto das zonas vocalizadas terem normalmente uma energia maior que as não vocalizadas e as pausas terem um nível mínimo de energia correspondente ao ruído ambiente e inerente ao processo de aquisição. O nível máximo de amplitude média de zonas de pausas pode ser medido no início de um trecho de fala iniciado com silêncio.

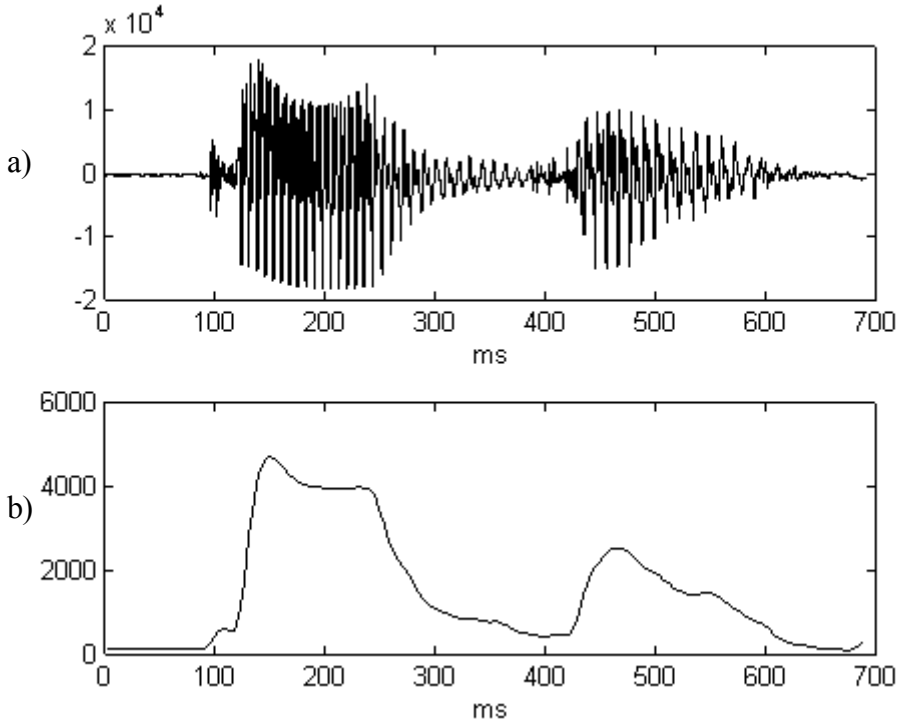


Figura 6.2 - Aplicação da amplitude média deslizante. a) sinal de fala correspondente a "tudo"; b) média deslizante com janela de comprimento 350 e espaçamentos de 50.

A prática demonstra que o cálculo da média das amplitudes centrada em amostras sucessivas não tem variações significativas. Então, o cálculo da média das amplitudes pode-se realizar centrado em amostras equiespaçadas de um determinado espaçamento para aliviar o seu peso computacional.

Este espaçamento não deve ser maior que o comprimento da janela para não deixar amostras fora do cálculo da média, nem é aconselhável um espaçamento superior a metade da janela, pois levaria a que algumas amostras entrassem no cálculo da média em duas janelas sucessivas enquanto outras amostras entrariam em apenas numa janela, dando assim pesos diferentes às amostras.

A função w() de pesagem temporal usada foi a janela de Hanning por apresentar melhores resultados.

A ferramenta criada é uma função no programa Matlab cujo código é apresentado no anexo B1. Esta função denominada *fmedia()*, tem como parâmetros de entrada o sinal que se pretende sujeitar à média deslizante, o comprimento da janela e o espaçamento entre amostras. Também se aplica esta função para outras situações além da amplitude do sinal de fala, como quando por exemplo, se pretende obter um alisamento. Esta função determina a média também nos extremos, com uma janela cada vez mais curta e igualmente espaçada desde a amostra correspondente a metade do comprimento da janela até à primeira amostra. O mesmo acontece no fim do sinal.

6.3 Energia Média Deslizante

A energia de um sinal pode ser estimada pela variância, que é o quadrado da diferença do valor da amostra e o seu valor médio. Então para sinais com valor médio nulo (caso dos sinais de fala), a energia de um curto período de tempo ("short-time") pode ser definida como a média do quadrado dos valores das amostras.

$$E(n) = \frac{1}{N} \sum_{m=-\infty}^{\infty} \left[x(n) \times w(n-m) \right]^2 \tag{6.3}$$

onde w() é a função de uma janela estável, que relaciona os valores de x(n) próximos de n, em que w(n-m) não é zero. Isto pode ser reescrito como

$$E(n) = \sum_{m=-\infty}^{\infty} x^2(m) \times h(n-m) \tag{6.4}$$

em que

$$h(n) = \frac{1}{N} w^2() \tag{6.5}$$

Da mesma forma que para a média deslizante, também aqui a energia deslizante E(n) pode ser estimada filtrando o quadrado do sinal com um filtro linear de resposta impulsional h(n), que teoricamente é o quadrado de uma função de janela apropriada.

Comparando a média deslizante com a energia deslizante de uma sequência de amostras do sinal de fala, verifica-se que a energia deslizante tem características que mostram uma maior diferença entre a fala vocalizada e a fala não vocalizada, e entre a fala não vocalizada e o silêncio. Então, como defende [Rowden 92], a estimação da energia média deslizante é preferível em aplicações como um detector de fala, usado em combinação habitual com detector de passagem por zero descrito na próxima secção.

Tal como no cálculo da amplitude média deslizante também o cálculo da energia média em amostras sucessivas não difere muito, pelo que mais uma vez para aliviar o cálculo computacional a energia média deslizante pode ser determinada centrada em amostras equiespaçadas de um determinado espaçamento. A natureza deste espaçamento obedece às mesmas restrições do espaçamento usada no cálculo da amplitude média deslizante.

Depreendem-se para esta ferramenta três aplicações importantes. Duas delas são as mesmas que para a amplitude média deslizante. Contudo, é preferível usar a amplitude média deslizante para a determinação do período fundamental e detecção de sincronismo, apesar dos resultados serem idênticos (figura 6.3), por ser mais leve computacionalmente. No entanto, para a determinação do modo de excitação e decisão de silêncio ou não, é preferível a energia média deslizante, por nesta ser mais facilmente visível as diferenças entre silêncio e fala não vocalizada e entre fala não vocalizada e fala vocalizada como mostra a figura 6.4.

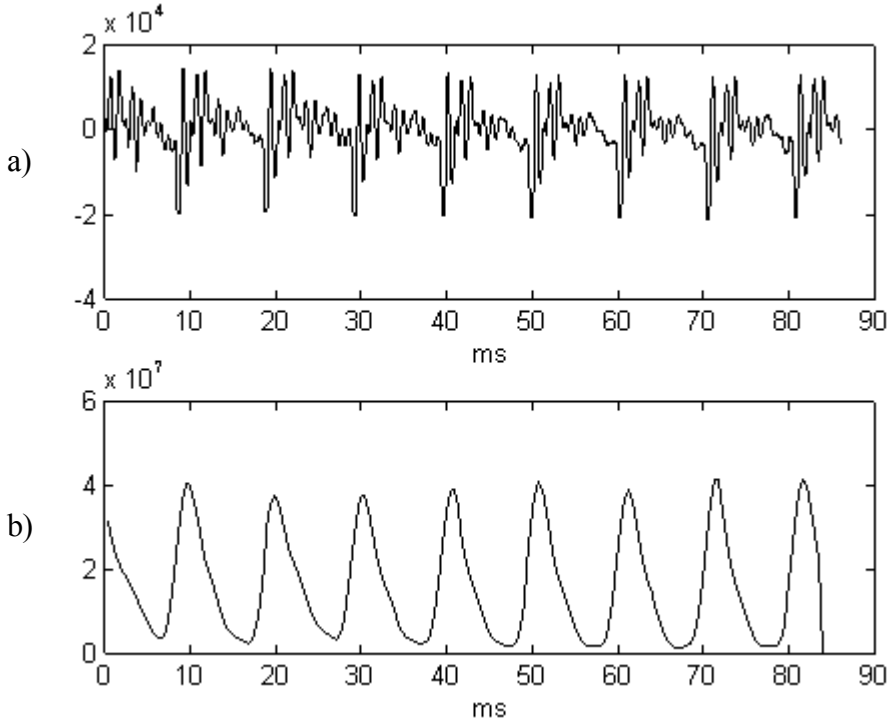


Figura 6.3 - Aplicação da energia média deslizante a um sinal de fala vocalizada. a) sinal de fala da vogal [a]; b) sinal alisado pela energia média deslizante com uma janela de 150 amostras e um espaçamento de 10 amostras.

A terceira aplicação desta ferramenta é mais direccionada para o domínio do reconhecimento da fala e ajuda a detectar o início e fim de palavra. Para este caso o comprimento da janela é 1, resultando num cálculo muito rápido para aplicações em tempo real.

A função desenvolvida no programa Matlab tem o nome de *fenergia()* e é apresentada no anexo B2. Os seus parâmetros de entrada são também o sinal sobre o qual será calculada a energia média deslizante, o comprimento da janela e o espaçamento.

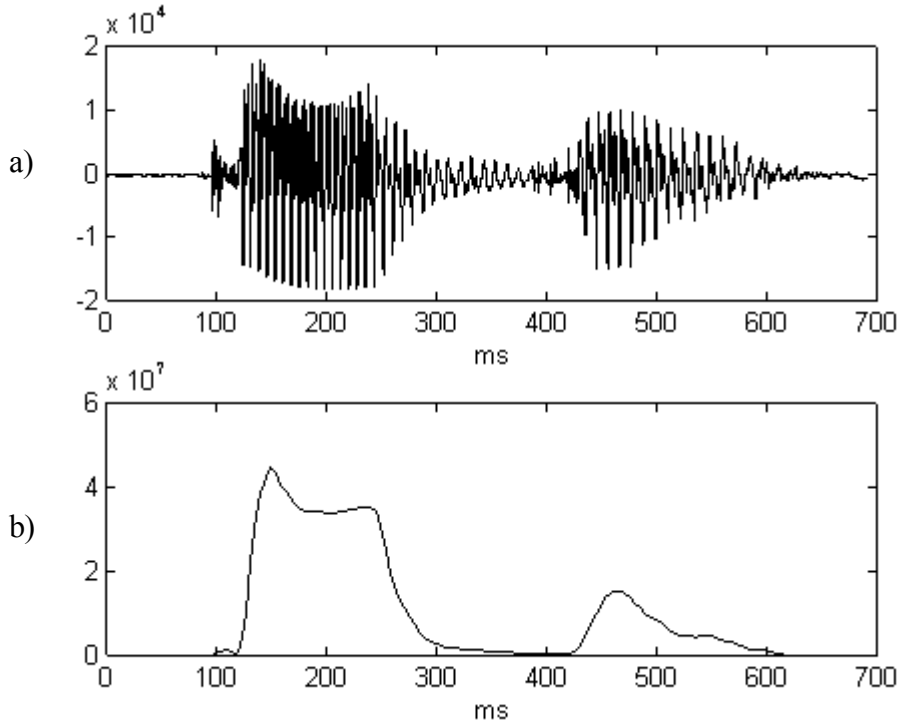


Figura 6.4 - Aplicação da energia média. a) sinal de fala correspondente a "tudo"; b) energia média deslizante com uma janela de comprimento 350 e espaçamento de 50 amostras.

6.4 Taxa de Passagem Por Zero

Uma importante diferença entre segmentos de fala vocalizada e segmentos de fala não vocalizada e silêncios é a razão a que o sinal passa por zero. Esta característica pode ser estimada pela função

$$Z(n) = \frac{1}{2N} \sum_{m=-\infty}^{\infty} \left| \text{sign}(x(m)) - \text{sign}(x(m-1)) \right| \times w(n-m) \tag{6.6}$$

onde

$$\text{sign}(x(m)) = \begin{cases} 1 & \text{se } x(n) \geq 0 \\ -1 & \text{se } x(n) < 0 \end{cases} \tag{6.7}$$

Mais uma vez, a taxa de passagem por zero é determinada com recurso à convolução de uma janela que desliza ao longo do sinal.

Também neste caso não faz muito sentido determinar a taxa de passagem por zero centrada em todas as amostras consecutivas, por a sua variação ser insignificante. Então o espaçamento entre as amostras sobre as quais se centra o cálculo da taxa de passagem por zero é também uma variável de controlo desta ferramenta.

A estimação de Z(n) pode ser significativamente afectada para a fala não vocalizada e para os silêncios por um pequeno desvio, "offset", na amplitude das amostras do sinal de fala por causa do seu baixo nível de amplitude. [Rowden 92] diz que o sinal deve ser filtrado por um filtro passa alto com uma frequência de corte aos 70 Hz a quando da sua aquisição no sentido de conseguir uma redução significativa destes efeitos. Nesta dissertação sugere-se, com os bons resultados apresentados na figura 6.5, que a taxa de passagem por zero seja aplicada, não directamente ao sinal, mas sim à derivada do sinal. Em que a derivada é determinada pela expressão

$$D(n) = x(n+1) - x(n) \tag{6.7}$$

atendendo a que as amostras estão separadas por um espaço temporal de Δt constante como é o caso.

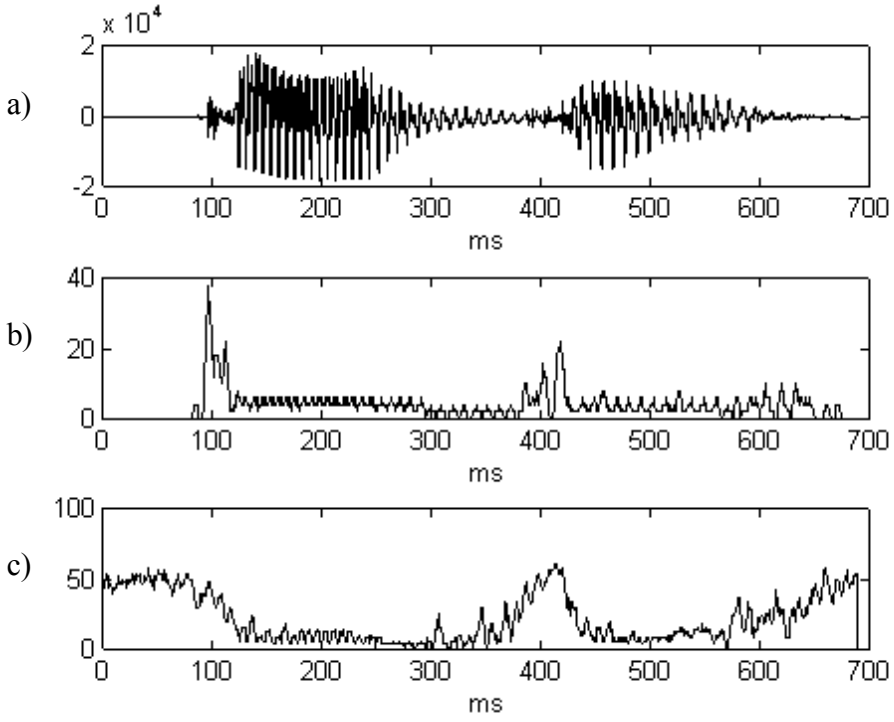


Figura 6.5 - Aplicação da taxa de passagem por zero. a) sinal de fala correspondente à locução "tudo"; b) taxa de passagem por zero aplicada directamente ao sinal de a) com uma janela de 50 amostras e um espaçamento de 10 amostras; c) taxa de passagem por zero aplicada à derivada do sinal de a) com uma janela de 50 e um espaçamento de 10 amostras.

É visível na figura 6.5 b) que a taxa de passagem por zero aplicada directamente ao sinal não resultou como se pretendia, já que esta taxa na zona de silêncio é nula quando deveria ser mais alta que na zona vocalizada. Este mau resultado deve-se ao facto de haver um pequeno "offset" na amplitude do sinal. Por outro lado, quando se aplicou a taxa de passagem por zero à derivada do sinal (figura 6.5 c), tornou-se perfeitamente possível fazer a distinção entre segmentos vocalizados de segmentos não vocalizados e silêncio.

No anexo B3 são apresentadas as funções que foram criadas no sistema Matlab para determinação da derivada e da taxa de passagem por zero com os nomes respectivamente *fderivad()* e *fzeros2()*.

6.5 Classificação Quanto ao Modo de Excitação / Segmentação

Como já se discutiu no capítulo 4, os modelos de parametrização dos sinais de fala são variantes no tempo considerando segmentos estáveis da ordem dos 10 a 30 ms. Também a Transformada de Fourier usada na análise dos sinais de fala tem como pressuposto ser aplicada a sinais estáveis. Torna-se então claro que a análise dos sinais de fala deve ser realizada segmento a segmento. O modo como o sinal de fala é particionado em segmentos é o que se pretende discutir nesta secção, bem como a classificação de cada segmento relativamente ao modo de excitação. A posterior análise de cada um destes segmentos é matéria de estudo dos capítulos seguintes.

Por segmentação designa-se aqui o acto de dividir a sequência de amostras do sinal de fala em pequenas quantidades de amostras contíguas designadas por segmentos com uma duração entre 10 a 30 ms. Os segmentos não têm que ser contíguos, pode haver sobreposição de segmentos.

Por classificação designa-se aqui o acto de classificar uma determinada quantidade de amostras do sinal de fala nos diferentes tipos: silêncio, fala vocalizada, fala não vocalizada e fala com excitação mista (glotal e ruído).

Encontram-se dois caminhos distintos relativamente ao modo de segmentar um sinal de fala. O primeiro, habitualmente usado, consiste em definir um comprimento fixo, para os segmentos a usar, satisfazendo as condições de o sinal se manter relativamente estável durante o período de tempo correspondente ao comprimento do sinal e o número de amostras ser confortável para um rápido e eficiente cálculo da FFT ("Fast Fourier Transform"). Este será o comprimento de uma janela de pesagem temporal que deslizará sobre o sinal com uma eventual sobreposição. Assim cada segmento a analisar consiste na multiplicação do sinal pela função da janela, no domínio dos tempos. Estes segmentos serão depois classificados quanto ao modo de excitação e analisados de acordo com o modelo definido para o tipo de segmento. Contudo, este caminho para a segmentação não fará uma correcta transição entre silêncio e fala, pois o ponto de transição entre o silêncio e a fala não é exactamente o fim de um segmento e início de outro, a menos que aconteça por coincidência. Um comportamento idêntico é verificado em relação a explosões[i] de curta duração que quando não "encaixam" num segmento são analisados em dois segmentos consecutivos que contêm silêncio no início do primeiro e no fim do segundo. Apesar de tudo estas contrariedades são minimizadas quando se usam segmentos de curta duração e funções de janelas sem transições bruscas (funções de janelas tipo Hanning ou Hamming) com uma sobreposição tal que garanta um peso idêntico para todas as amostras (50%).

O segundo processo para segmentar o sinal, de certo modo sugerido aqui, consiste em estudar o sinal como um todo para classificar e distinguir com rigor as zonas de silêncio, as zonas de fala vocalizada, fala não vocalizada e mista. Em cada zona é realizada a segmentação por um processo semelhante ao anterior mas com janelas de

[i] Transições de curta duração que ocorrem normalmente após o período de oclusão de algumas consoantes ([t] [p] [k]).

comprimento ajustável e procurando-se que estas tenham um comprimento idêntico e obedecendo mais uma vez a que o sinal se mantenha relativamente estável durante este período de tempo. É ainda desejável que o número de amostras seja confortável para o cálculo da FFT. Tal como no primeiro caso, também é recomendável que as funções das janelas não tenham transições bruscas e sejam sobrepostas de forma a garantir um peso idêntico para todas as amostras.

No processo de classificação das zonas em diferentes tipos de sinal há partes curtas que não são possíveis de classificar com muita certeza. Estas partes ficam classificadas como não definidas e no processo de segmentação são agrupadas às zonas contíguas, já que se trata normalmente de partes que fazem a transição entre dois tipos diferentes de zonas. No processo de decisão a qual das zonas contíguas, no caso de serem de tipos diferentes, estas partes devem ser agrupadas, deve-se ter em atenção a que na transição da fala vocalizada para a fala não vocalizada e vice versa há uma zona de excitação mista. Se não for este o caso podem-se usar estas partes para juntar à zona contígua cuja dimensão seja mais crítica para a escolha da dimensão do segmento ou segmentos a usar.

Este segundo processo de segmentação tem o inconveniente, que poderá não ser visto como tal, de os segmentos terem comprimentos variáveis. Este facto pode dificultar um pouco o processo de análise no que diz respeito ao ajuste de variáveis de controlo que sejam dependentes do comprimento do segmento a analisar.

Ressalva-se que a resolução frequencial da aplicação da FFT a segmentos de comprimentos diferentes pode não ser afectada se se recorrer à técnica "zero padding", isto é, realizar FFT's do mesmo comprimento juntando zeros ao segmento até igualar o comprimento da FFT que deve ser maior ou igual ao máximo comprimento dos segmentos.

Os dois métodos descritos para segmentar um sinal de fala conduzem a duas filosofias opostas para realizar a segmentação/classificação. Assim o primeiro método remete para que seja realizada a segmentação e depois para cada segmento classifica-se o seu tipo. Ao contrário, o segundo método classifica primeiro todo o sinal de fala e só depois faz a segmentação nas diferentes zonas classificadas do sinal.

Contrariando a ideia de dois procedimentos opostos para cada tipo de segmentação foi desenvolvido um sistema de classificação inspirado num algoritmo de discriminação de silêncio versos fala, apresentado por [Rabiner 78], que genericamente faz uma partição do sinal de fala em partes muito pequenas, classifica cada uma dessas partes e depois segmenta de acordo com o método pretendido.

A dificuldade de descriminar (classificar) os sinais de fala torna-se tanto maior quanto menor for a relação sinal ruído. Assim, mais uma vez se chama a atenção para a necessidade de se seguirem as recomendações do capitulo 3 relativamente à aquisição do sinal, nomeadamente a utilização de material de boa qualidade e a redução do ruído de fundo (sala insonorizada).

Habitualmente as dificuldades de classificação surgem em sons do tipo:

1. Fricativas não vocalizadas ([f] [s] [•]) no início e fim.
2. Início e fim de explosões das oclusivas ([p] [t] [k]).
3. Final das nasais.
4. Fricativas vocalizadas que perdem a vocalização no final de palavras.

O fluxograma do algoritmo desenvolvido é apresentado na figura 6.6.

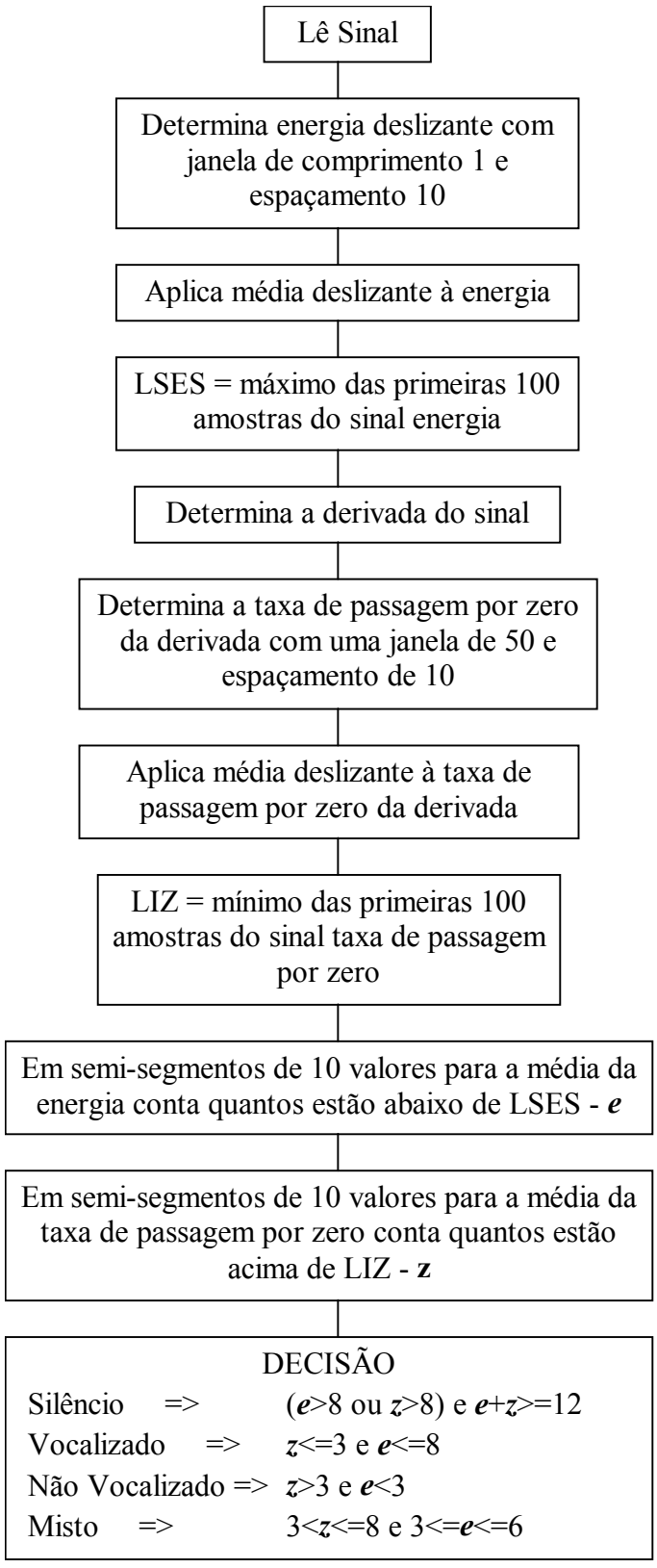


Figura 6.6 - Algoritmo para classificação de semi-segmentos de sinais de fala.

Este algoritmo baseia-se em três valores extremamente importantes: a energia deslizante, a taxa de passagem por zero e um domínio de decisão baseado no resultado de dois vectores obtidos pelas duas ferramentas anteriores.

O programa de realização deste algoritmo em Matlab é apresentado no anexo B4 com o nome *classif1.m*.

O algoritmo começa por ler o sinal que se pretende classificar. A esse sinal aplica a energia deslizante com uma janela de apenas 1 amostra e espaçamento de 10. A dimensão deste novo sinal, energia, é 10 vezes inferior ao comprimento do sinal original. Seguidamente aplica a média deslizante ao sinal energia com uma janela de comprimento 5 e espaçamento 1 para alisar este sinal. Depois usando novamente o sinal original é determinada a sua derivada e sobre esta a taxa de passagem por zero com uma janela de 50 amostras e um espaçamento de 10 amostras. O sinal de saída, zero, é submetido também a um alisamento com a média deslizante com comprimento de janela 5 e espaçamento unitário.

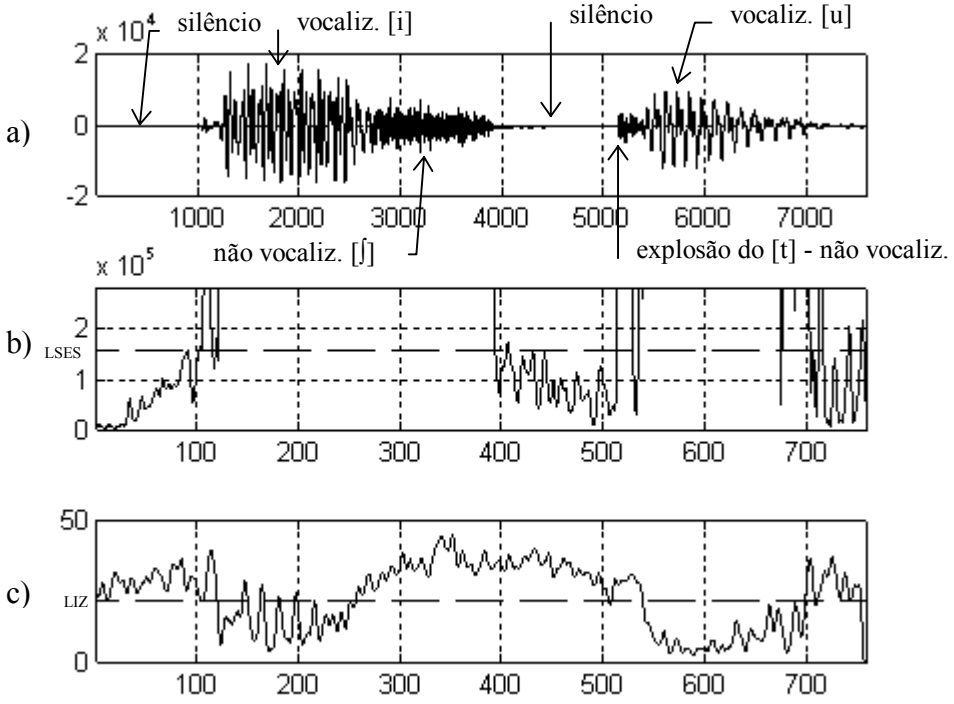


Figura 6.7 - Aplicação da TPZ e energia média deslizante. a) Sinal de fala correspondente a "isto"; b) energia media deslizante relativamente a LSES; c) taxa de passagem por zero da derivada relativamente a LIZ.

É suposto que as primeira 1000 amostras do sinal original a analisar correspondem a silêncio[i]. É sabido que este sinal correspondente a silêncio é um sinal de ruído com baixa energia e uma elevada taxa de passagem por zero. É então estabelecido um nível máximo de energia correspondente ao silêncio, LSES, que é o máximo do sinal de energia depois de alisado, nos primeiros 100 elementos, correspondentes às primeiras 1000 amostras do sinal original. É também estabelecido o nível mínimo de

[i] Na aquisição do sinal deve-se ter em conta este facto iniciando o armazenamento um pouco antes do início, visível, da fala.

passagens por zero, LIZ, nos primeiros 100 elementos do sinal zero depois de alisado, correspondentes às primeiras 1000 amostras do sinal original. O comportamento dos sinais energia e zero depois de alisados relativamente aos níveis máximo e mínimo respectivamente do sinal numa zona de silêncio, ao longo do sinal de fala "isto" é apresentado na sequência de imagens da figura 6.7.

Na sequência de imagens da figura 6.7 é claro que os sinais energia e zero são indicadores do tipo de sinal em cada instante.

Prosseguindo com o algoritmo, para cada sequência de 10 valores do sinal energia, é contabilizado na variável *e* o número de elementos abaixo do limite LSES. Para cada sequência de 10 elementos do sinal zero, é contabilizado na variável ***z*** quantos estão acima do limite LIZ. Assim, se ***e*** tem valores altos implica baixa energia, se ***z*** tem valores baixos implica taxa de passagem por zero baixa. Para cada par de valores ***e*** e ***z*** correspondem então 10 elementos dos sinais zero e energia e 100 amostras do sinal original. Finalmente, baseado em cada par de valores ***e*** e ***z*** é tomada uma decisão para classificar as 100 amostras do sinal original com os códigos: 0 - não definido, 1 - silêncio, 2 - som não vocalizado, 2.5 - som que pode ser não vocalizado ou misto (ver domínio de decisão), 3 - som de excitação mista, 4 - som vocalizado.

Tecendo agora um comentário aos critérios de decisão, estabeleceu-se um "domínio de decisão" tendo como vector no eixo das ordenadas os valores possíveis para ***z*** de 0 a 10 e o vector no eixo das abcissas com os valores possíveis para ***e*** de 0 a 10 como na figura 6.8. Empiricamente e com a experiência, estabeleceram-se zonas no domínio de decisão para o silêncio, sons não vocalizados, sons vocalizados e sons de excitação mista. Restam ainda no domínio, zonas não preenchidas que são consideradas como som não definido. Há uma zona do domínio, o rectângulo ***e***<=3 e 3<***z***<=8, que é comum aos sons vocalizados e de excitação mista, pelo que com este processo não é possível destinguir estes sons de um destes tipos de excitação. Os semi-segmentos correspondentes a esta zona são classificados com o código 2,5. No processo de segmentação, estes segmentos, serão normalmente considerados não vocalizados a menos que estejam entre segmentos de excitação mista, sendo então classificados como pertencentes a este tipo.

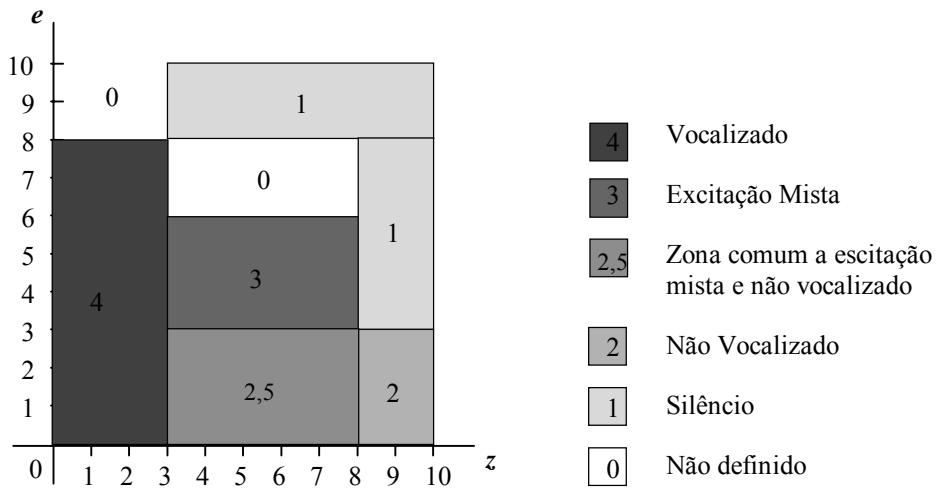


Figura 6.8 - "Domínio de decisão" para classificar sons baseado na energia e taxa de passagem por zero.

As zonas de classificação de cada tipo de som estabelecidas são:

Silêncio	$(e + z) \geq 12 \ \wedge \ (e > 8 \ \vee \ z > 8)$
Vocalizado	$z \leq 3 \ \wedge \ e \leq 8$
Não Vocalizado	$z > 3 \ \wedge \ e < 3$
Excitação Mista	$3 < z \leq 8 \ \wedge \ 3 < e \leq 6$
Zona comum a não vocalização e excitação mista	$e < 3 \ \wedge \ 3 < z \leq 8$

Estas zonas foram estabelecidas empiricamente, pelo que se admitem ajustes dos seus limites no sentido de conseguir um aperfeiçoamento da decisão de classificação. Contudo, a posição de cada tipo de som no "domínio de decisão" não é alterável já que aos sons vocalizados corresponde uma energia alta (***e*** pequeno) e tpz (taxa de passagem por zero) baixa (***z*** pequeno); para sons mistos corresponde uma energia alta (***e*** grande) e uma tpz alta (***z*** grande); os sons não vocalizados têm uma energia alta (***e*** pequeno) e tpz alta (***z*** grande); o silêncio terá energia baixa (***e*** grande) e tpz grande (***z*** grande). As zonas centrais do domínio são menos habituais, razão pela qual não se detectam facilmente as zonas de excitação mista.

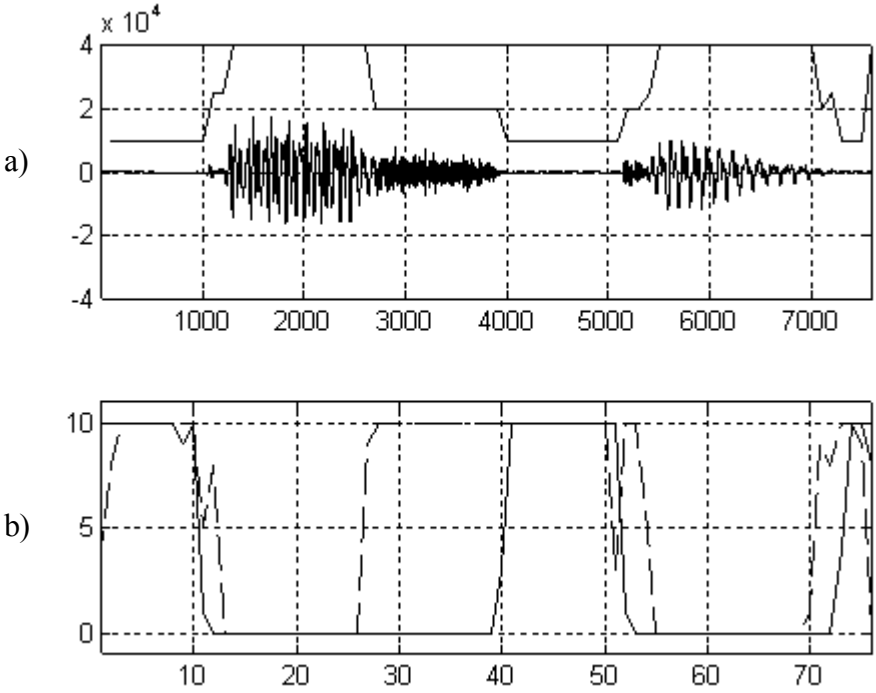


Figura 6.9 - Classificação do sinal "isto". a) sinal "isto" e o código do tipo de sinal a menos do factor de escala de 10^4 . Código 0 - não definido, 1 - silêncio, 2 - não vocalizado, 2,5 - não vocalizado ou misto, 3 - misto, 4 - vocalizado. b) a traço contínuo o vector ***e*** e a traço intermitente o vector ***z***.

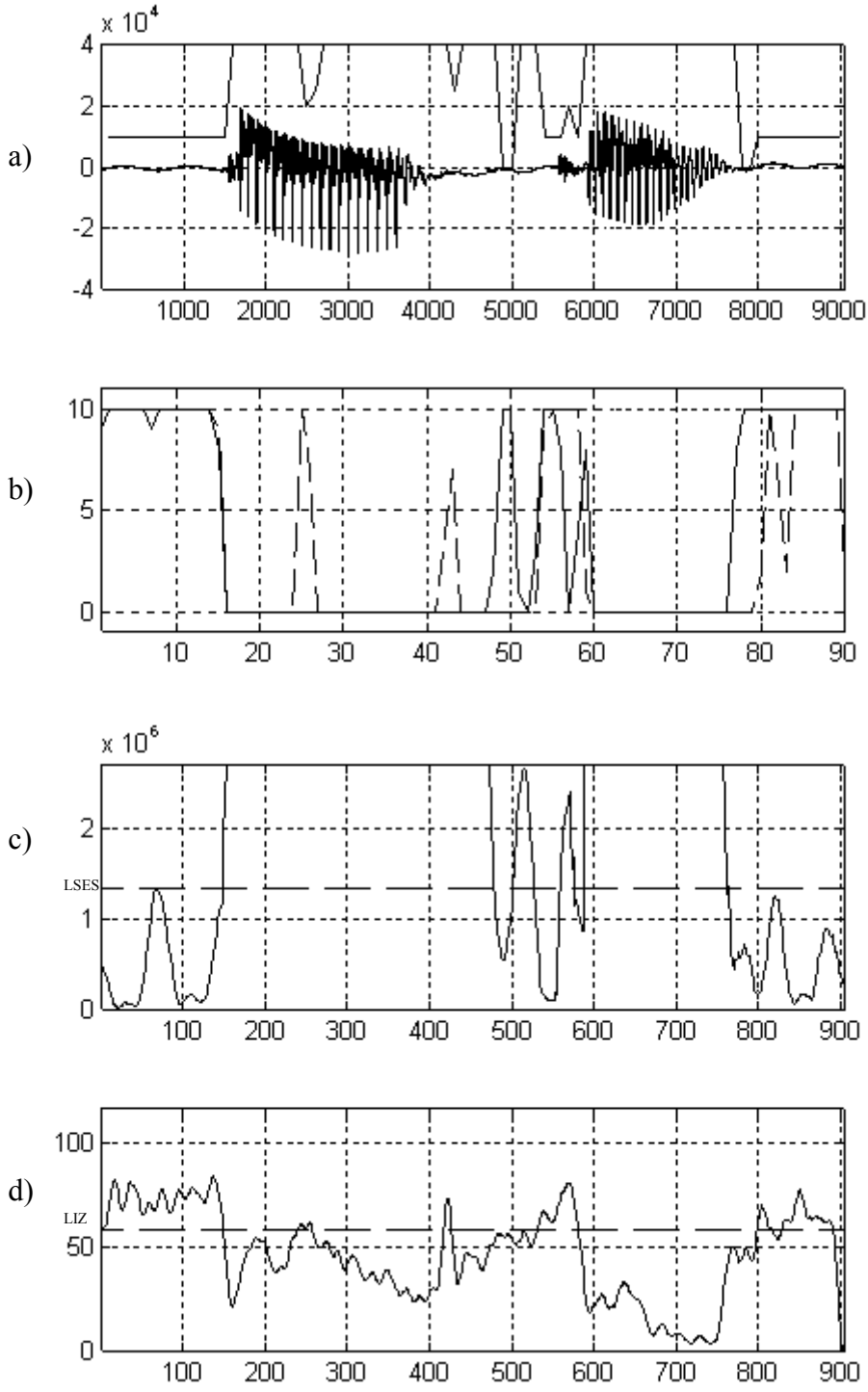


Figura 6.10 - Classificação do sinal "pato". a) sinal "pato" e o código do tipo de sinal a menos do factor de escala de 10^4 . Código 0 - não definido, 1 - silêncio, 2 - não vocalizado, 2.5 - não vocalizado ou misto, 3 - misto, 4 - vocalizado. b) a traço contínuo o vector ***e*** e a traço descontínuo o vector ***z***. c) energia do sinal relativamente a LSES. d) taxa de passagem por zero relativamente a LIZ.

Pelo facto de este algoritmo não reconhecer correctamente as zonas de excitação mista, devido à existência de uma zona comum para os sons de excitação mista e não

vocalizados, admite-se uma evolução, talvez criando mais vectores no domínio de decisão. Sugerem-se como possibilidades a criação de um valor determinado da taxa de passagem por zero para os sons não vocalizados, ou para estes mesmos sons o estabelecimento de um limite de energia.

Neste estado de desenvolvimento, se não se considerar excitação mista, pelas razões apontadas, os resultados conseguidos pelo algoritmo são apresentados nas figuras 6.9 e 6.10 para as palavras "isto" e "pato" respectivamente. Nestas figuras são também apresentados os vectores ***e*** e ***z*** ao longo do sinal e na figura 6.10 apresenta-se também a energia média relativa a LSES e a taxa de passagem por zeros relativa a LIZ. Nestas figuras, mais correctamente para a figura 6.9, é visível uma separação das zonas de silêncio das zonas vocalizadas e não vocalizadas dos sinais. Na oclusão do [t] de pato na fig. 6.10, o silêncio não está correctamente classificado por existir um "offset" de amplitude provocando uma energia superior a LSES.

6.6 Transformada de Fourier de Curta Duração "Short Time"

Sendo objecto de estudo na bibliografia [Proakis 92], [Bracewell 86], [Griffin 84], [Papoulis 84] entre outros, não será aqui desenvolvido o algoritmo da FFT ("Fast Fourier Transform") nem as suas propriedades. Contudo, esta ferramenta é de tão grande importância nesta área de análise que se deixa um breve comentário relativamente à sua utilização.

Sendo as representações de Fourier apropriadas para sinais periódicos, transitórios ou ruído estacionário, não é directamente aplicável à representação dos sinais de fala cujas propriedades variam marcadamente com o tempo. No entanto, já foi visto, que o princípio de análise de curta duração "short-time" é uma aproximação válida para o processamento da fala.

Assim cria-se o conceito de representação de Fourier dependente do tempo [Rabiner 78], motivada pela necessidade de representação espectral que reflicta as propriedades variantes no tempo das formas de onda da fala. Uma representação útil da transformada de Fourier dependente do tempo é

$$X_n(e^{jw}) = \sum_{m=-\infty}^{\infty} w(n-m)x(m)e^{-jwm} \qquad (6.8)$$

em que w(n-m) é uma janela que determina a porção de sinal de entrada analisado no índice temporal ($_n$) particular.

As janelas usadas neste trabalho para esta análise foram as janelas de Hamming e de Hanning com uma sobreposição de 50%, o que dá um peso idêntico a todas as amostras como o confirma a figura 6.11 para as duas janelas.

A largura da janela a usar é estabelecida pelo comprimento de cada segmento definido na secção anterior. No entanto, dependendo do que se pretende analisar, a frequência fundamental dos sons vocalizados ou o trato vocal, são desejáveis comprimentos diferentes para facilitar a respectiva análise. Assim, para analisar a frequência fundamental são desejáveis janelas com largura de 4 ou mais períodos fundamentais para que a sua estrutura periódica apareça bem marcada na análise e se evidencie relativamente ao trato vocal. Se por outro lado, se pretende analisar o trato vocal, são

desejáveis janelas com um comprimento de 1 ou 2 períodos fundamentais para que a estrutura periódica destes não apareça vincada em relação à estrutura do trato vocal.

A este método de processamento da FFT do sinal de fala convoluido com a janela temporal para a representação das suas propriedades variantes no tempo chamam-se vulgarmente "Short Time Fourier Transform", ou traduzido à letra Transformada de Fourier de curta duração.

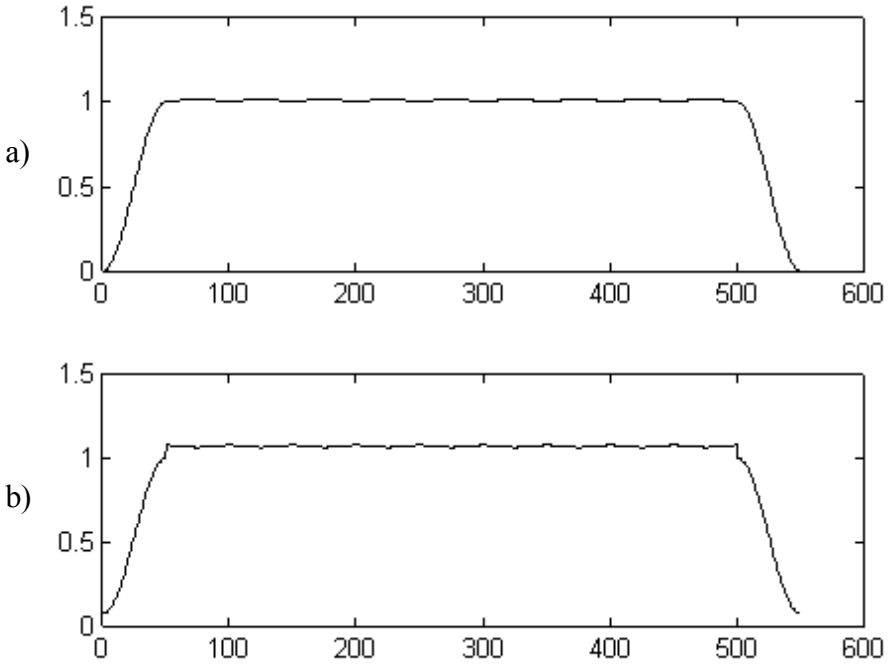


Figura 6.11 - Peso de cada uma das amostras com a aplicação de 50% de sobreposição às janelas a) Hanning, b) Hamming.

Os cálculos da transformada de Fourier são realizados com recurso à função fft() da "toolbox Signal Processing" do sistema Matlab.

6.7 Chirp Z Transform

Esta é uma técnica de passagem da representação do sinal no domínio temporal para o domínio Z das frequências.

Foi usada no algoritmo de extracção automática de formantes pelo método de análise cepstral com a finalidade de realizar o "zoom" de frequências. Para o seu uso recorreu-se à função CZT() da "toolbox Signal Processing" do programa Matlab.

A função G=CZT(X,K,W,A) usa as variáveis:

G - vector com os K elementos de saída da transformada do sinal de entrada X.

X - sinal de entrada.

K - número de pontos em que se pretende calcular a transformada.

W - espaçamento entre pontos ao longo do contorno espiral do plano de interesse.

A - O ponto complexo de início desse contorno.

O desenvolvimento teórico desta transformada é apresentado por [Schafer 70], da qual se apresentam as vantagens:

- O número de amostras do sinal no domínio dos tempos não tem que igualar o número de pontos da transformada nas frequências.

- O ponto do domínio Z a que se inicia o cálculo da transformada é arbitrário. Isto permite centrar a análise na região de frequências de interesse.

- O espaçamento frequencial das amostras espectrais é arbitrário, o que permite obter uma resolução frequencial tão fina quanto se pretenda.

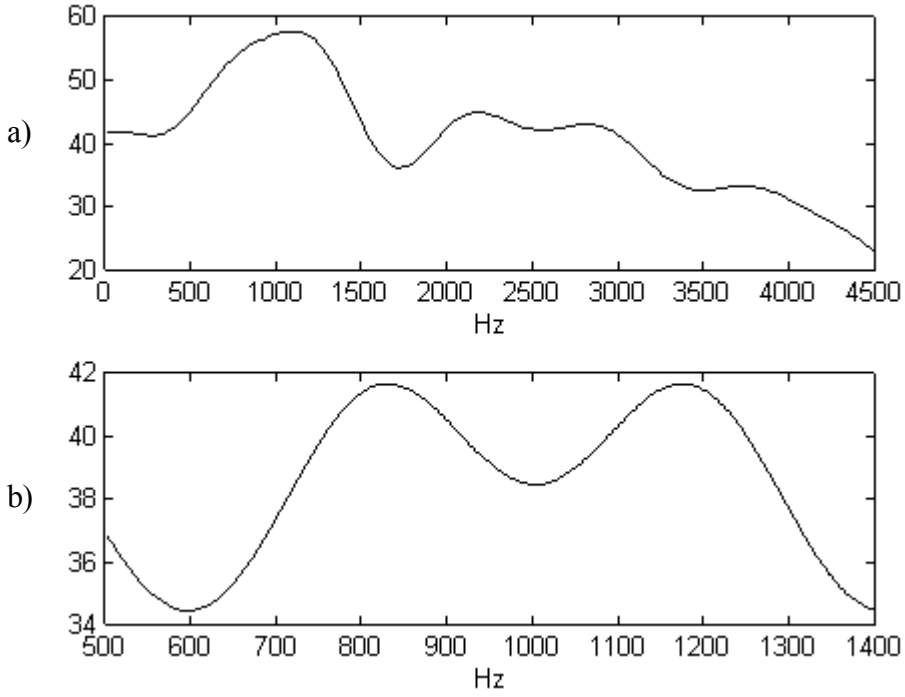


Figura 6.12 - Exemplo do "zoom" realizado pelo CZT. a) espectro alisado pelo cepstro de um segmento de fala; b) "zoom" com CZT na zona de frequências de 500 a 1400Hz.

Esta transformada é aplicada no algoritmo de extracção automática dos formantes pelo método do cepstro para se conseguir um "zoom" em frequências do espectro alisado pelo cepstro quando dois formantes aparecem suficientemente próximos ao ponto do espectro alisado não conseguir distinguir as suas frequências. Nestes casos, a função da janela de "lifteragem"[i] usada no cepstro, para separar as características do

[i] Janela usada no método do cepstro para separar gradualmente as altas das baixas quefrências e consequentemente as características espectrais do trato vocal das da frequência fundamental, alisando assim o espectro do trato vocal tanto quanto se pretenda. Ver secção seguinte.

trato vocal das características da frequência fundamental, deve ser mais larga, para que o espectro não seja tão alisado e permitir destinguir os dois formantes. Um exemplo destes casos é apresentado na figura 6.12 em que no espectro alisado pelo cepstro na figura 6.12a não se distinguem o primeiro e segundo formantes próximos na zona de frequências dos 1000Hz e na figura 6.12b depois de realizado o "zoom" do espectro naquela zona de frequências, já se distinguem as frequências dos dois formantes. F_1=820Hz e F_2=1180Hz.

6.8 Cepstro

Análise cepstral é o nome dado a uma gama de técnicas que utilizam funções que podem ser consideradas como "espectro do espectro logarítmico".

Esta ferramenta já abordada no 4º capítulo em que se referiram as aplicações na área da análise de sinais de fala, nomeadamente das propriedades que permitem realizar a separação da componente fundamental da componente do trato vocal, determinação da frequência fundamental e alisamento do envelope espectral do trato vocal.

Apresenta-se aqui uma pequena abordagem do cepstro, referindo o estudo teórico mais profundo a [Proakis 92], [Noll 67] e [Rabiner 78].

Tendo sido inicialmente proposto como a melhor alternativa às funções de autocorrelação para a detecção de ecos em sinais sísmicos, pode definir-se como a transformada inversa de Fourier do logaritmo do espectro.

A análise cepstral divide-se em dois métodos, o cepstro de potências e o cepstro complexo.

A definição do cepstro de potências é

$$C_{AA}(\sigma) = F^{-1}\left\{\log S_{AA}(f)\right\} \tag{6.9}$$

Onde F é a transformada de Fourier e $S_{AA}(f)$ é o auto espectro médio de potências, bilateral:

$$S_{AA}(f) = \overline{\left|F\left\{a(t)\right\}\right|^2} \tag{6.10}$$

O sinal analítico correspondente pode obter-se a partir do espectro logarítmico unilateral de potências

$$\hat{C}_{AA}(\sigma) = F^{-1}\left\{T_{AA}(f)\right\} \tag{6.11}$$

com

$$T_{AA}(f) = \begin{cases} 2\log S_{AA}(f) & f > 0 \\ \log S_{AA}(f) & f > 0 \\ 0 & f < 0 \end{cases} \tag{6.12}$$

a parte real de $\hat{C}_{AA}(\sigma)$ é igual à parte real de $C_{AA}(\sigma)$.

O parâmetro σ na definição é uma variável temporal, embora seja chamada quefrency.

A definição do cepstro complexo

$$C_A(\sigma) = F^{-1}\{\log A(f)\} \tag{6.13}$$

sendo A(f) o espectro complexo de a(t):

$$A(f) = F\{a(t)\} = A_R(f) + jA_I(f) \tag{6.14}$$

ou

$$A(f) = |A(f)|e^{j\phi(f)} \tag{6.15}$$

resultando

$$\log A(f) = \log|A(f)| + j\phi(f) \tag{6.16}$$

Quando a(t) é real, A(f) será par conjugada e o cepstro complexo é real.

O cepstro complexo $C_A(\sigma) = F^{-1}\{\log|A(f)| + j\phi(f)\}$ alegadamente mais poderoso que o cepstro de potências exige que a função de fase $\phi(t)$ seja uma função contínua, em vez da função apresentada em módulo 2π, como normalmente se usa e se obtém nos cálculos directos com as equações 6.14 e 6.15. Quando assim acontece deve-se usar o desempacotamento de fase para passar a função do módulo 2π para uma função contínua.

Nas funções de fase mínima (funções que não apresentam pólos nem zeros no semi-plano direito de Laplace), a fase não precisa de ser medida.

O cepstro complexo de uma função de fase mínima é causal (existe só para quefrências positivas), dado que as partes real e imaginária da respectiva transformada de Fourier estão relacionadas por uma transformada de Hilbert [Freitas 93].

Para funções de fase mínima o cepstro complexo pode obter-se do cepstro de potências duplicando os valores de quefrências positivos e anulando os de quefrências negativas.

Uma propriedade do cepstro que lhe permite um grande número de aplicações é a possibilidade de separar efeitos da fonte e dos caminhos de transmissão, isto é, efectua a desconvolução.

$$B(f)=A(f).H(f) \tag{6.17}$$

e

$$\log B(f)=\log A(f) + \log H(f) \tag{6.18}$$

e

$$F^{-1}\{\log B(f)\} = F^{-1}\{\log A(f)\} + F^{-1}\{\log H(f)\} \tag{6.19}$$

o mesmo pode ser escrito para o cepstro de potências mostrando que os efeitos da fonte e da transmissão são aditivos no cepstro de potência.

A subtracção do cepstro de potências da fonte permitirá conhecer o cepstro só da transmissão.

As aplicações para o cepstro vão desde a detecção e remoção de ecos, estabelecimento de propriedades de uma superfície reflectora, passando pela análise da fala até ao diagnóstico em máquinas.

Esta ferramenta foi usada recorrendo à função rceps() da "toolbox Signal Processing" do Matlab.

O alisamento espectral realizado pelo método do cepstro recorre a uma "lifteragem" linear. Isto é, o sinal do cepstro é sujeito a uma função de "filtragem" nas quefrências para remover a componente espectral relativa ao período fundamental.

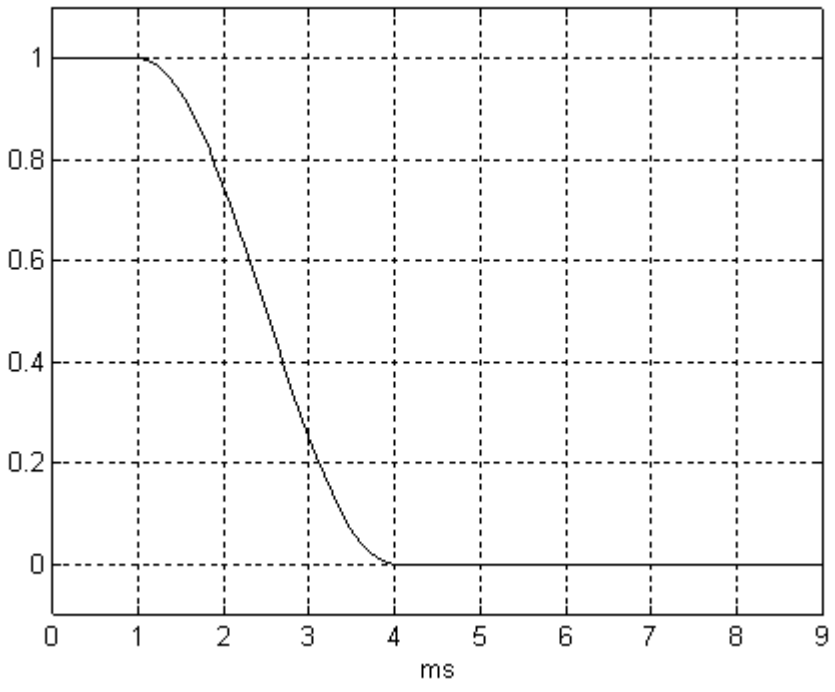


Figura 6.13 - Função da janela l(n) usada no alisamento espectral pelo método do cepstro.

A componente devida ao modelamento da forma de onda glotal e do trato vocal está concentrada na região de quefrências inferiores a σ e a componente devida ao período fundamental ocorre a quefrências superiores ou iguais a σ, em que σ é a componente

do período fundamental do segmento analisado. O período fundamental pode ser determinado procurando no cepstro um pico forte na região acima de um σ_{min} esperado para o período fundamental. O envelope espectral pode ser obtido por um "short pass lifter" da amplitude logarítmica da transformada discreta de Fourier. No método do cepstro esta filtragem é realizada multiplicando o sinal do cepstro por uma função de "short pass liftering", l(n), da forma

$$l(n) = \begin{cases} 1 & n < \sigma_1 \\ \frac{1}{2}\left\{1 + \cos\left[\pi(n - \sigma_1)/\Delta\sigma\right]\right\} & \sigma_1 \le n < \sigma_1 + \Delta\sigma \\ 0 & n \ge \sigma_1 + \Delta\sigma \end{cases} \qquad (6.20)$$

em que $\sigma_1 + \Delta\sigma$ é menor que o período fundamental mínimo esperado.

A figura 6.13 mostra a função da janela com σ_1=1ms , $\Delta\sigma$=3ms, frequência de amostragem de 11.025 Khz e uma janela de 400 amostras determinada pela função *flifter*() desenvolvida em Matlab e cujo código se apresenta no anexo B5.

Quanto maior é $\sigma_1 + \Delta\sigma$ menos alisado será o espectro, sendo também verdadeiro o inverso. Uma correcta escolha destes parâmetros é portanto extremamente importante para se obterem bons resultados com o alisamento espectral. Esta importância é acrescida quando as formantes são directamente extraídas do espectro alisado.

Não se poderá deixar de referir que devido à componente da frequência fundamental aparecer separada no cepstro manifestando-se pelo aparecimento de um pico às quefrências correspondentes à duração do período fundamental, pode-se avaliar da vocalização ou não do segmento em causa pela existência ou não do referido pico, já que este aparece apenas para a fala vocalizada.

6.9 Espectrógrafo

Para estudar as características acústicas do sinal de fala várias técnicas são geralmente utilizadas. Uma classe de técnicas que tem tradicionalmente desempenhado um papel importante baseia-se nas representações tempo-frequência e, em especial no espectrograma. Por espectrograma designa-se uma imagem bidimensional que permite visualizar o conteúdo energético do sinal de fala no tempo e na frequência. O cálculo do espectrograma é feito utilizando um banco de filtros passa-banda centrados em frequências equiespaçadas, cobrindo toda uma gama de frequências com interesse para a análise do sinal. O sinal de fala é filtrado por cada um desses filtros e a amplitude da saída representada sob a forma de níveis de cinzento no ponto do plano tempo-frequência correspondente à frequência central do filtro e ao instante de tempo em que a amplitude é medida. Embora inicialmente fossem obtidos de forma analógica, os espectrogramas são actualmente calculados em computadores digitais, sendo necessário interpolar de alguma forma os valores do espectrograma referentes a localizações discretas no tempo e na frequência por forma a obter-se uma representação bidimensional contínua. Um parâmetro que altera profundamente as características do espectrograma, e consequentemente a sua função na análise do sinal de fala, é a largura de banda dos filtros utilizados. Duas soluções típicas são os espectrogramas de banda larga usando tipicamente filtros com largura de banda a -3 dB de 300 Hz e os espectrogramas de banda estreita usando larguras de banda de 45

Hz. Enquanto os espectrogramas de banda estreita permitem o seguimento fácil das harmónicas do sinal de fala em zonas vocalizadas, os espectrogramas de banda larga permitem o acompanhamento das formantes dos sinais de fala como se verifica na figura 6.14 para o sinal correspondente à locução "ama".

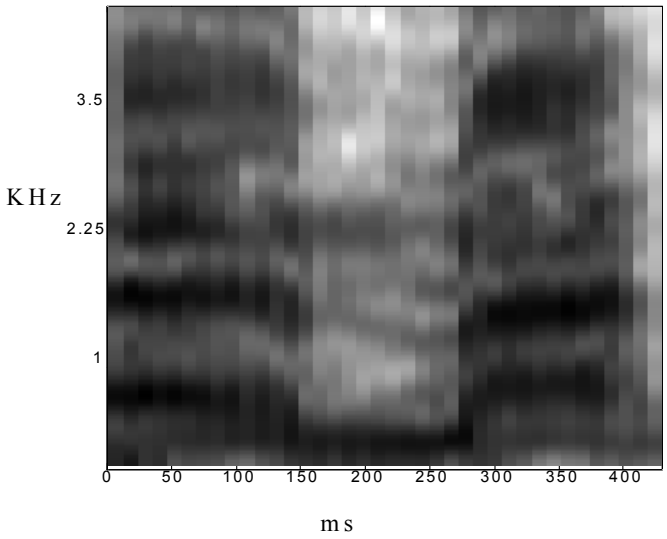


Figura 6.14 - Espectrograma de banda larga do sinal de fala "ama" obtido com alisamento espectral e a função *espectro()*.

A necessidade de apresentar e comparar sinais neste trabalho levou ao desenvolvimento de um espectrógrafo digital implementado computacionalmente.

Os espectrogramas de banda larga apresentados pela função *espectro()* desenvolvida no Matlab, cujo código vem no anexo B6 são baseados, não num banco de filtros passa-banda com as frequências centrais equiespaçadas e larguras de banda a -3 dB de 300 Hz, como é convencional, mas sim pela sequência de envelopes espectrais dos segmentos de sinais de fala obtidos pelo processo de análise cepstral apresentado no capítulo 4, ou pela sequência das funções de transferência do modelo do trato vocal dos sinais de fala obtidos pela análise por predição linear referida nos capítulos 4 e 7.

A função *espectro()* tem como parâmetros de entrada, a matriz com os valores de amplitude das frequências dos espectros alisados ou funções de transferência dos modelos ao longo do tempo, a frequência de amostragem, a escala do eixo do tempo e a escala do eixo das frequências.

6.10 Codificação por Predição Linear (LPC)

Tendo sido apresentado no capítulo 4 o desenvolvimento teórico do modelo de predição linear quer pelo método da matriz autocorrelação quer pelo método da matriz covariância, será agora referida a implementação computacional destes dois métodos aplicada aos sinais de fala.

A dimensão das matrizes correlação e covariância é de $p \times p$ em que p é o número de pólos para o modelo.

6.10.1 Matriz Autocorrelação

Esta matriz é construída no início da função *fcorprd2()* apresentada no anexo B7.

Já se mostrou que a matriz autocorrelação é uma matriz Toeplitz, isto é, simétrica e com todos os elementos iguais ao longo de uma diagonal paralela à diagonal principal.

Todos os elementos desta matriz são pertencentes aos elementos de ordem 0 a p, do vector de autocorrelação do sinal de entrada como mostra a expressão

$$\begin{bmatrix} R_n(0) & R_n(1) & R_n(2) & ... & R_n(p-1) \\ R_n(1) & R_n(0) & R_n(1) & ... & R_n(p-2) \\ R_n(2) & R_n(1) & R_n(0) & ... & R_n(p-3) \\ ... & ... & ... & ... & ... \\ ... & ... & ... & ... & ... \\ R_n(p-1) & R_n(p-2) & R_n(p-3) & ... & R_n(0) \end{bmatrix} \tag{6.21}$$

A função AR=xcorr(x) do programa Matlab realiza a autocorrelação do vector x de comprimento n no vector AR (par e centrado em n) de comprimento 2n-1. A ordenação dos elementos em AR é de 1 a 2n-1 em vez de -n a n como na definição da função autocorrelação por os vectores no Matlab não terem índices negativos. Assim os elementos que interessa reter, da autocorrelação de 0 a p, será no vector AR de n a n+p. Eliminando os n-1 primeiros elementos em AR ficam agora ordenados neste vector de 1 a p+1 correspondendo no vector de autocorrelação aos elementos de 0 a p.

Seguidamente é construída a matriz de autocorrelação atendendo à expressão

$$\mathrm{ACORR}(i,j) = \mathrm{AR}(|i-j|+1) \qquad \begin{matrix} 1 \le i \le p \\ 1 \le j \le p \end{matrix} \tag{6.22}$$

Os elementos do vector do lado direito do sinal de igualdade da equação 4.50

$$\begin{bmatrix} R_n(1) \\ R_n(2) \\ R_n(3) \\ ... \\ ... \\ R_n(p) \end{bmatrix} \tag{6.23}$$

são os elementos do vector AR de 2 a p+1 respectivamente.

6.10.2 Matriz Covariância

Esta matriz é construída no início da função *fcovpr2()* apresentada no anexo B8.

Na equação 4.56 já foi apresentada a matriz covariância que se repete aqui como sendo

$$\begin{bmatrix} \phi_n(1,1) & \phi_n(1,2) & \phi_n(1,3) & \dots & \phi_n(1,p) \\ \phi_n(2,1) & \phi_n(2,2) & \phi_n(2,3) & \dots & \phi_n(2,p) \\ \phi_n(3,1) & \phi_n(3,2) & \phi_n(3,3) & \dots & \phi_n(3,p) \\ \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots \\ \phi_n(p,1) & \phi_n(p,2) & \phi_n(p,3) & \dots & \phi_n(p,p) \end{bmatrix} \tag{6.24}$$

em que

$$\phi_n(i,j) = \sum_{m=0}^{n-1} s_n(m-i)s_n(m-j) \qquad \begin{matrix} 1 \le i \le p \\ 1 \le j \le p \end{matrix} \tag{6.25}$$

Como foi visto em 4.5.2 o cálculo de $\phi_n(i,j)$ para todos os valores de i e j são necessários os valores de $s_n(m)$ no intervalo $-p \le m \le c-1$. No entanto, como é sabido os valores do sinal a analisar, x, existem apenas entre 1 e c, então é necessário recorrer a um artifício de cálculo que consiste em considerar o comprimento c do sinal x, apenas como o verdadeiro comprimento-p, sendo o primeiro elemento de $s_n(m)$ o elemento do vector x de ordem p+1 e os elementos à sua esquerda entre 1 e p inclusivé, considerados como sendo os elementos de -p a 0 de $s_n(m)$ (figura 6.15).

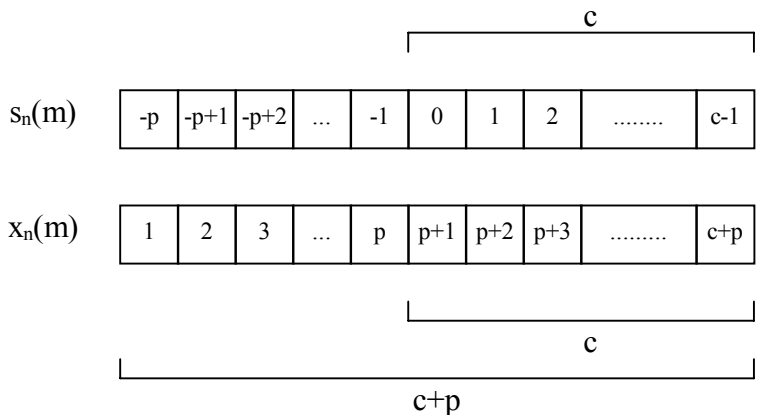


Figura 6.15 - Artifício de cálculo para a determinação da matriz covariância.

Correspondendo à expressão

$$\phi_n(j,i) = \mathrm{SUM}(x(p+1-j:\mathrm{length}(x)-j).\times x(p+1-i:\mathrm{length}(x)-i)) \qquad \begin{matrix} j = 1,\ 2,\ \dots,\ p \\ i = 1,\ 2,\ \dots,\ p \end{matrix} \tag{6.26}$$

Os elementos do lado direito do sinal de igualdade da equação 4.56

$$\begin{bmatrix} \phi_n(1,0) \\ \phi_n(2,0) \\ \phi_n(3,0) \\ \dots \\ \dots \\ \phi_n(p,0) \end{bmatrix} \tag{6.27}$$

são obtidos com a mesma equação 6.26 com i=0.

6.11 Sintetizador

Foi desenvolvido computacionalmente um sintetizador simples com a finalidade de testar os parâmetros dos sinais analisados. Este sintetizador obedece a um modelo para sons vocalizados de 4 formantes com implementação série.

O programa em Matlab que realiza a síntese chama-se *sintese.m* e é apresentado no anexo B9. Este programa recebe os parâmetros do modelo do sinal em vectores (4 formantes, 4 larguras de banda, amplitude, frequência fundamental e duração). A frequência de amostragem é alterável. Baseado na frequência de amostragem, frequência fundamental e no comprimento da janela a função *fgerimp()*, anexo B10, gera um segmento, com o mesmo comprimento da janela[i], de impulsos glotais. A geração de impulsos glotais seria pacífica se se pretendesse um sinal de duração contínua, contudo neste caso não é o que acontece, já que se quer gerar impulsos glotais em segmentos de um comprimento finito e deseja-se que o próximo segmento comece no ponto seguinte ao que terminou o segmento anterior para não causar descontinuidades na amplitude do sinal de impulso glotal ao longo da sequência de segmentos. Para solucionar este problema a função *fgerimp()* gera não só o sinal do impulso glotal como também um sinal de resto que contém o final do último impulso glotal que excede o segmento. Este resto será o início do próximo segmento de impulsos glotais. A figura 6.16 mostra uma sequência de segmentos de impulsos glotais onde se pode verificar a continuidade do sinal entre segmentos. Os impulsos glotais são baseados na equação 4.12 que modela o impulso glotal.

Seguidamente cada um destes segmentos é multiplicado pelo ganho, Av, e submetido a uma sequência de 4 filtros que modelam os 4 formantes e respectivas larguras de banda de acordo com o modelo da equação 4.7. Depois, são sujeitos a um filtro que modela o efeito de radiação nos lábios de acordo com a expressão 4.8. Finalmente, cada segmento é multiplicado por uma função de janela Hanning e sobreposto em 50% ao final do vector que constituirá o sinal sintetizado. O procedimento de multiplicação por uma função de janela e sobreposição é realizado para minimizar possíveis defeitos em alguns segmentos que são excitados pelos filtros somente a partir do aparecimento do primeiro impulso glotal. O sinal sintetizado pode ser gravado em ficheiros do tipo WAV para ser ouvido com a placa de som usada (SOUND BLASTER).

[i] A duração temporal de cada segmento é metade do número de amostras vezes o intervalo entre amostras devido ao uso de sobreposição de 50%.

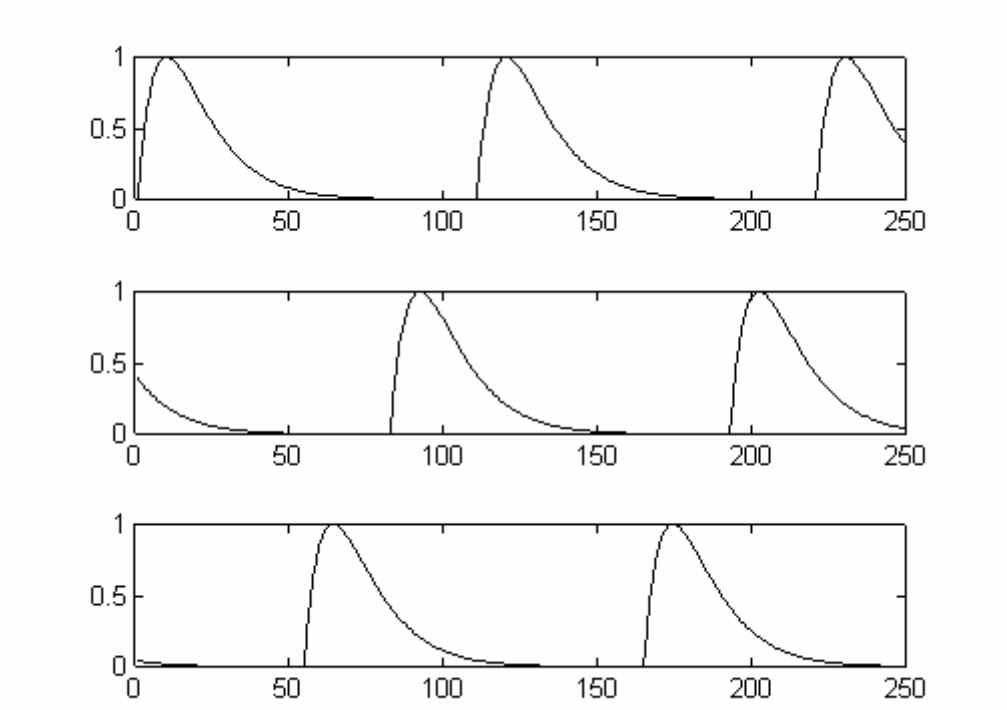


Figura 6.16 - Sequência de segmentos de impulsos glotais gerados pela função *fgerimp()*.

Porque não é possível aqui ouvir um desses sinais, ainda no programa síntese, é alisado o espectro do sinal sintetizado pelo método do cepstro e com recurso à função espectro(), já apresentada, é realizado o espectrograma do sinal. Assim mostra-se na figura 6.17b o espectrograma de um hipotético sinal sintetizado com uma frequência fundamental contínua de 100 Hz e ganho Av=1. A variação das frequências dos formantes é apresentada na figura 6.17a com larguras de banda constantes: B1=30, B2=50, B3=60, B4=90 Hz.

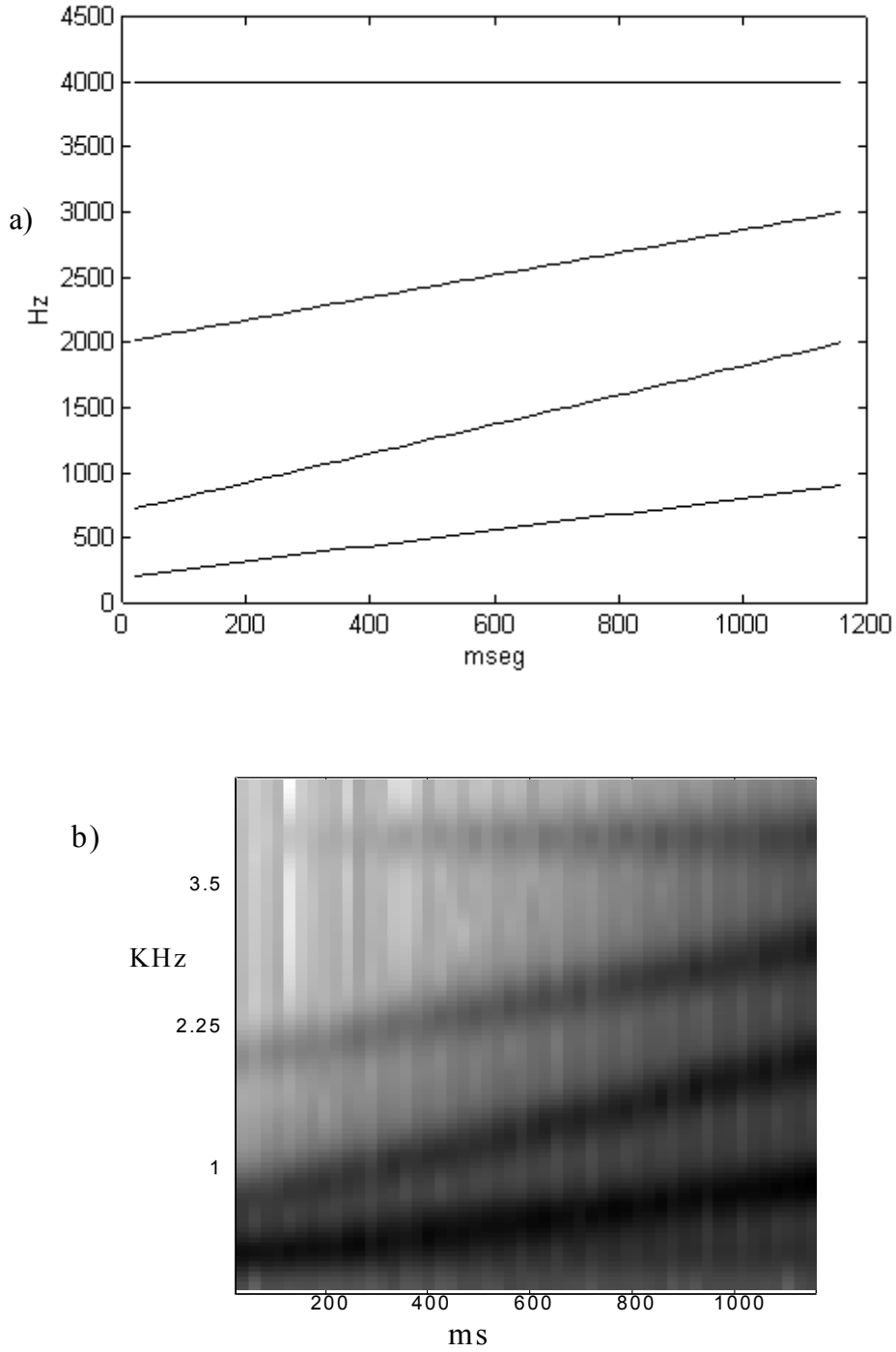


Figura 6.17 - Sinal sintetizado. a) Variação da frequência dos 4 formantes ao longo do tempo; b) espectrograma.

CAPÍTULO 7

DETERMINAÇÃO AUTOMÁTICA DOS PARÂMETROS DOS SINAIS DE FALA

7. DETERMINAÇÃO AUTOMÁTICA DOS PARÂMETROS DOS SINAIS DE FALA

7.1 Introdução

Este capítulo pode ser entendido como o culminar ou o ponto de chegada de todo o estudo desenvolvido nesta dissertação. Aqui far-se-á uso das ferramentas atrás descritas, aplicadas aos sinais adquiridos, recorrendo aos modelos discutidos para extracção da sequência de parâmetros que podem ser usados em sistemas de conversão texto-fala.

O desenvolvimento, implementação e optimização de alguns dos algoritmos aqui descritos tomou grande parte do tempo e esforço dedicado a este trabalho.

Assim o estudo dedicado à determinação da frequência fundamental reveste-se de um elevado interesse já que é a variação deste parâmetro, frequência fundamental, ou o seu inverso, período fundamental, que determina a entoação de um qualquer texto ou frase. Realça-se mais uma vez que a correcta entoação e prosódia num conversor texto-fala são o elemento da mais valia para estes sistemas, que lhes permitirão ser aceites com agrado por parte das pessoas ouvintes.

O sub-capítulo dedicado à extracção automática de parâmetros dos sinais de fala é o fulcro desta dissertação que tinha como objectivo inicial o estudo, desenvolvimento e aplicação de modelos matemáticos para análise de sinais de fala e concepção de um conjunto consistente de ferramentas instrumentais vocacionadas para a extracção dos parâmetros a usar em síntese de fala. Foi dedicada especial atenção aos sinais de fala vocalizada por estes serem os sons que mais influência têm na qualidade da fala sintetizada. Foram desenvolvidos com elevado acuidade três métodos para extracção automática dos parâmetros do modelo para fala vocalizada (método do cepstro e LPC's pelas matrizes autocorrelação e covariância), sem se deixar de tocar o método de análise síncrona com o período fundamental nem o estudo de fala não vocalizada.

Os três primeiros métodos são incorporados no mesmo corpo de programa, distinto do corpo de programa usado para a análise síncrona com o período fundamental, devido aos diferentes métodos de segmentação usados. Nos três primeiros casos foram usados segmentos de comprimento fixo com sobreposição de 50% enquanto no último caso a segmentação é pendente do início e duração de cada impulso glotal.

Um sistema completo de análise de sinais de fala deveria usar o modelo mais adequado a cada tipo de sinal de fala para analisar e modelizar parametricamente cada segmento previamente classificado do sinal de fala. Isto exige sistemas para extracção automática dos parâmetros dos modelos considerados para cada tipo de sinais de fala (vocalizada, não vocalizada e excitação mista). Neste trabalho foram desenvolvidos algoritmos para extracção dos parâmetros dos modelos de fala vocalizada, reconhecendo-se a necessidade de estudo mais profundo de sistemas para extracção dos parâmetros dos modelos de fala não vocalizada e de excitação mista.

O sintetizador desenvolvido neste trabalho e discutido no capítulo anterior foi usado no âmbito deste capítulo para sintetizar a sequência de parâmetros extraídos automaticamente dos sinais de fala para comparação dos resultados, bem como para sintetizar uma sequência de parâmetros com uma variação conhecida e analisar o sinal

sintetizado para testar a qualidade dos resultados obtidos, por comparação dos parâmetros extraídos automaticamente por análise com os parâmetros sintetizados.

7.2 Determinação da Frequência Fundamental

7.2.1 Introdução

O estudo da frequência fundamental dos sinais de fala tanto na síntese como na análise tem uma importância relevante ao nível supra-segmental.

Já foi bastas vezes referida a crescente importância da prosódia nos sinais de fala para que os sistemas de análise e síntese deixem de produzir fala robotizada, sendo a natural variação da frequência fundamental um parâmetro relevante para o estudo e imposição da correcta entoação.

Durante o trabalho foram estudadas e implementadas algumas das mais importantes técnicas para determinação automática da frequência fundamental com processamento tanto no domínio das frequências como no domínio temporal.

Nesta secção faz-se referência à determinação deste parâmetro com recurso ao cepstro aproveitando o facto de a técnica do cepstro ser usada para extracção automática dos formantes. A utilização de técnicas de predição linear neste trabalho leva a que se faça uma referencia obrigatória da determinação deste parâmetro com recurso a este processamento. Finalmente é apresentada uma técnica com processamento exclusivamente no domínio temporal baseada no algoritmo de [Gold 69].

7.2.2 Método do Cepstro

Já foi referido nos capítulos 4 e 6 a capacidade do cepstro separar os efeitos da fonte e do meio de transmissão. A aplicação do cepstro aos sinais de fala vocalizada, como se pode ver na fig. 4.15 referente ao cepstro do sinal de fala da vogal [a], leva a que os efeitos da fonte (frequência fundamental) apareçam separados dos efeitos do trato vocal, uma vez que constituem famílias harmónicas diferentes e consequentemente situam-se a quefrências diferentes. Quando o cepstro é aplicado a segmentos de sinais de fala vocalizada com mais de um período fundamental, aparece evidenciado no gráfico de quefrências um pico ao valor de quefrência igual à duração do período fundamental.

A determinação do período fundamental pelo método do cepstro não é mais do que determinar para cada segmento o valor de quefrências a que este pico ocorre. Este deve ser procurado acima de uma valor considerado mínimo para a ocorrência do período fundamental, pois para os valores mais baixos das quefrências do espectro existem outros picos maiores mas que dizem respeito ao trato vocal. O inverso do período fundamental é a frequência fundamental relativa ao segmento em causa.

O programa de determinação automática dos parâmetros com recurso à análise cepstral, *fformvo2*, apresentado adiante e cujo código vem no anexo B13, também determina automaticamente ao longo dos sinais de fala a frequência fundamental por este método.

A figura 7.1 mostra a sequência de valores da frequência fundamental ao longo do sinal de fala correspondente à locução da palavra "mola" pelo locutor 1, determinada pelo programa *fformvo2* com recurso ao método do cepstro.

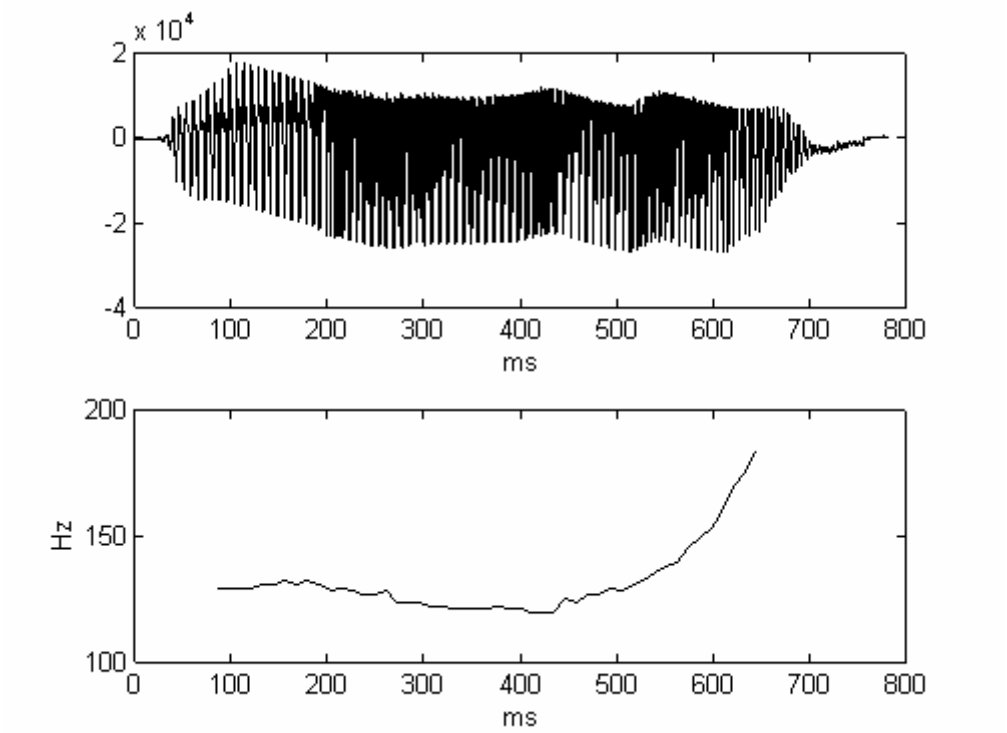


Figura 7.1 - Variação da frequência fundamental. Em cima o sinal de fala correspondente à locução de "mola" pelo locutor 1; Em baixo a variação da frequência fundamental ao longo do sinal.

7.2.3 Determinação da Frequência Fundamental no Erro Residual de Predição Linear

Quando a análise de sons vocalizados é realizada com recurso ao método da predição linear pode-se determinar a frequência fundamental analisando o sinal do erro residual já que teoricamente este sinal é a excitação do sistema a menos de um factor de ganho G.

O modelo só com pólos considerado pode ser representado pela figura 7.2.

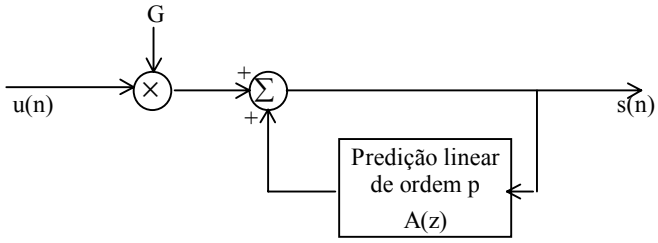


Figura 7.2 - Modelo de predição linear só com pólos.

Sendo

$$A(z) = 1 - \sum_{k=1}^{p} \alpha_k z^{-k} \qquad (7.1)$$

e a função de transferência

$$H(z) = \frac{G}{A(z)} \tag{7.2}$$

Para o sistema da figura 7.2 as amostras do sinal de fala s(n) estão relacionadas com o sinal de excitação u(n) pela equação diferencial

$$s(n) = \sum_{k=1}^{p} a_k s(n-k) + Gu(n) \tag{7.3}$$

A saída do sistema de predição linear com os coeficientes α_k é

$$\tilde{s}(n) = \sum_{k=1}^{p} \alpha_k s(n-k) \tag{7.4}$$

Sendo o erro residual de predição definido por

$$e(n) = s(n) - \tilde{s}(n) = s(n) - \sum_{k=1}^{p} \alpha_k s(n-k) \tag{7.5}$$

Se o sinal de fala obedece ao modelo e $\alpha_k=a_k$, então substituindo a_k por α_k em 7.3 e substituindo s(n) em 7.5 fica

$$e(n) = Gu(n) \tag{7.6}$$

Então a determinação da frequência fundamental no sinal de erro residual para a fala vocalizada consiste em detectar os picos mais salientes dentro de um espaçamento considerado razoável para o período fundamental. O correcto espaçamento entre picos corresponde ao período fundamental e o seu inverso à frequência fundamental.

7.2.4 Processamento no Domínio Temporal

Apesar de já terem sido apresentados dois métodos para a determinação da frequência fundamental, nunca é demais a discussão de outro, já que não se conhecem esquemas sem limitações e, como afirma [Rowden 92] não se pode esperar presentemente um detector de frequência fundamental com resultados perfeitamente satisfatórios para um grande número de falantes, aplicações e ambientes de recolha do sinal.

Este método para determinação da frequência fundamental, recorre a um processamento exclusivamente no domínio temporal. Os seus resultados podem ser bons para diferentes falantes e ambientes de recolha do sinal.

O esquema de processamento deste método pode ser facilmente implementado em "hardware" levando a uma solução com resultados em tempo real.

Este algoritmo proposto por [Gold 69] baseia-se nos seguintes princípios:

1. O sinal de fala é tratado para criar um número de sequências de impulsos que retenham a periodicidade do sinal original e se descarte das características irrelevantes para o processo de detecção do período fundamental.

2. Este processamento permite que simples detectores do período fundamental possam ser usados para estimar o período de cada sequência de impulsos.

3. A estimativa de vários destes detectores simples do período fundamental é logicamente combinada para determinar o período da forma de onda da fala.

O esquema usado é representado pela figura 7.3.

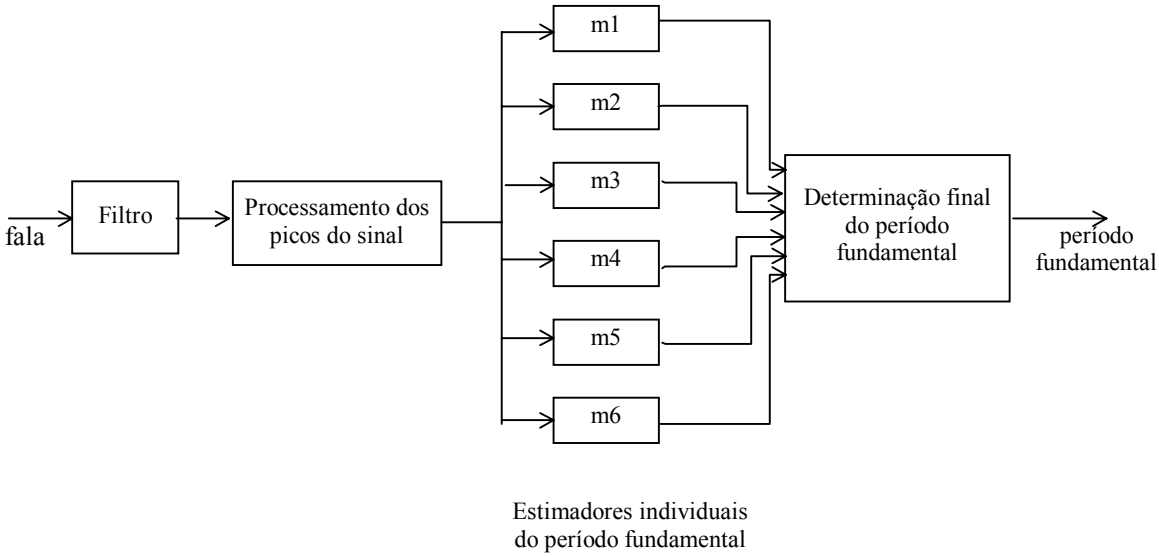


Figura 7.3 - Diagrama de blocos do processamento paralelo no domínio temporal do detector de frequência fundamental.

A forma de onda da fala deve ser amostrada a uma frequência suficiente para dar uma resolução temporal para o período fundamental adequada. A amostragem a 11.025 Khz usada permite uma resolução do período fundamental de 0,09 ms.

O primeiro bloco consiste num filtro passa baixo para alisar o sinal de fala com uma frequência de corte cerca dos 900 Hz. Em alguns casos é necessário um filtro passa banda com frequências de passagem entre os 100 e os 900 Hz para remover a componente DC e possível ruído da rede (50 Hz) do sinal. Este bloco é implementado recorrendo mais uma vez à amplitude média deslizante com uma janela de sensivelmente metade do período fundamental previsto e um espaçamento de 1 amostra (quanto maior for o espaçamento, menor será a resolução do período fundamental), seguido pela passagem do sinal pela função detrend() do Matlab para eliminar a componente DC residual.

A seguir à filtragem são localizados os "picos e vales" (máximos e mínimos locais) bem como as suas amplitudes, e derivadas algumas sequências de impulsos do sinal filtrado. Cada sequência é composta por impulsos de amplitude positiva na localização dos picos e vales. Os seis casos propostos e usados são:

1. m1(n): Ocorrência de um impulso na localização de cada pico com a mesma amplitude deste.

2. m2(n): Ocorrência de um impulso na localização de cada pico com amplitude igual à diferença de amplitudes entre o pico e o vale anterior.

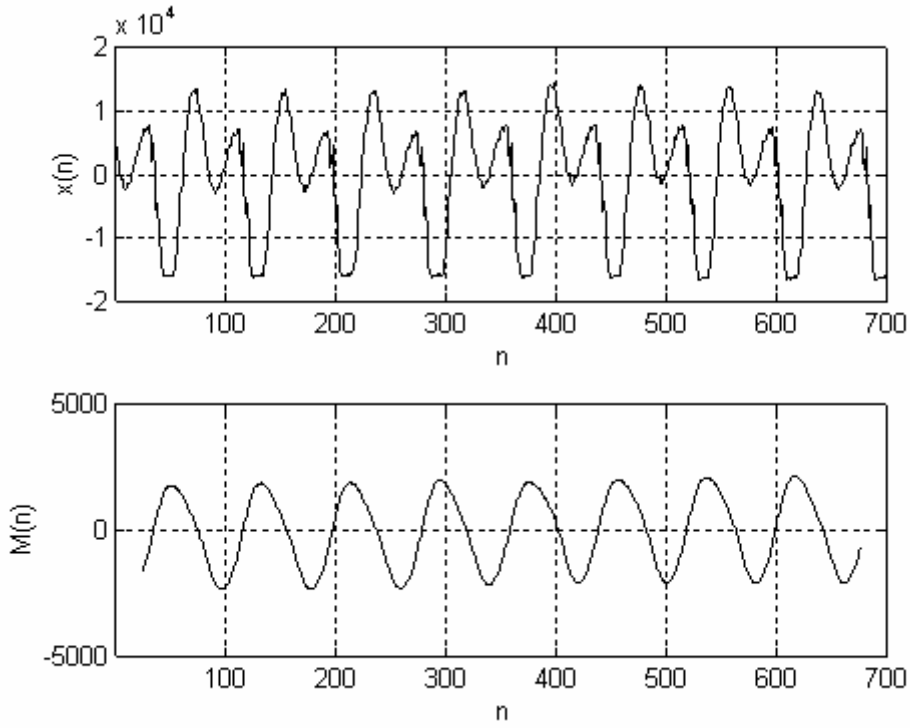
3. m3(n): Ocorrência de um impulso na localização de cada pico com amplitude igual à diferença entre as amplitudes do próprio pico e do pico anterior. Se esta diferença for negativa o impulso terá amplitude nula.

4. m4(n): Ocorrência de um impulso na localização de cada vale com amplitude simétrica da amplitude do vale.

5. m5(n): Ocorrência de um impulso na localização de cada vale com amplitude igual ao simétrico da amplitude do próprio vale, mais a amplitude do pico anterior.

6. m6(n): Ocorrência de um impulso na localização de cada vale com amplitude igual ao simétrico da amplitude do próprio vale, mais a amplitude do vale anterior. Se esta diferença for negativa o impulso terá amplitude nula.

A figura 7.4 mostra um exemplo para as seis sequências m1 a m6 aplicadas ao sinal [i] alisado.



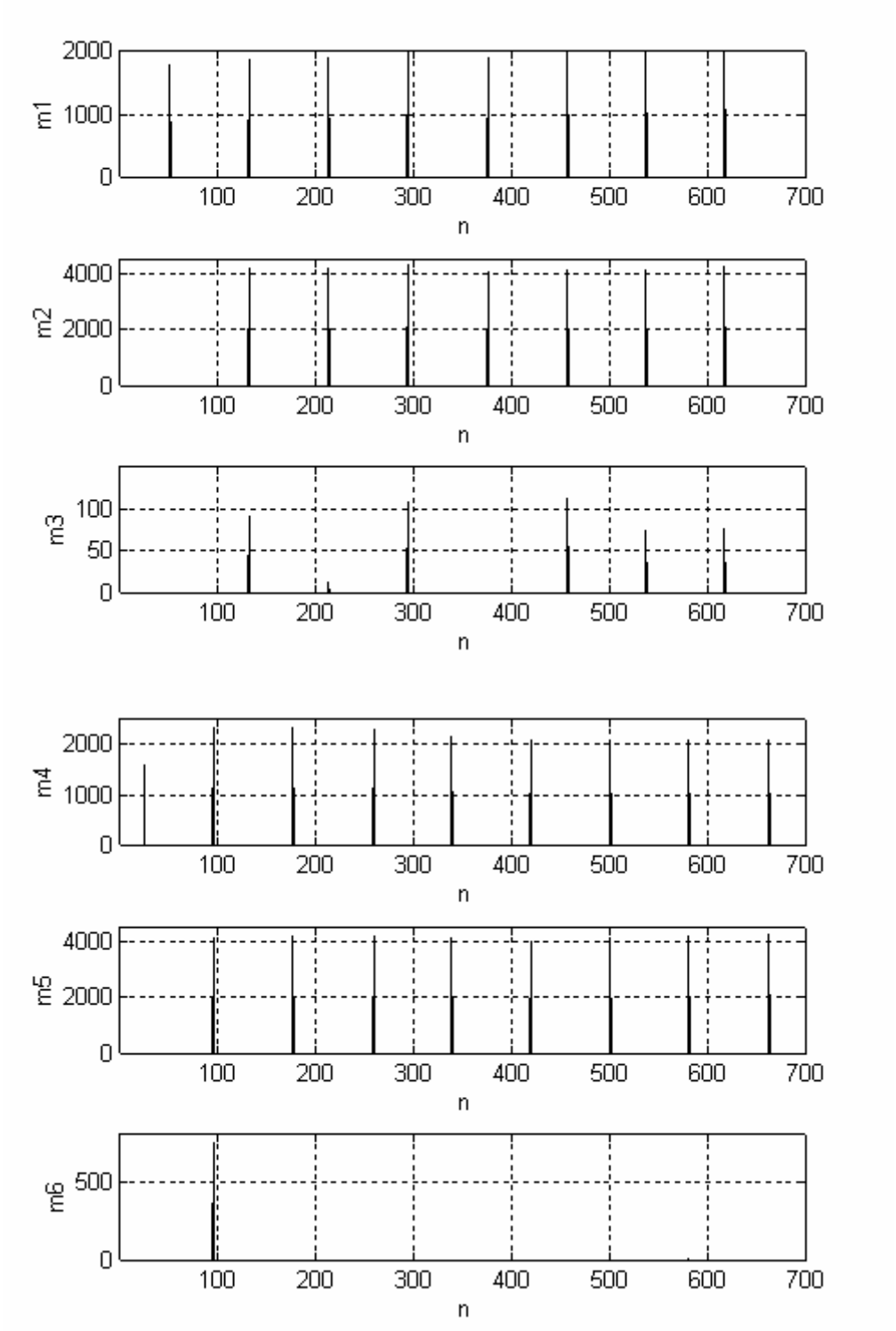


Figura 7.4 - As seis sequências m1 a m6 de impulsos obtidas a partir do sinal [i] alisado com uma janela de média de 80 amostras e espaçamento unitário.

O propósito destas seis sequências de impulsos é realizar estimativas simples do período numa curta duração do sinal. Seguidamente, algumas sequências de impulsos são processadas por um sistema não linear variante no tempo. Quando um impulso de amplitude suficiente é detectado na entrada do sistema a sua saída toma o valor da amplitude do impulso e mantém-se durante um intervalo τ durante o qual nenhum

impulso pode ser detectado. No final do intervalo τ a saída inicia um decaimento exponencial. Quando um impulso excede o nível do decaimento da saída o processo é repetido. A razão do decaimento e o intervalo τ são dependentes da estimativa do último período fundamental. Estes valores são críticos para um bom funcionamento do sistema. O intervalo τ foi estabelecido como sendo 1/3 da duração do último período estimado. O resultado deste processo é uma espécie de alisamento da sequência de impulsos produzindo uma sequência quase periódica de impulsos, como mostra a figura 7.5. A média da largura de cada impulso é estimada como sendo o período fundamental estimado pelo estimador em causa.

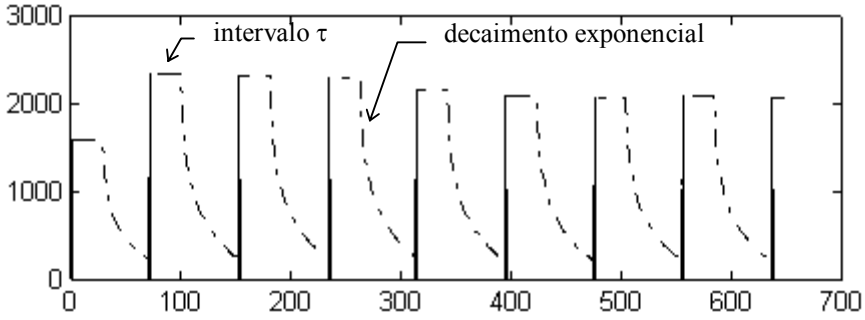


Figura 7.5 - Operação básica de cada estimador individual do período fundamental no domínio temporal.

Na implementação deste algoritmo, esta técnica foi aplicada a 4 estimadores (m1, m2, m4 e m5).

Aos estimadores m3 e m6 foi aplicada uma operação diferente. A partir da posição de cada estimativa de impulso é procurado novo impulso no intervalo entre sensivelmente 85% e 115% do último período fundamental estimado, para diante. A média da largura destes impulsos é também estimada como sendo o novo período fundamental estimado pelo estimador em causa.

Os resultados dos estimadores m3 e m6 são normalmente idênticos aos dos outros estimadores. Mas crê-se que a riqueza de variação de estimadores é salutar, pois quando alguns falham em sinais menos claros, estimadores diferentes podem dar bons resultados. Isto acontece para estes estimadores com operações diferentes em alguns sinais.

Na figura 7.6 apresentam-se resultados dos seis estimadores para a vogal [i] pronunciada pelo locutor 2. Os valores obtidos pela frequência fundamental são confirmados quando se utiliza o método do cepstro tanto para este exemplo como para outros sinais.

A função desenvolvida em Matlab que implementa este método apresenta-se no anexo B11 com o nome *fpitch3()*.

Conhecidas as seis estimativas para a frequência fundamental e o seu valor anterior, basta agora criar um processo de decisão para chegar ao valor mais provável da frequência fundamental, quando as estimativas diferem umas das outras. Este processo de decisão não foi no entanto implementado nem usado neste trabalho por imposições de tempo.

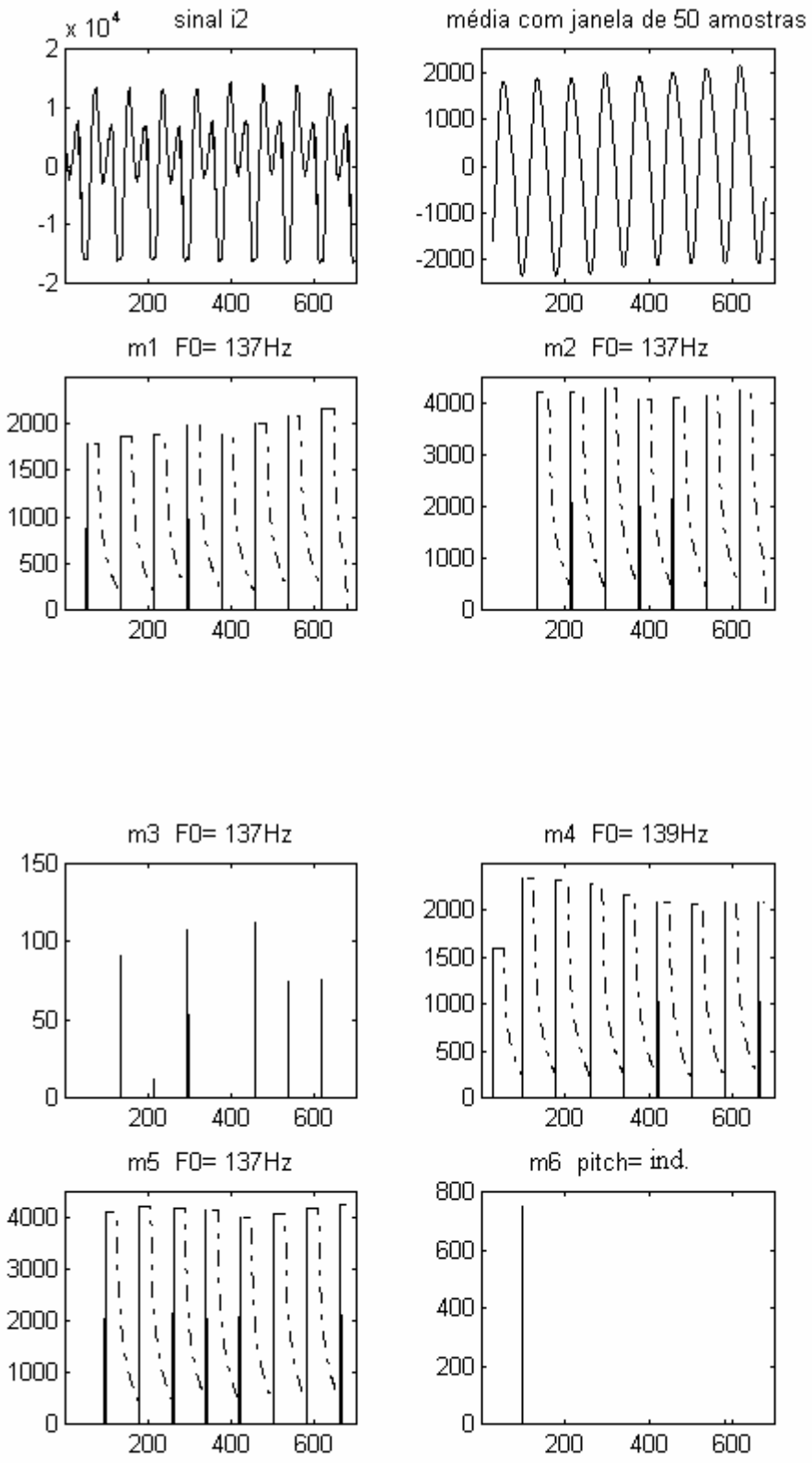


Figura 7.6 - Frequência fundamental estimada pelos seis estimadores para o sinal [i].

7.3 Determinação Automática dos Parâmetros Para a Fala Vocalizada

Foram estudados e implementados quatro métodos diferentes (cepstro, LPC (Linear Predictive Code) pela matriz autocorrelação, LPC matriz covariância e análise síncrona com o período fundamental pelo método de predição linear), para extracção automática dos formantes, larguras de banda, frequência fundamental, e amplitude. Em todos os métodos foi considerado um modelo de 4 formantes.

Para os três primeiros casos é idêntico o processamento a termo longo, segmentação do sinal em troços de um comprimento fixo com sobreposição de 50%, , ao contrário do método de análise síncrona em que os segmentos são determinados pelo período fundamental do sinal.

Assim, foi criada uma estrutura para o processamento dos métodos implementados com segmentos de comprimento fixo. Esta estrutura realiza a segmentação do sinal considerado vocalizado, sendo os parâmetros extraídos segmento a segmento por uma função que implementa o método previamente escolhido para a análise realizando depois, um pós-processamento no sentido de corrigir possíveis erros na sequência de parâmetros do sinal de fala a longo termo.

O cálculo da frequência fundamental pelos métodos atrás referidos, bem como da amplitude do sinal é também realizado segmento a segmento pela função de extracção dos parâmetros.

7.3.1 Estrutura de Análise

Foi desenvolvida uma estrutura de análise para os sinais de fala vocalizada que engloba o método de análise cepstral, e a codificação por predição linear pelos métodos das matrizes autocorrelação e covariância.

Este desenvolvimento foi realizado em Matlab e o código do programa é apresentado no anexo B12 com o nome *falacont.m*.

A figura 7.7 mostra o fluxograma da estrutura de análise para os sinais vocalizados implementada no programa *falacont.m*.

O programa permite ao utilizador escolher o sinal a analisar de um dos sinais previamente armazenados. Se a frequência desse sinal for superior a 20 Khz (22.05 Khz para os sinais adquiridos) realiza uma decimação 2:1 (com filtragem anti-aliasing), cuja função do filtro foi analisada no capítulo 3. O sinal é então apresentado ao utilizador para que este escolha o início e o fim da parte do sinal que pretende analisar. Seguidamente o utilizador escolhe qual dos métodos de análise deseja usar (análise cepstral, predição linear pelo método da matriz autocorrelação ou predição linear pelo método da matriz covariância). Depois, começando na amostra que foi escolhida para início da análise e até à amostra correspondente ao fim da análise é iniciada a sequência de selecção de segmentos com um comprimento pré-definido, análise dos segmentos pelo método seleccionado e armazenamento dos parâmetros de cada segmento bem como do seu envelope espectral. Mais pormenores relativamente a cada método de análise são descritos nas secções seguintes. É também determinada a energia do segmento do sinal recorrendo à expressão

$$E = \frac{1}{(\text{fim} - \text{inicio} + 1)} \sum_{m=\text{inicio}}^{\text{fim}} x^2(m) \tag{7.7}$$

em que inicio e fim são respectivamente o primeiro e último elemento do segmento do sinal.

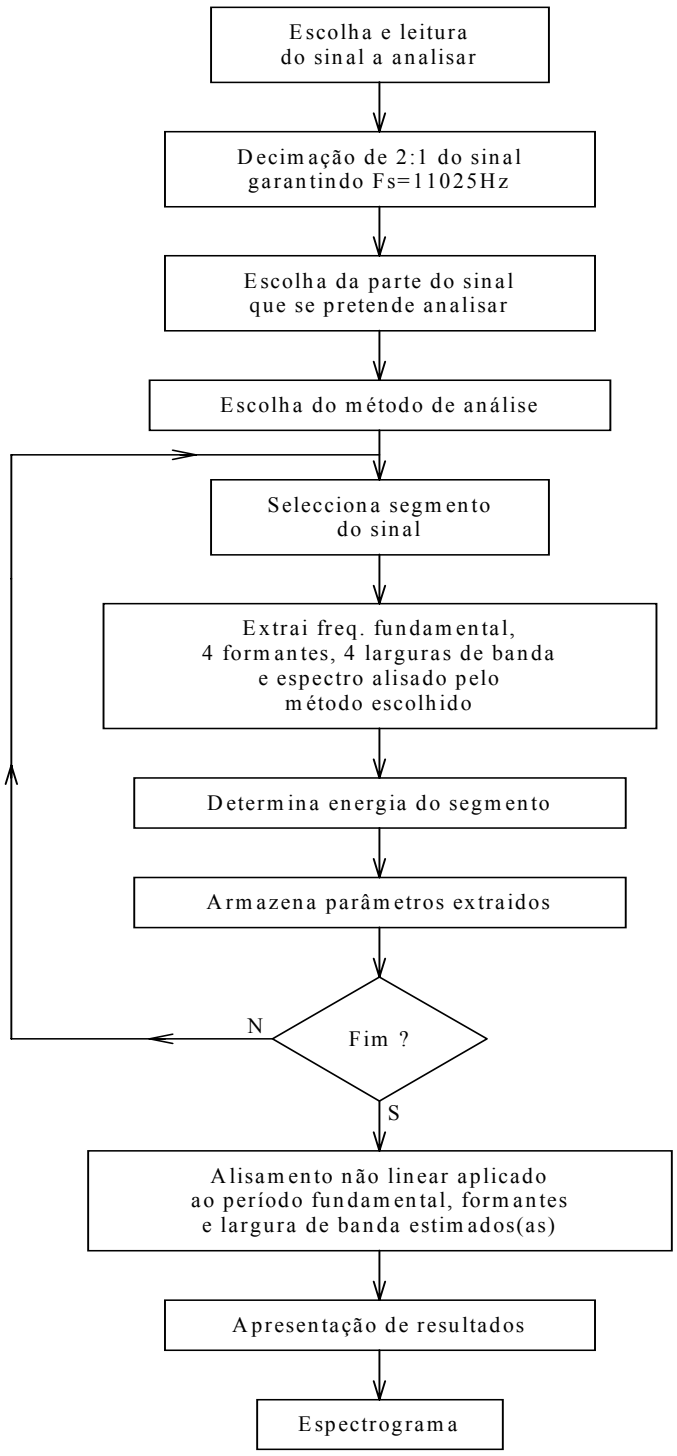


Figura 7.7 - Fluxograma da estrutura de análise para o modelo de sinais vocalizados.

Finalmente, para a sequência relativa a cada parâmetro extraído automaticamente (excepto a energia) ao longo do sinal é aplicado um alisamento não linear para corrigir eventuais pontos fora da trajectória estimada para esse parâmetro. Remete-se a discussão deste alisamento não linear para uma secção adiante.

Uma vez concluída a análise do sinal de fala são apresentados graficamente ao longo do tempo os valores para a frequência fundamental, formantes e larguras de banda,

bem como o espectrograma determinado pela sequência de envelopes espectrais e apresentado pela função *espectro()* já anteriormente apresentada.

7.3.1.1 Método de Análise Cepstral

O sistema implementado, testado e aqui apresentado de extracção automática dos 4 formantes, respectivas larguras de banda e frequência fundamental dos sinais de fala vocalizada pelo método de análise cepstral é baseado no sistema desenvolvido por [Schafer 70] para extracção do período fundamental e das 3 primeiras frequências formantes.

O processamento descrito nesta secção para a extracção da frequência fundamental, dos 4 formantes e respectivas larguras de banda implementado na função *fformvo2()* em Matlab é apresentado no anexo B13. Esta função serve o corpo principal do programa de extracção de parâmetros para a fala vocalizada, sendo chamada para extrair esses parâmetros segmento a segmento.

A estimação da frequência fundamental e alisamento do espectro logarítmico é obtida com recurso à análise cepstral por processos já descritos anteriormente. As frequências formantes são estimadas no envelope espectral usando constrições das gamas de frequências dos formantes e níveis relativos das amplitudes dos picos espectrais às frequências formantes. Estas constrições permitem a detecção de casos em que dois formantes têm frequências tão próximas que não permitem a sua determinação no envelope espectral inicial. Nestes casos, um novo processamento de análise espectral ("Chirp Z Transform"), permite a computação eficiente de um novo espectro de banda mais estreita para a resolução das frequências dos formantes.

O alisamento do envelope espectral e determinação da frequência fundamental é realizado segundo o diagrama de blocos da figura 7.8.

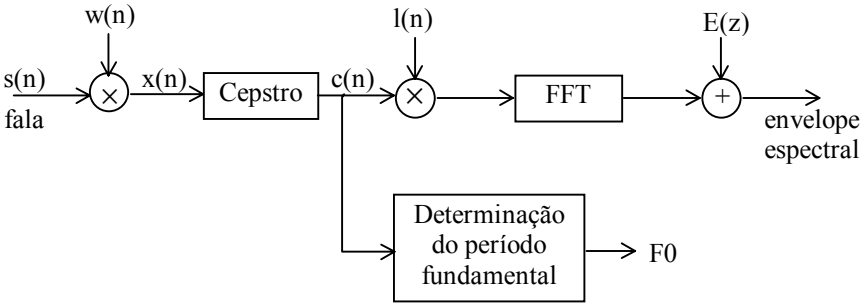


Figura 7.8 - Diagrama de blocos para determinação da frequência fundamental e envelope espectral.

O sinal de fala, s(n), é convoluido com uma função de janela temporal (melhores resultados com a janela de Hamming, embora, se obtenham também bons resultados com a janela de Hanning). Para o sinal resultante, x(n), é calculado o cepstro, cuja fundamentação já foi desenvolvida nos capítulos 4 e 6. O cepstro, c(n), é usado para determinação da frequência fundamental, F0, como foi discutido no início deste capítulo. Ainda a este sinal, c(n), é aplicada a função de "lifter" l(n), que realiza uma "lifteragem" linear, também já discutida no capítulo anterior, com a finalidade de remover do cepstro as componentes de mais altas quefrências relativas à frequência

fundamental. O sinal resultante é passado para o domínio das frequências pela aplicação da FFT, obtendo-se o espectro alisado. Contudo, este espectro é relativo ao trato vocal e aos efeitos de radiação nos lábios bem como aos efeitos de modelação da forma de onda glotal. Na tentativa de anular estes efeitos é somada a função de equalização das frequências E(z), com a forma da figura 7.9, cuja função de transferência é apresentada na expressão 7.8.

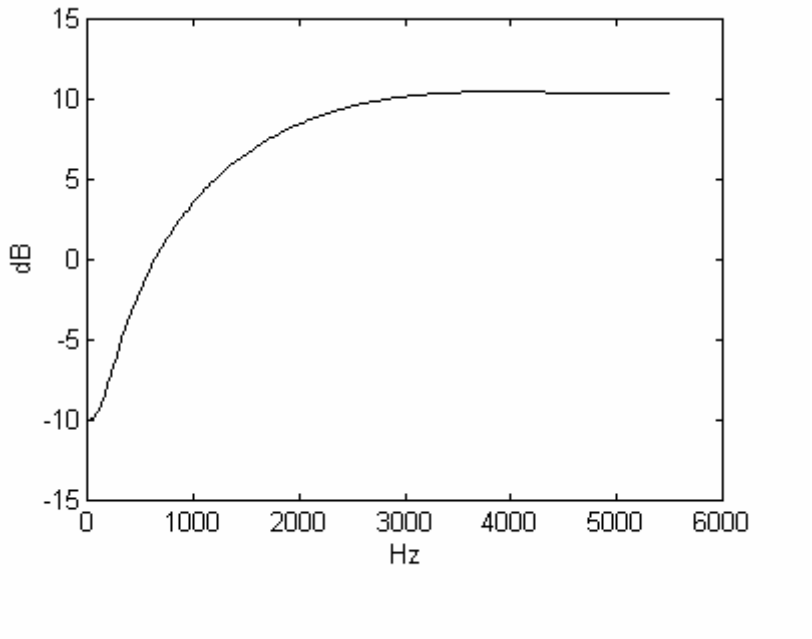


Figura 7.9 - Função de equalização das amplitudes dos formantes.

$$E(z) = \left(\frac{1 - e^{-aT} z^{-1}}{1 - e^{-aT}}\right)\left(\frac{1 + e^{-bT} z^{-1}}{1 + e^{-bT}}\right) \tag{7.8}$$

Esta função visa, segundo [Flanagan 64], equalizar a amplitude dos formantes, removendo aproximadamente as contribuições da forma de onda glotal e do efeito de radiação.

A função, realizada em Matlab, que implementa esta função de equalização é apresentada no anexo B14 com o nome *equaliz1()*.

Os valores a e b dependem do aparelho fonador do falante que produz a fala a analisar e da frequência de amostragem. Neste trabalho, para uma frequência de amostragem de 11.025 Khz foram usados a=882π e b=11025π. Estes valores foram escolhidos de forma a que a função seja centrada em torno de 0 dB, valendo aos 0 Hz -XDB e aos 4000 Hz +XDB (foi usado XDB=10).

Depois de adicionada a função E(z) ao espectro alisado pelo cepstro obtém-se finalmente a envolvente espectral do trato vocal.

As frequências dos formantes são obtidas por um processo de "pegar os picos" no envelope espectral impondo constrições às gamas de frequências dos formantes e suas amplitudes relativas.

No que diz respeito às gamas de frequência para os formantes não existe um consenso entre os autores de diversos sistemas de análise e de síntese que usam também esta constrição. As regiões de frequências aqui adoptadas foram: $200 \le F1 \le 1000$; $550 \le F2 \le 2500$; $1100 \le F3 \le 3000$ e $2700 \le F4 \le 4500$ (figura 7.10). No estabelecimento destas regiões deve-se garantir que não há zonas de frequências pertencentes a mais do que duas regiões de formantes sob pena de se tornar bastante complexa a extracção dos formantes a partir do envelope espectral.

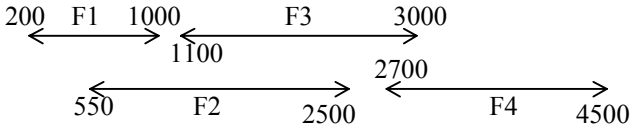


Figura 7.10 - Lugares das frequências dos 4 formantes.

Uma propriedade pertinente da fala na estimação dos formantes é a relação entre as frequências formantes e as amplitudes relativas dos picos formantes no envelope espectral. Para o modelo usado nesta análise, particularizado na figura 7.11, são impostas algumas constrições nas formas do envelope espectral. Em particular, é claro que fornecidas as frequências formantes e as larguras de banda, as amplitudes relativas dos picos do espectro ficam completamente especificadas. O conhecimento da relação entre os níveis dos formantes pode ser muito útil no processo de selecção de formantes a partir dos picos espectrais, pelo que é aconselhável um estudo sobre esta relação de amplitudes.

Neste trabalho os valores para a relação de amplitudes dos picos das formantes foram estabelecidos atendendo a [Schafer 70], em que se atribui uma maior importância à relação entre as amplitudes de F2 e F1 sendo a curva limite da relação apresentada na figura 7.12.

O nível da medida usado é Δ_{21} definido como

$$\Delta_{21} = 20\log_{10}\left|H\left(e^{j2\pi F_2 T}\right)\right| - 20\log_{10}\left|H\left(e^{j2\pi F_1 T}\right)\right| \tag{7.9}$$

em que F1 e F2 são as frequências do primeiro e segundo formantes e $|H()|$ é a amplitude do espectro alisado. A curva da figura 7.12 é determinada pela função criada no Matlab *ftresh2()*, cujo código se apresenta no anexo B15.

As curvas limites das relações entre F3 e F2 bem como entre F4 e F3 têm um valor contínuo nas frequências valendo respectivamente -34.8 e -48.8. Os níveis de medida usados são Δ_{32} e Δ_{43} respectivamente, definidos pela generalização da expressão 7.9.

$$\Delta_{i(i-1)} = 20\log_{10}\left|H\left(e^{j2\pi F_i T}\right)\right| - 20\log_{10}\left|H\left(e^{j2\pi F_{i-1} T}\right)\right| \tag{7.10}$$

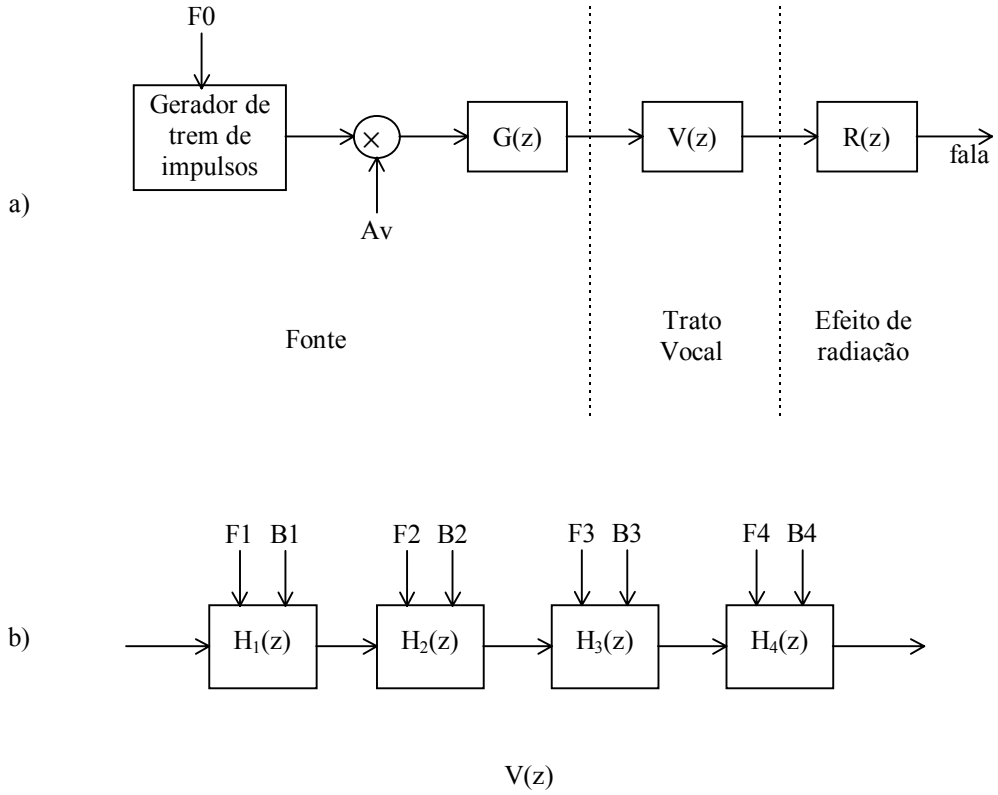


Figura 7.11 - a) Modelo digital para a fala vocalizada. b)Diagrama detalhado do modelo de trato vocal.

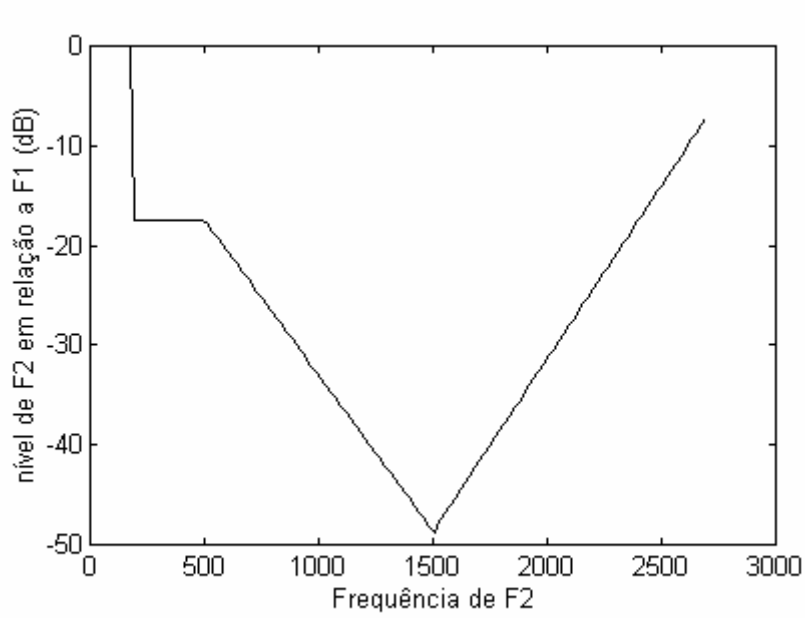
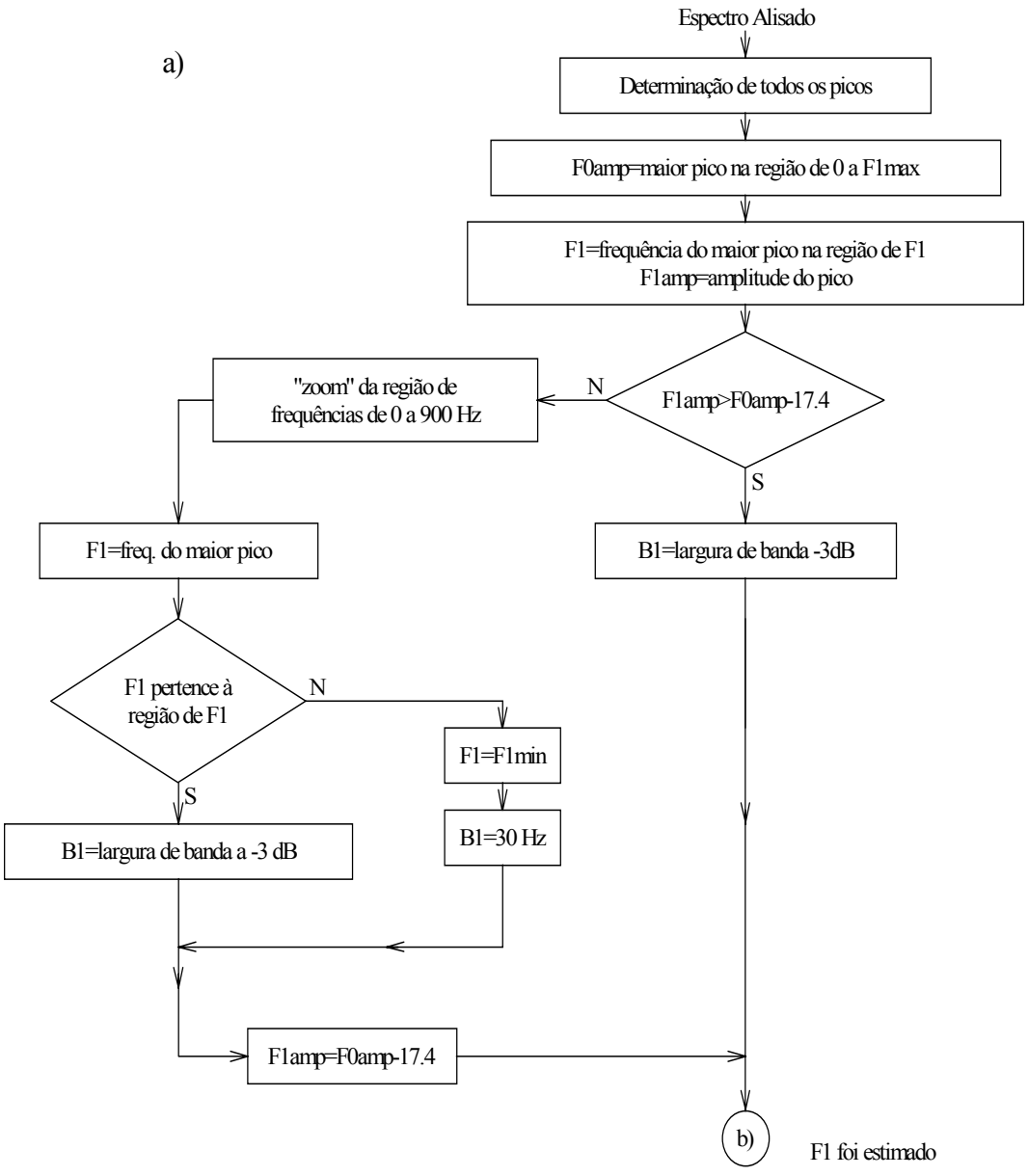


Figura 7.12 - Curva limite da relação das amplitudes logarítmicas entre F2 e F1.

Um pico na região de frequências de F2 só será considerado como F2 se Δ_{21} exceder a curva da figura 7.12.

Concretizando, no processo de extracção da frequência formante i, depois de encontradas as frequências formantes de ordem mais baixa, o máximo pico do espectro alisado dentro da região de frequências de Fi só será considerado como o formante Fi se $\Delta_{i(i-1)}$ estiver acima da curva limite para o formante i. Esta constrição impede que um pico do espectro alisado que tenha uma amplitude tão pequena, que, se a diferença entre a amplitude do próprio pico e a amplitude do pico do formante anterior, não exceder a curva limite do formante em análise, não seja considerado como o pico relativo a esse formante.

O fluxograma para a extracção dos formantes a partir do envelope espectral é apresentado na figura 7.13.



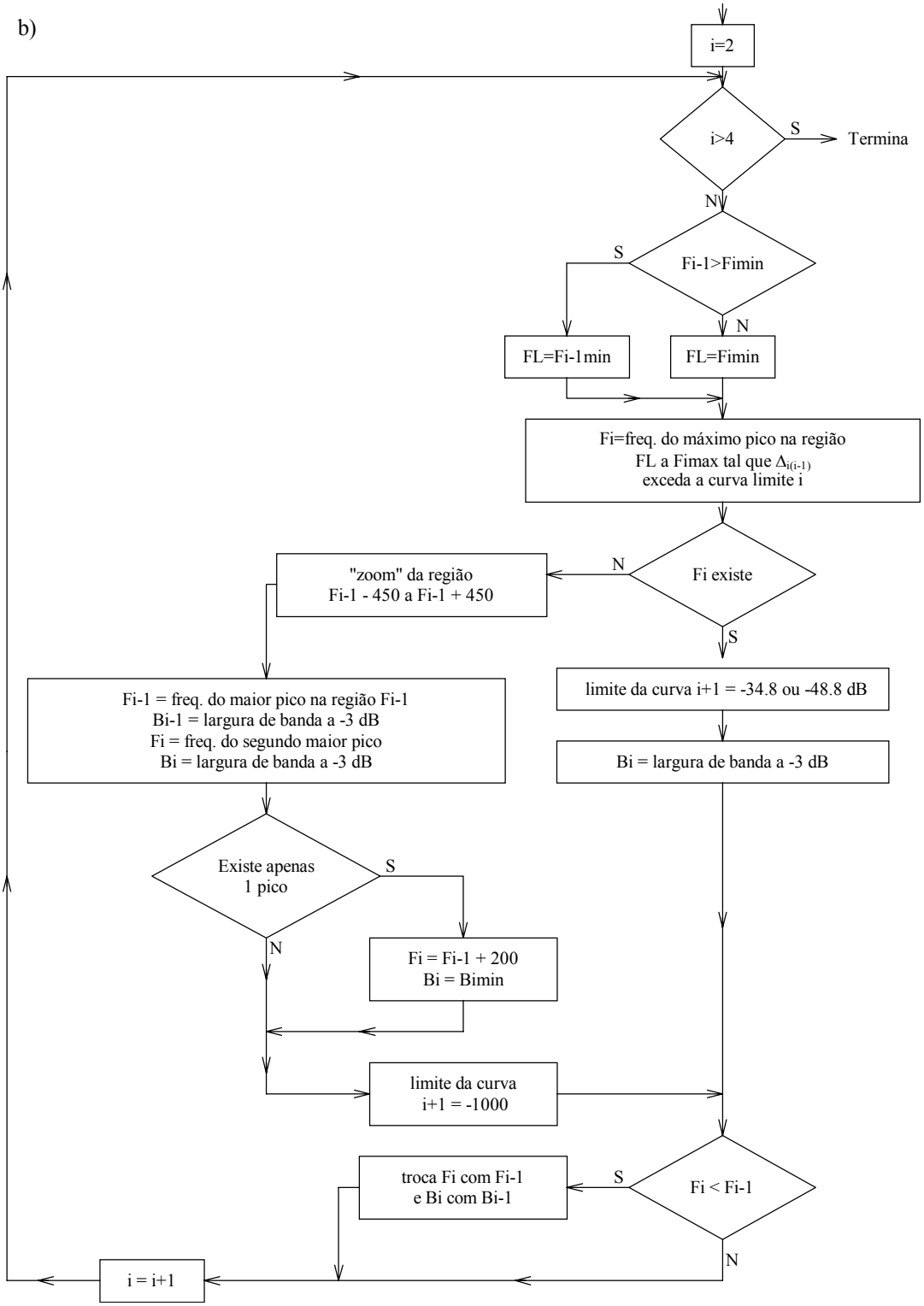


Figura 7.13 - Algoritmo usado para extrair os formantes e larguras de banda do envelope espectral. a) Extracção do primeiro formante e largura de banda; b) extracção dos formantes e larguras de banda i, com i=2,3,4.

O algoritmo está dividido em duas partes fig. 7.13a e 7.13b já que a parte a) relativa à extracção do primeiro formante e largura de banda difere ligeiramente da parte b) onde são extraídos em cada ciclo os formantes e larguras de banda 2, 3 e 4.

Partindo do espectro alisado, são determinados os picos e respectivas amplitudes com recurso à função implementada em Matlab com o nome *fpicos()* apresentada no anexo B16.

A iniciar o processo, é procurado o pico mais alto na gama de 0 a F1max onde F1max é o limite superior da região de frequências para o formante F1. A amplitude deste pico é atribuída a F0amp. Geralmente, este valor ocorre para um pico na região de frequências de F1, que será mais tarde escolhido como o formante F1. Contudo, existe por vezes, um forte pico abaixo de F1min, o limite mínimo da região F1, que é devido à forma de onda da fonte glotal. Nestes casos, pode haver, ou não, um pico evidente acima de F1min que seria F1. No sentido de evitar a escolha de um pico ilegítimo de baixa amplitude ou possivelmente o pico de F2 como F1, quando na verdade o pico relativo a F1 e o pico relativo à fonte não se distinguem, é requerido que o pico na região F1 esteja menos de 17.4 dB abaixo de F0amp para poder ser considerado como um possível pico de F1. A frequência do pico de mais alto nível na região de F1 que excede este nível mínimo é considerado como o formante F1. A amplitude deste pico é guardada como F1amp. Se nenhum pico na região F1 obedece a esta constrição, então a região de 0 900 Hz é expandida usando o algoritmo de CZT aplicado ao cepstro filtrado com uma janela de alisamento espectral com τ_1 maior (3 miliseg. em vez de 2 miliseg.), como foi discutido no capítulo anterior. Nesta secção do espectro ampliado com maior resolução, é pesquisado o maior pico na região de F1. A localização deste pico é aceite como F1. Se o "zoom" realizado não conseguiu separar os picos relativos a F1 e à fonte de excitação, então F1 toma o valor de F1min.

A quantidade F1amp é usada na estimação de F2. Se o pico F1 é muito baixo em frequência e não se distingue claramente do pico relativo à excitação da fonte, então F1amp é igualado a (F0amp-17.4). Esta atribuição é realizada para efectivamente baixar a curva limite usada na pesquisa de F2.

A largura de banda é obtida com recurso à função criada no Matlab *flbanda()* apresentada no anexo B17. Conhecida a frequência do formante, esta função procura, para a esquerda ou para a direita, a distância da próxima frequência cuja amplitude do espectro se situa a 3 dB abaixo da amplitude do pico do formante. A largura de banda toma o valor de 2 vezes essa distância em Hz[i].

A extracção dos formantes 2, 3 e 4 é iniciada pela comparação da estimação da frequência do formante anterior, F_{i-1}, com Fimin, em que Fimin é o limite inferior da região de frequências do formante i. Se F_{i-1} é menor que Fimin então a região a analisar será apenas de Fimin a Fimax. Contudo, se F_{i-1} foi estimado com uma frequência superior a Fimin, é possível que o pico respeitante a F_i tenha sido estimado como F_{i-1}. Então, a combinação das regiões F_{i-1} e F_i, de F(i-1)min a Fimax é pesquisada para assegurar que se é este o caso, então o pico correspondente a F_{i-1} será estimado como sendo o possível F_i. Depois do formante i ter sido estimado, F_i é comparado com F_{i-1} e os seus valores são trocados caso F_i seja menor que F_{i-1}.

O processo de decisão de qual dos picos pertencentes à região de frequências em análise corresponde ao formante i é idêntico ao caso de análise do primeiro formante salvo as diferenças que se apontam:

[i] A determinação da largura de banda desta forma, incorre no erro de considerar f0=(f1+f2)/2, em vez de correctamente se usar $f0=\sqrt{f1f2}$, em que f0 é a frequência central do pico, f1 a frequência inferior (a -3 dB) e f2 a frequência superior (a -3 dB). No entanto, o erro cometido nesta aproximação é inferior à resolução frequencial usada, não resultando portanto, em perda de rigor dos resultados obtidos.

A região de expansão será agora de F_{i-1} - 450 a F_{i-1} + 450. Na região expandida são procurados os dois maiores picos para F_{i-1} e F_i Caso exista apenas um pico, então F_i tomará o valor de F_{i-1} + 200 e B_i será de 50, 60, ou 90 para os casos de F2, F3 e F4 respectivamente. A curva limite para F_{i+1}, no caso da expansão, toma um valor de tal ordem elevado (-1000) que não faz sentido essa curva. Na extracção de F4, os valores da curva limite i+1 não são atribuídos obviamente por não serem necessários. No caso do processamento do segundo formante a curva limite usada é a referida na figura 7.12, variando portanto com a frequência do pico respeitante a F2. As larguras de banda são estimadas pelo mesmo processo descrito anteriormente.

No capítulo seguinte será realizada uma avaliação mais profunda dos resultados deste algoritmo para a fala contínua. No entanto, deixam-se aqui registados os valores das formantes e larguras de banda obtidos para um segmento de fala da vogal [a], cujo espectro alisado é representado na figura 7.14, como um simples exemplo dos resultados deste processamento. Assim as frequências formantes obtidas foram de F1=781 Hz, F2=1169 Hz, F3=2208 Hz, F4=2765 Hz e as respectivas larguras de banda B1=194 Hz, B2=194 Hz, B3=194 Hz, B4=280 Hz.

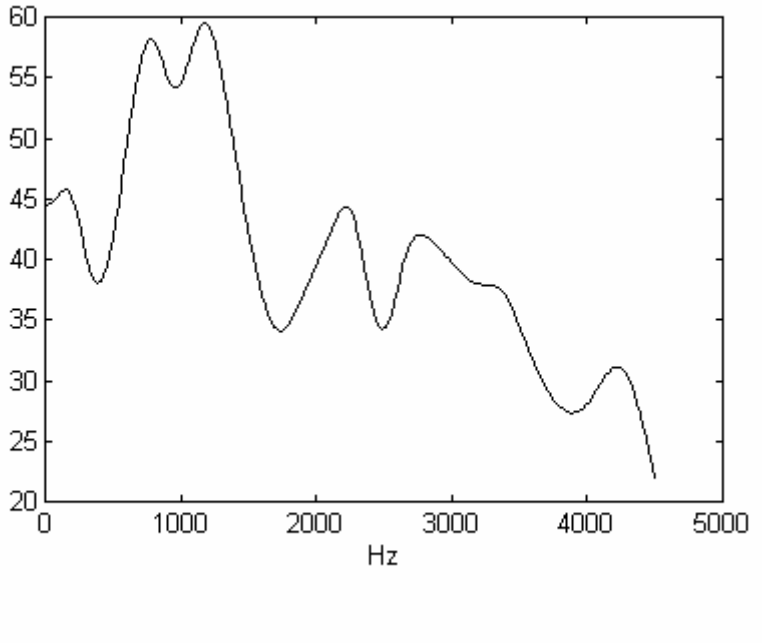


Figura 7.14 - Espectro alisado pelo método do cepstro de um segmento da vogal [a].

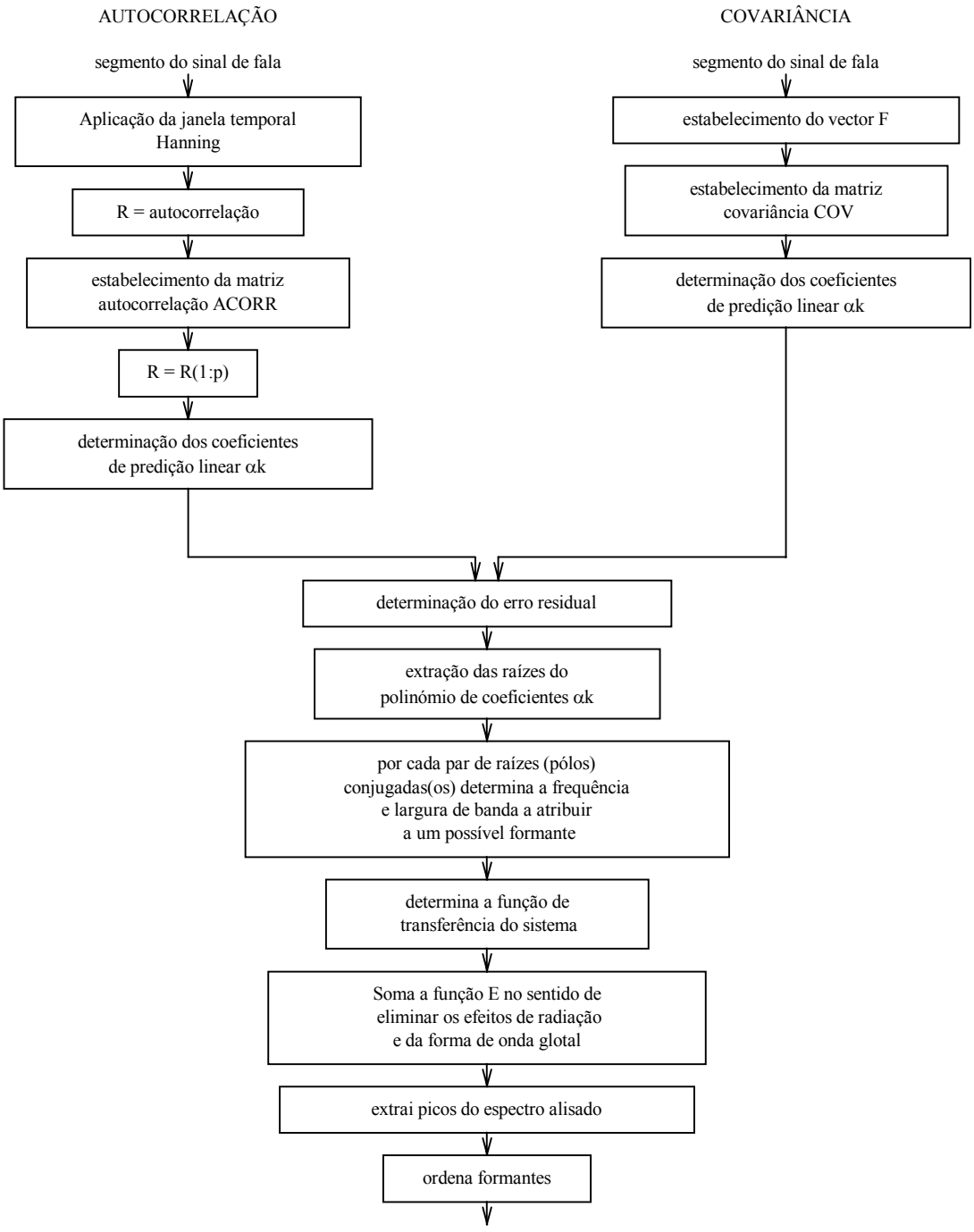
7.3.1.2 Método da Predição Linear - Matriz Autocorrelação e Matriz Covariância

Os sistemas implementados, testados e aqui apresentados de extracção automática das 4 formantes e respectivas larguras de banda em segmentos de sinais de fala vocalizada pelo método da predição linear para um modelo só com pólos, cujo estudo teórico já foi desenvolvido no capítulo 4, recorrem à matriz autocorrelação e matriz covariância, em que os processos para determinação dessas matrizes foram já discutidos no capítulo 6.

O processamento descrito nesta secção para extracção automática dos parâmetros é implementado, para os métodos da matriz autocorrelação e da matriz covariância, respectivamente pelas funções *fcorprd2()* e *fcovpr2()* desenvolvidas no programa

Matlab e apresentadas nos anexos B7 e B8 respectivamente. Cada uma destas funções serve o corpo principal do programa de extracção de parâmetros para a fala vocalizada, sendo chamadas para extrair esses parâmetros segmento a segmento.

O fluxograma dos dois métodos aqui apresentados é o mesmo depois que se obtêm os coeficientes de predição linear α_k diferentemente pelos métodos da matriz autocorrelação e da matriz covariância. A figura 7.15 representa o fluxograma destes métodos aqui estudados para extracção automática dos 4 formantes e respectivas larguras de banda.



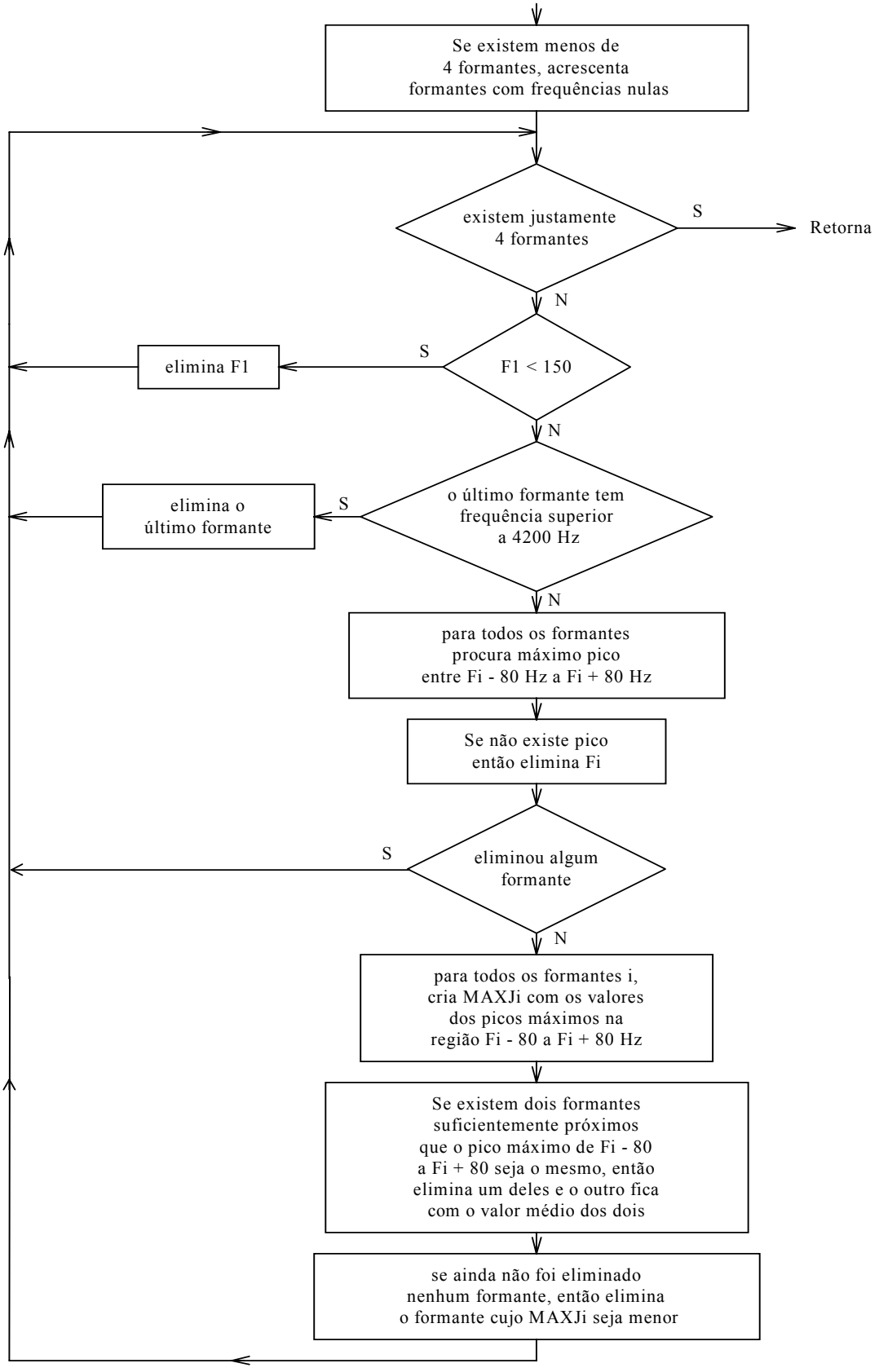


Figura 7.15 - Fluxograma do processamento de extracção automática dos formantes e larguras de banda pelo método de predição linear - matriz autocorrelação e matriz covariância.

O processamento para o método da matriz autocorrelação exige que o número de amostras temporais do segmento a analisar seja suficientemente longo, da ordem de alguns períodos fundamentais, para assegurar resultados credíveis. Então, é aplicada uma janela de Hanning para realizar uma pesagem das amostras temporais, simulando de certa forma um aumento do comprimento do segmento. O sinal de saída é depois correlacionado consigo próprio, para construir a matriz autocorrelação ACORR da equação 6.21 e o vector R da equação 6.23 pelo processo descrito no capítulo anterior. Os coeficientes de predição linear estimados α_k, com k=1... p, sendo p o número de pólos são obtidos pela multiplicação matricial da inversão da matriz ACORR pelo vector R.

No caso do método da matriz covariância, a obtenção de resultados credíveis não condiciona o número de amostras do segmento de fala a analisar. Assim, é dispensável a utilização de uma função de janela temporal. O processo é iniciado pelo cálculo do vector F da equação 6.26 e da matriz covariância COV, pelo processo descrito no capítulo anterior. Estes elementos são usados para a estimação dos coeficientes de predição linear α_k, com k=1...p, sendo p o número de pólos, pela operação multiplicação matricial da inversão da matriz COV pelo vector F.

Deste ponto em diante, o processo seguido pelos dois métodos para a extracção dos formantes é igual.

Com base nos coeficientes de predição linear estimados é determinado o erro residual pela expressão 4.29. Os coeficientes do filtro inverso de predição linear da equação 4.30 são então 1-α_k. Seguidamente, são extraídas as raízes, do polinómio com estes coeficientes, que são os pólos da função de transferência do sistema. Cada par de pólos complexos conjugados é considerado como um possível formante. O número de possíveis formantes é variável, dependendo do número de pólos considerados para o modelo e das características do trato vocal do segmento em análise, obrigando a uma posterior selecção de justamente 4 formantes. A frequência de cada formante e respectiva largura de banda são determinados pela localização do par de pólos respectivo.

Para cada par de pólos

$$z = e^{sT} \tag{7.11}$$

em que

$$s = -\pi B \pm j2\pi F \tag{7.12}$$

e F a frequência natural do par de pólos e B a largura de banda.

Fica

$$\begin{aligned} z &= e^{-\pi BT} e^{\pm j2\pi FT} \\ &= e^{-\pi BT}\left[\cos(\pm 2\pi FT) + j\,\mathrm{sen}(\pm 2\pi FT)\right] \\ &= e^{-\pi BT}\left[\cos(2\pi FT) \pm j\,\mathrm{sen}(2\pi FT)\right] \end{aligned} \tag{7.13}$$

sendo

$$|z| = e^{-\pi BT} \tag{7.14}$$

$$\mathrm{Re}(z) = e^{-\pi BT}\cos(2\pi FT) \tag{7.15}$$

$$\mathrm{Im}(z) = e^{-\pi BT}\mathrm{sen}(2\pi FT) \tag{7.16}$$

Fazendo a largura de banda estimada

$$\hat{B} = (-F_s/\pi)\ln|z| \tag{7.17}$$

em que Fs é a frequência de amostragem = 1/T, e substituindo a expressão 7.14 fica

$$\begin{aligned}\hat{B} &= \frac{-1}{\pi T}\ln\left(e^{-\pi BT}\right)\\ &= \frac{-1}{\pi T}(-\pi BT)\\ &= B\end{aligned} \tag{7.18}$$

Confirmando assim correcta a expressão 7.17 para a estimação da largura de banda.

A frequência formante estimada será

$$\hat{F} = \frac{1}{2\pi T}\mathrm{arctg}\left(\frac{\mathrm{Im}(z)}{\mathrm{Re}(z)}\right) \tag{7.19}$$

fazendo novamente Fs=1/T e substituindo as expressões 7.15 e 7.16, fica

$$\begin{aligned}\hat{F} &= \frac{1}{2\pi T}\mathrm{arctg}\left(\frac{e^{-\pi BT}\mathrm{sen}(2\pi FT)}{e^{-\pi BT}\cos(2\pi FT)}\right)\\ &= \frac{1}{2\pi T}2\pi FT\\ &= F\end{aligned} \tag{7.20}$$

Confirmando a expressão 7.19 para a estimação das frequências formantes.

Após a estimação das frequências formantes e respectivas larguras de banda de cada par de pólos complexos conjugados, é determinada a função de transferência do sistema recorrendo à expressão 4.25. À curva da função de transferência, com uma forma idêntica à do envelope espectral do trato vocal, é somada a curva de equalização das amplitudes dos formantes, da figura 7.9, na tentativa de anular os efeitos de radiação e da forma de onda glotal, tal como para o método do cepstro. Da função de transferência resultante são extraídos os picos com a função *fpicos()* do

anexo B16. São ordenados os possíveis formantes no vector FORM por ordem crescente das suas frequências. Se o número de possíveis formantes obtidos for inferior a 4, então são acrescentados formantes e larguras de banda com valores nulos, até que o seu número seja 4, para satisfazer a imposição de dimensão dos vectores de saída das funções *fcorprd2()* e *fcovpr2()*. Se por outro lado o número de possíveis formantes for superior a 4 então, é iniciado um processo de eliminação de possíveis formantes e respectivas larguras de banda até que restem apenas 4 considerados como verdadeiros formantes. Este processo a seguir descrito elimina apenas 1 par, possível formante e respectiva largura de banda, sendo repetido até o número de formantes ser 4.

Se o primeiro possível formante tem uma frequência inferior a 150 Hz então é eliminado e o primeiro possível formante passa a ser o seguinte, senão, verifica se a frequência do último possível formante é superior a 4200 Hz. Em caso afirmativo, elimina-o e repete o processo. Em caso negativo, verifica para cada possível formante se não existe algum pico no espectro dentro de uma gama de frequências F_i-gama a F_i+gama (gama=80 Hz). Quando aparece o primeiro caso elimina o possível formante respectivo e inicia o processo. Se não encontrar nenhum possível formante sem um pico na respectiva gama de frequências então, para cada possível formante i determina o valor da frequência e de amplitude do máximo pico dentro da sua gama de frequências. Se existirem dois formantes suficientemente próximos que o pico máximo da região de frequências seja o mesmo então elimina um deles e o outro fica com uma frequência que será a média da frequência dos dois (o mesmo sucede para as larguras de banda). Finalmente, se ainda não foi eliminado nenhum formante então, elimina o que tiver o pico máximo na respectiva gama de frequências com menor amplitude, terminando o processo.

Este processo é então repetido tantas vezes quantos possíveis formantes além de 4 tiverem sido inicialmente estimados. Quando um formante é eliminado também o é a sua largura de banda.

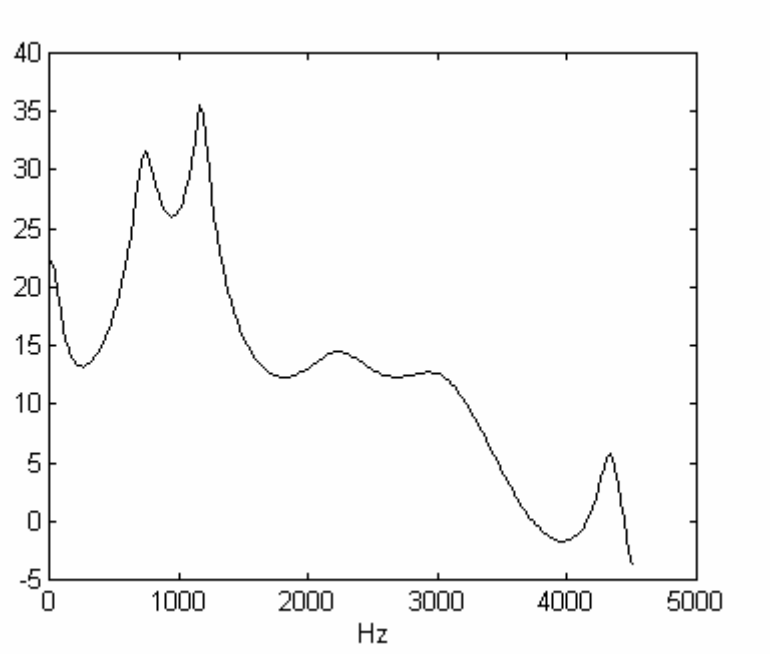


Figura 7.16 - Função de transferência do trato vocal obtida pelo método da matriz autocorrelação para um segmento da vogal [a].

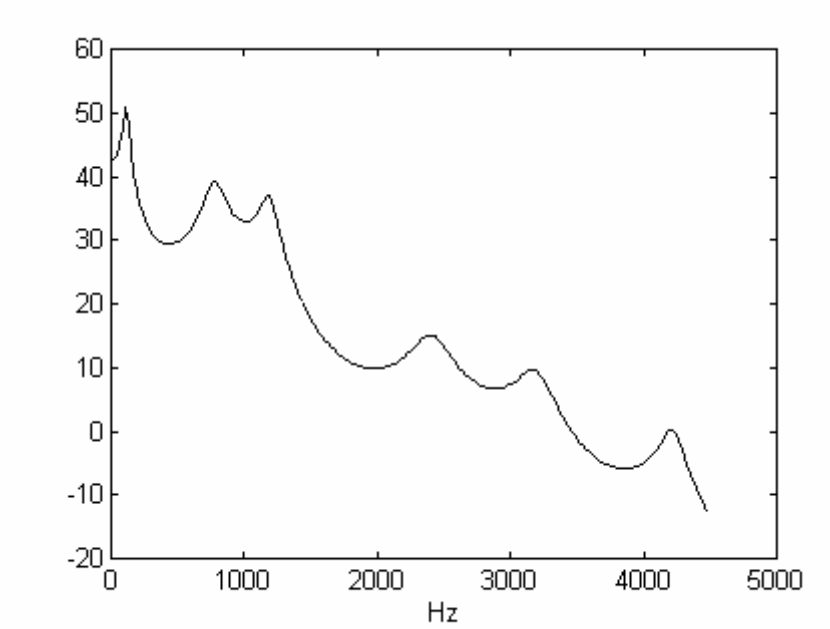


Figura 7.17 - Função de transferência do trato vocal obtida pelo método da matriz covariância para um segmento da vogal [a].

Deixa-se para o capítulo seguinte uma avaliação mais profunda dos resultados deste algoritmo para os dois métodos em fala contínua. Contudo, apresenta-se aqui nas figuras 7.16 e 7.17 a função de transferência obtida para um segmento de fala da vogal [a]. Os parâmetros estimados. apresentam-se na tabela 7.1.

Tabela 7.1 - Parâmetros estimados para um segmento da vogal [a] pelo método de predição linear, matrizes autocorrelação e covariância.

(Hz)	F1	B1	F2	B2	F3	B3	F4	B4
autoc.	707	125	1148	83	2192	564	2978	692
covar.	752	154	1171	121	2387	287	3172	258

Não se pode deixar de reparar na convergência de valores estimados para os parâmetros pelos três métodos já descritos.

7.3.2 Análise Síncrona Com o Período Fundamental

O método de análise síncrona com o período fundamental já discutido no capítulo 4, baseia-se na análise de segmentos de fala vocalizada com a duração de 1 período e com início no impulso glotal ou um pouco mais à frente com o intuito de que no segmento analisado estejam presentes apenas as características do trato vocal.

Como se compreende, este método segue uma segmentação diferente da usada nos métodos anteriores, com uma duração variável e síncrona com o período fundamental. Razão pela qual se separa este método dos anteriores pois está sujeito a um processamento distinto do imposto pelo corpo principal do programa principal de análise discutido anteriormente.

O número de amostras do segmento a analisar depende da duração do período fundamental, sabendo-se à partida que será sempre pequeno. Isto impõe algumas limitações quanto ao método usado para analisar cada segmento de um período fundamental de fala. Assim, foi implementado o método de análise por predição linear pela matriz covariância, por ter um melhor comportamento, que os outros métodos, quando usado com um sinal de tamanho mais reduzido. A utilização da análise síncrona com o período fundamental obriga a um qualquer procedimento que reconheça exactamente o início de cada impulso glotal para precisamente realizar o sincronismo.

O processo mais preciso para um total conhecimento do impulso glotal recorre ao uso de um electroglotógrafo, já referido no capítulo 2, com o qual se pode ter uma "imagem" do impulso glotal pela medição da variação da impedância eléctrica entre os extremos da glote.

Contudo, há outros métodos mais ou menos sofisticados para determinação do sincronismo com o período fundamental recorrendo apenas ao sinal de fala [McAulay 86].

Neste trabalho também foi desenvolvido um algoritmo para detecção do sincronismo do período fundamental da fala vocalizada.

O algoritmo é implementado na função *fdetsinc()* em Matlab e apresentada no anexo B18.

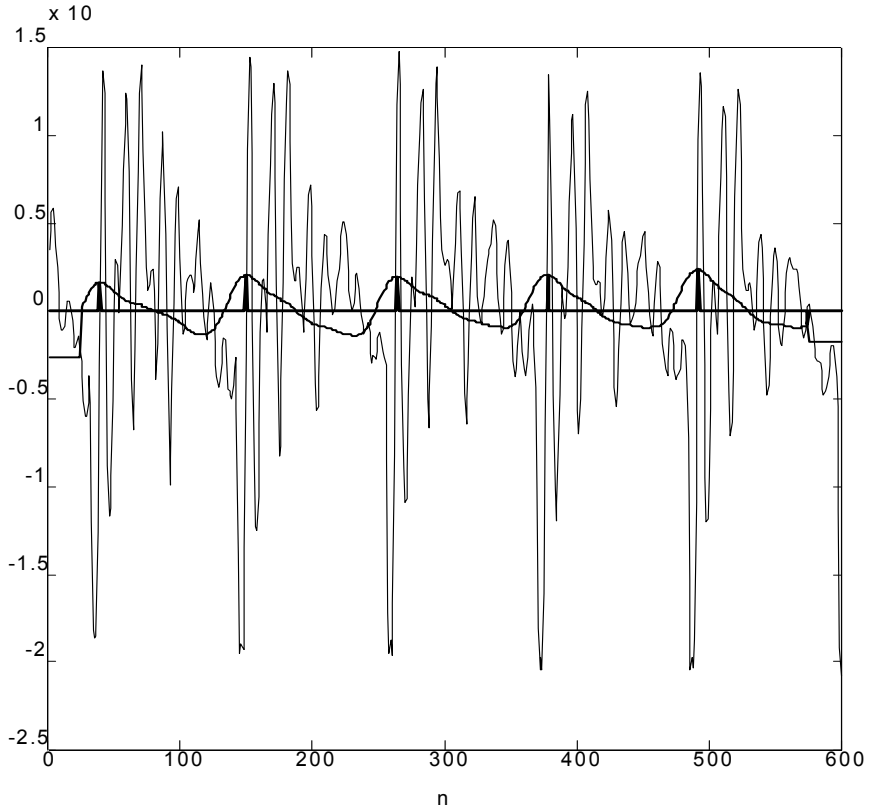


Figura 7.18 - Detecção de sincronismo realizada com a função *fdetsinc()* aplicada ao sinal [a]. A traço fino o sinal de fala, a traço médio a média deslizante e a traço grosso os picos marcando o suposto início do impulso glotal.

O algoritmo muito simplesmente determina a média deslizante do sinal com um espaçamento unitário, para uma maior precisão, e com um comprimento de janela de 50 amostras. O resultado é um sinal alisado do qual se determinam os picos. Estes picos, quando o comprimento da janela média é correctamente escolhido, coincidem com o início do impulso glotal.

A figura 7.18 mostra um exemplo da detecção de sincronismo marcada com picos no suposto início do impulso glotal realizada pela função *fdetsinc()*.

Este algoritmo tem resultados bons quando aplicado a sinais de fala nitidamente vocalizados. Uma escolha acertada do comprimento da janela que realiza a média deslizante é muito importante para a obtenção de bons resultados. Infelizmente, o que é um bom comprimento desta janela para um sinal nem sempre o é para outro, podendo este aspecto ser visto como o "calcanhar de Aquiles" deste procedimento básico.

O algoritmo desenvolvido para a análise síncrona com o período fundamental é apresentado na figura 7.19. Este algoritmo foi implementado no programa *falasinc.m* em Matlab (anexo B19).

Este programa, tal como o *falacont*, permite ao utilizador escolher o sinal de fala a analisar de entre um dos sinais previamente armazenados. Se a frequência desse sinal for superior a 20 Khz (caso dos sinais recolhidos a 22.05 Khz) realiza uma decimação 2:1 com filtragem anti-aliasing, cuja função do filtro foi analisada no capítulo 3. O sinal é então apresentado ao utilizador para que este escolha o início e fim da análise. Seguidamente, com recurso à função *fdetsinc()* é determinado o vector M contendo os picos correspondentes ao início dos impulsos glotais. Carrega apontador, ini, com o índice do vector do sinal de fala correspondente ao início do primeiro impulso glotal. Este, corresponde ao índice do primeiro elemento não nulo do vector M. Depois, e repetidamente até ao último período fundamental da parte do sinal escolhido é realizada ciclicamente a sequência de operações iniciada pela atribuição ao apontador fim, do índice do vector do sinal de fala correspondente ao início do próximo impulso glotal e portanto fim do impulso glotal actual. Este corresponde ao próximo, depois de ini, elemento não nulo do vector m. Assim está definido o segmento a analisar que será desde o início, ini, ao índice fim. Alguns autores [Miyoshi 87] segmentam o sinal, não desde o início do impulso glotal, mas sim mais à frente, supostamente depois do sinal do impulso glotal (ver figura 6.16) se anular, com a intenção de não haver qualquer efeito deste na análise do trato vocal. Contudo, a solução para determinação do sincronismo com o período fundamental não permite o conhecimento do fim do impulso glotal. Ainda assim, foi testada a análise com início do segmento progressivamente mais adiante do início do impulso glotal mas com resultados cada vez menos bons. Esta deterioração da qualidade dos parâmetros obtidos à medida que o início do segmento avança em relação ao início do impulso glotal pode dever-se ao facto de simultaneamente haver uma diminuição do comprimento do segmento com repercussões na estabilidade e precisão do modelo usado na análise.

Cada segmento do sinal é então submetido à função *fcovpr2()* para extracção dos parâmetros (frequência fundamental, 4 formantes, 4 larguras de banda) e respectiva função de transferência pelo método de predição linear, matriz covariância. É também determinada a energia do sinal recorrendo à expressão 7.7.

Seguidamente, são armazenados os parâmetros do segmento nos vectores de parâmetros na posição correspondente ao segmento. O apontador, ini, para o segmento seguinte é igualado ao apontador fim do segmento actual. Se o segmento actual ainda não corresponde ao último impulso glotal então é repetido este ciclo.

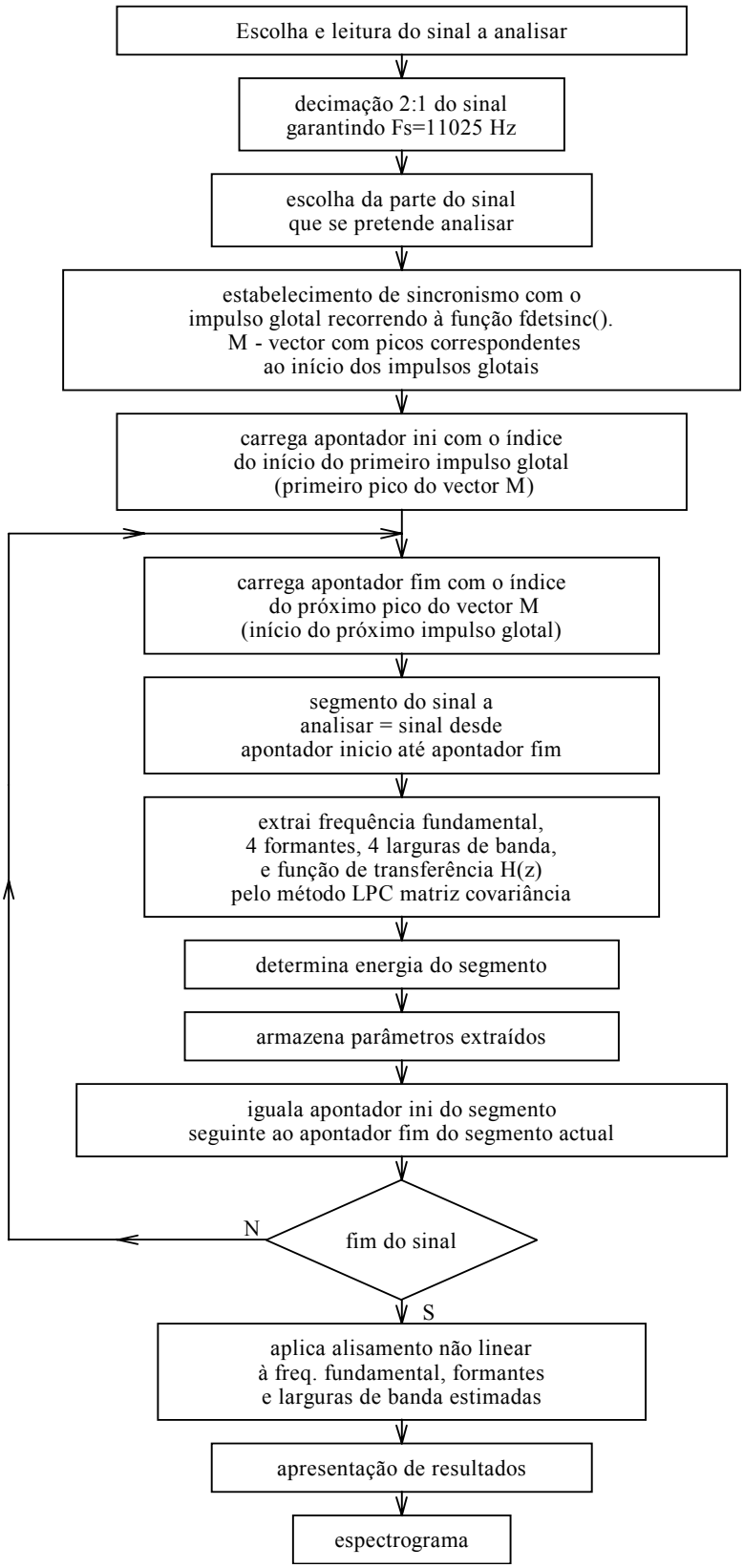


Figura 7.19 - Fluxograma do algoritmo desenvolvido para extracção automática dos parâmetros do modelo de fala vocalizada por um processo de análise síncrona com o período fundamental.

Finalmente, para a sequência de cada parâmetro extraído automaticamente ao longo do sinal é aplicado um alisamento não linear para corrigir eventuais pontos fora da

trajectória estimada para esse parâmetro. Este alisamento é discutido na secção seguinte.

Uma vez concluída a análise do sinal de fala são apresentados graficamente ao longo do tempo os valores para a frequência fundamental, formantes e larguras de banda, bem como o espectrograma determinado pela sequência de funções de transferência e apresentado pela função *espectro()* já anteriormente apresentada.

Remete-se para o próximo capítulo uma análise dos testes e resultados deste algoritmo. No entanto, apresenta-se na figura 7.20 a função de transferência obtida com este processamento para um segmento do sinal da vogal [a]. Para este segmento os parâmetros obtidos foram F1=746Hz, F2=1157Hz, F3=2464Hz, F4=3692Hz, B1=287Hz, B2=182Hz, B3=-72Hz, B4=61Hz. O valor negativo de B3, indica que o 3º formante tem os pólos fora do circulo unitário. Isto evidencia a estabilidade crítica a que está sujeito este sistema quando o número de amostras do segmento é reduzido.

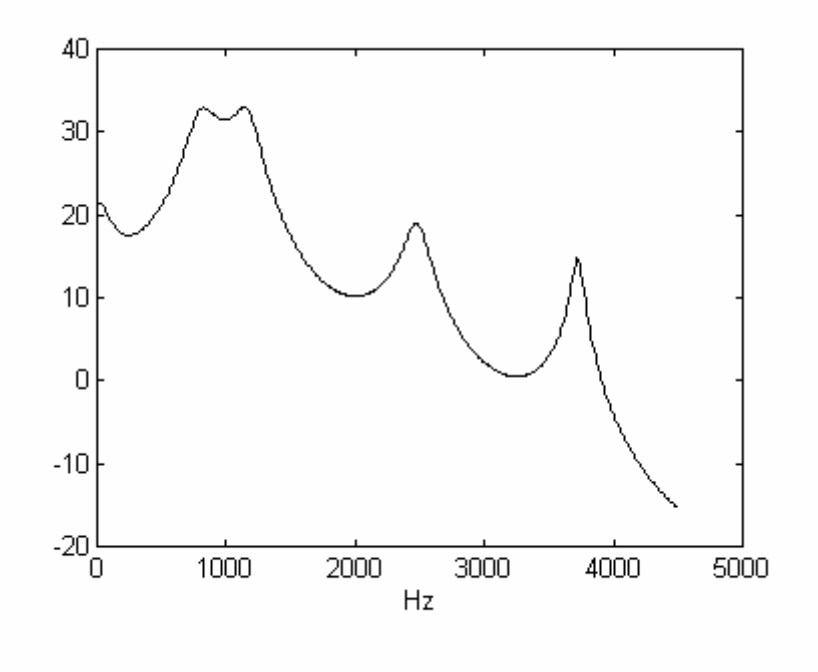


Figura 7.20 - Função de transferência do trato vocal para um segmento de fala da vogal [a] obtida pelo algoritmo descrito para análise síncrona.

7.3.3 Alisamento Não Linear Aplicado à Sequência de Parâmetros Estimados

Nos processos analisados para extracção automática dos parâmetros não foi tomada nenhuma atenção especial para tirar vantagem do facto de que esses parâmetros devem variar de uma forma contínua com o tempo. Os processos implementados conduzem a situações de dificuldade para recuperar de estimativas erradas de um dos parâmetros torna-se então inevitável que erros "grosseiros" possam ocorrer na estimativa destes parâmetros. Um meio de corrigir de certa forma esses erros "grosseiros", isto é, pontos que estejam claramente fora de uma linha, é a imposição de uma condição de continuidade.

A imposição de continuidade é incorporada nos sistemas de análise pelo recurso a uma simples operação de alisamento não linear aplicada à sequência de valores da

frequência fundamental, formantes e larguras de banda. As situações consideradas erradas e corrigidas pelo algoritmo são apresentadas na figura 7.21.

São consideradas 4 situações de existência de erro. Situação A para 1 ou 2 pontos abaixo da linha e situação B para 1 ou 2 pontos acima da linha. Na figura 7.21, para cada situação apresenta-se a sequência de valores dos parâmetros e em baixo a derivada dessa sequência em relação ao intervalo -ΔT a +ΔT considerado como o limite máximo aceitável para a variação dos parâmetros. Assim a situação A 1 ponto é detectada quando para um valor da sequência de parâmetros está mais de ΔT abaixo dos pontos anterior e posterior. A situação B 1 ponto para o mesmo caso, mas quando o valor do parâmetro se encontra ΔT acima do ponto anterior e posterior. Nestes dois casos o valor do parâmetro é corrigido como sendo a média dos dois pontos vizinhos mais próximos. A situação A 2 pontos é detectada quando um valor da sequência de parâmetros está ΔT abaixo do ponto anterior e o ponto seguinte está também ΔT abaixo do seu ponto seguinte. A situação B 2 pontos é idêntica à de A 2 pontos salvo que os desvios são considerados para cima e não para baixo. Nestes dois casos, os valores para os dois pontos fora da linha são determinados por uma interpolação entre os pontos correctos anterior e posterior.

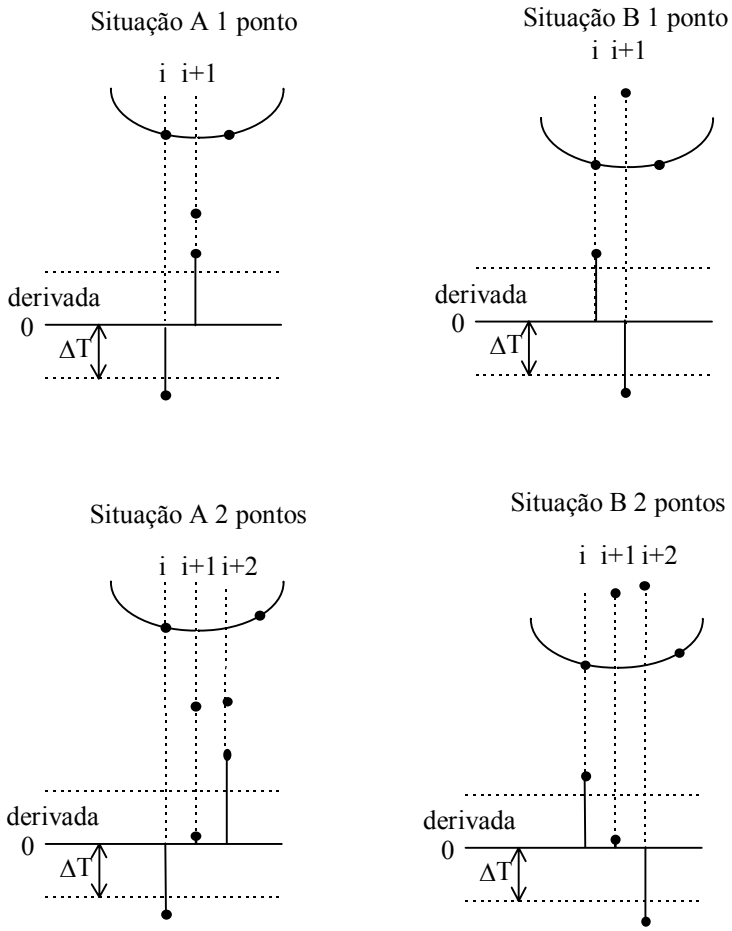


Figura 7.21 - Situações de parâmetros fora de uma linha consideradas como incorrectas.

Este algoritmo de alisamento não linear é implementado na função *fcorrec()* no Matlab e apresentada no anexo B20.

Os valores de ΔT usados pelos programas que recorrem a este alisamento (*falacont.m* e *falasinc.m*) são:

Tabela 7.2 - ΔT usado para realizar o alisamento não linear.

Parâmetro	**ΔT**
período fundamental	1 ms
F1	100 Hz
F2	150 Hz
F3	200 Hz
F4	250 Hz
B1	100 Hz
B2	150 Hz
B3	200 Hz
B4	250 Hz

7.4 Análise de Fala Não Vocalizada

O estudo da análise de fala não vocalizada foi considerado neste trabalho segundo dois modelos diferentes para estes sinais. Para o modelo com 4 formantes a análise pode ser realizada pelos dois métodos de predição linear, desenvolvidos para a analise de fala vocalizada exactamente da mesma forma. O sinal de erro residual aparecerá com uma forma ligeiramente diferente já que não devem ocorrer os picos correspondentes ao impulso glotal. Contudo, este modelo apesar de ser usado por alguns autores (MULTIVOX), não se adapta com rigor aos sinais de fala excitados por ruído (não vocalizados). Isto reflecte-se no erro residual que tem para estes sinais uma energia maior correspondendo portanto a um erro maior devido ao modelo não ser correcto para este sinal. O modelo correcto obriga ao uso de um zero o que complica sobremaneira o método da predição linear.

Assim, foi implementado um algoritmo para análise dos sinais de fala não vocalizada baseado no modelo apresentado em 4.3.1, com um pólo e um zero, no programa *falanvoc.m* em Matlab cujo código vem no anexo B21.

Este algoritmo lê o sinal pretendido, dá ao utilizador a possibilidade de seleccionar o início da parte do sinal de fala não vocalizada que se pretende analisar e selecciona um segmento de um comprimento pré-estabelecido a partir do início escolhido. Seguidamente determina a função de alisamento espectral pelo método do cepstro. Depois, encontra os picos da função de alisamento espectral e atribui a F_p , frequência do pólo, a frequência do pico com maior amplitude entre 1000 e 4000 Hz. A frequência do zero, F_z é determinada recorrendo às expressões 4.15 e 4.16.

Na figura 7.22 apresenta-se o espectro alisado de um segmento de duração 36 ms do sinal de fala correspondente ao som [ʃ] da palavra "isto". Os parâmetros extraídos pelo algoritmo foram F_p=2240Hz, F_z=550Hz.

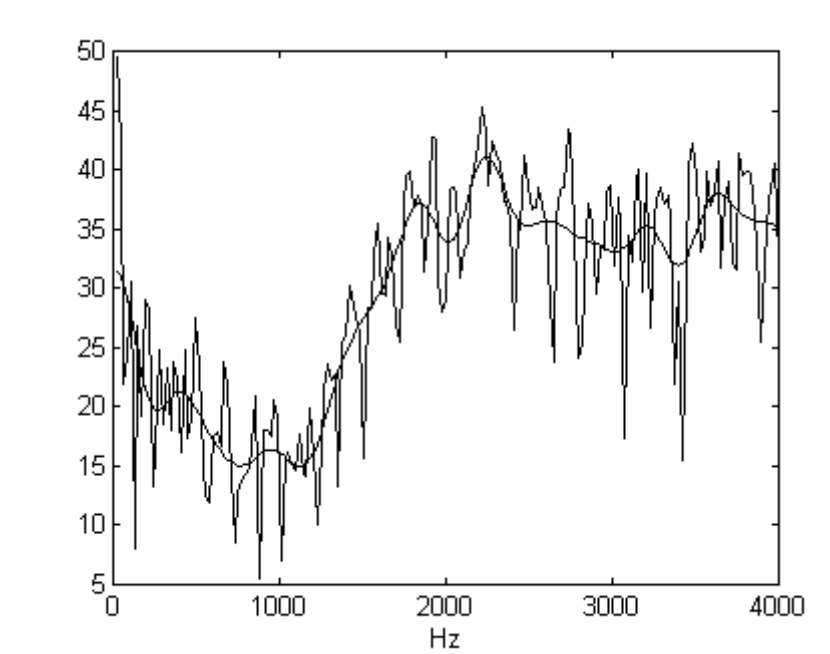


Figura 7.22 - Espectro e espectro alisado pelo método do cepstro de um segmento de fala do som do fonema [ʃ] retirado da palavra "isto".

CAPÍTULO 8

TESTE E AVALIAÇÃO DOS RESULTADOS

8. TESTE E AVALIAÇÃO DOS RESULTADOS

8.1 Introdução

Neste capítulo procura-se fazer uma apresentação de alguns dos resultados obtidos neste trabalho que informem ao leitor sobre a qualidade dos sistemas de análise e de síntese desenvolvidos.

Os resultados intermédios das ferramentas que foram usadas nos sistemas de análise foram já apresentados no momento da discussão da própria ferramenta. Evita-se repetir aqui esses resultados para não tornar demasiado pesada a leitura. Assim, serão apresentados apenas os resultados finais dos sistemas usados para extracção automática de parâmetros da forma que melhor permita a sua avaliação e comparação.

Não é possível a apresentação de resultados objectivos finais do sistema de conversão texto-fala para o português por não se terem ainda realizado testes de inteligibilidade com um grande número de pessoas.

8.2 Conversor Texto-Fala para o Português

Os resultados já obtidos relativamente a este sistema são baseados na opinião do autor e de outras pessoas próximas conhecedoras do estado de desenvolvimento do sistema.

Avaliando separadamente os blocos constituintes do conversor dependentes da língua pode-se afirmar que:

1. A conversão em conjunto de caracteres internos do MULTIVOX que inclui as rotinas de conversão de números em códigos fonéticos, e conversão de abreviaturas, acrónimos e distinção de palavras homógrafas atingiram o nível de funcionamento pretendido não se julgando necessários mais desenvolvimentos.

2. O grupo de fonemas base implementados no sistema está também completo.

3. As regras de conversão grafema-código de fonema estão em fase adiantada de desenvolvimento, cobrindo já uma percentagem muito elevada de conversões com sucesso, não sendo graves alguns erros que ainda existem. Contudo, reconhece-se a necessidade de mais alguma dedicação a este bloco para se atingirem resultados perfeitos.

4. A preparação prosódica ao nível de códigos fonéticos está bem ao nível da acentuação das palavras e entoação das frases reconhecendo-se a necessidade de melhoria especialmente ao nível da programação rítmica para que o sistema produza fala natural.

5. Os blocos respeitantes às regras de concatenação das estruturas acústicas dos fonemas e o inventário de 255 unidades acústicas mínimas estão intimamente relacionados e vão sendo gradualmente melhorados em conjunto, sendo difícil chegar a um ponto e concluir-se como completo o desenvolvimento destes dois blocos. No entanto, estão numa fase de desenvolvimento em são perceptíveis os sons produzidos, necessitando todavia de mais desenvolvimento para uma maior inteligibilidade, da qual não se pode separar o desenvolvimento prosódico.

6. Os grupos de regras de entoação e, sobreposição de elementos de prosódia e correcção de intensidade, necessitam de mais desenvolvimento já que foram implementadas pouco mais do que algumas regras básicas já descritas. Um bom desempenho destes blocos podem dar uma mais valia a este sistema de conversão texto fala.

O sistema neste momento produz fala extremamente clara para quem está habituado a ouvi-lo. Para quem não está habituado ao sistema a fala produzida soa ao princípio um pouco robótica, sendo no entanto, inteligível.

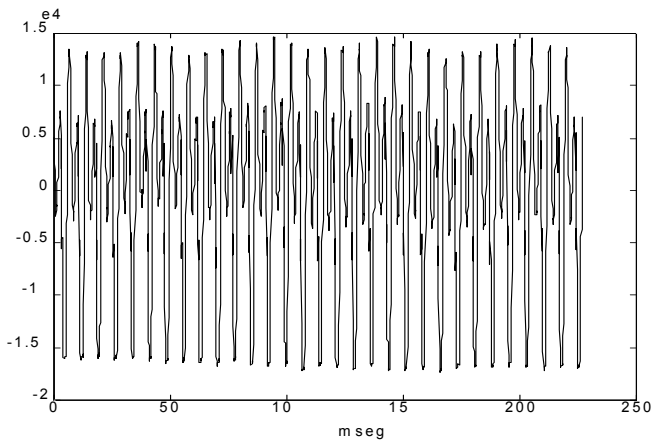
8.3 Determinação da Frequência Fundamental com Processamento no Domínio Temporal

O processamento realizado no domínio temporal para determinação da frequência fundamental apresentado no capítulo anterior é direccionado para uma fácil implementação em "hardware", permitindo assim, a determinação deste parâmetro em tempo real com uma muito boa precisão para sinais recolhidos nos mais diversos ambientes e independentemente do falante.

8.4 Extracção Automática de Parâmetros dos Sinais de Fala

Uma correcta avaliação dos diferentes sistemas usados (análise cepstral, predição linear pelos métodos das matriz autocorrelação e covariância e análise síncrona com o período fundamental) para extrair automaticamente os parâmetros (frequência fundamental, 4 formantes, 4 larguras de banda e amplitude) e os envelopes espectrais dos sinais de fala vocalizada só pode ser realizada perante os resultados para diferentes sinais. Apesar de terem sido feitos testes para os diferentes métodos com os mais variados sinais, apresentam-se aqui apenas os resultados obtidos pelos métodos usados para a vogal [i] locuccionada continuamente e para a palavra "ama" pronunciados pelo locutor 2 e para um sinal sintetizado com uma variação conhecida dos seus parâmetros, apresentados na figura 8.1. Os resultados a seguir apresentados para os diferentes métodos dizem respeito à variação das 4 frequências formantes, respectivas larguras de banda e espectrograma. Para o método de análise cepstral também é apresentada a variação da frequência fundamental. A variação da amplitude dos segmentos ao longo do sinal não é apresentada por não haver qualquer ambiguidade na sua determinação.

a)



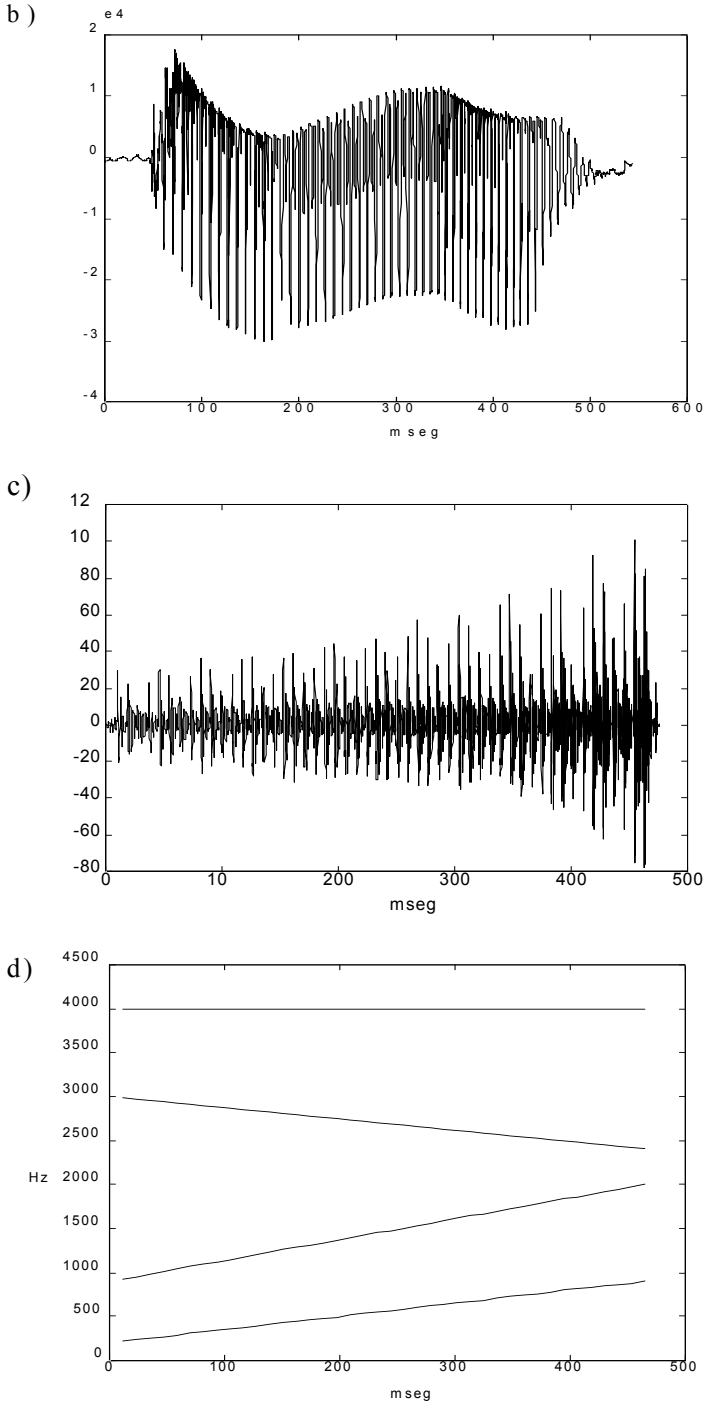


Figura 8.1 - Sinais usados na análise. a) sinal [i]. b) sinal correspondente à locução da palavra "ama". c) sinal sintetizado com larguras de banda constantes B1=80Hz, B2=90Hz, B3=100Hz, B4=120Hz, frequência fundamental de 100 Hz e amplitude unitária. d) variação das frequências formantes do sinal sintetizado.

Foram escolhidos os mesmos sinais para análise pelos diferentes métodos para permitir uma comparação dos resultados obtidos por estes e assim avaliar comparativamente cada um dos métodos.

Existem algumas variantes de ajuste de cada método como sejam o comprimento dos segmentos usados, utilização ou não da função janela de pesagem temporal, e em caso

afirmativo, que janela usar, número de pólos do modelo de predição etc. Deixa-se aqui um apontamento de quais as variantes óptimos para análise dos sinais de fala nas condições usadas e já descritas neste trabalho.

8.4.1 Método de Análise Cepstral

As figuras 8.2, 8.3 e 8.4 apresentam os parâmetros: frequência fundamental, formantes e larguras de banda extraídos automaticamente por este métodos bem como os espectrogramas para os sinais [i], "ama" e sinal sintetizado respectivamente.

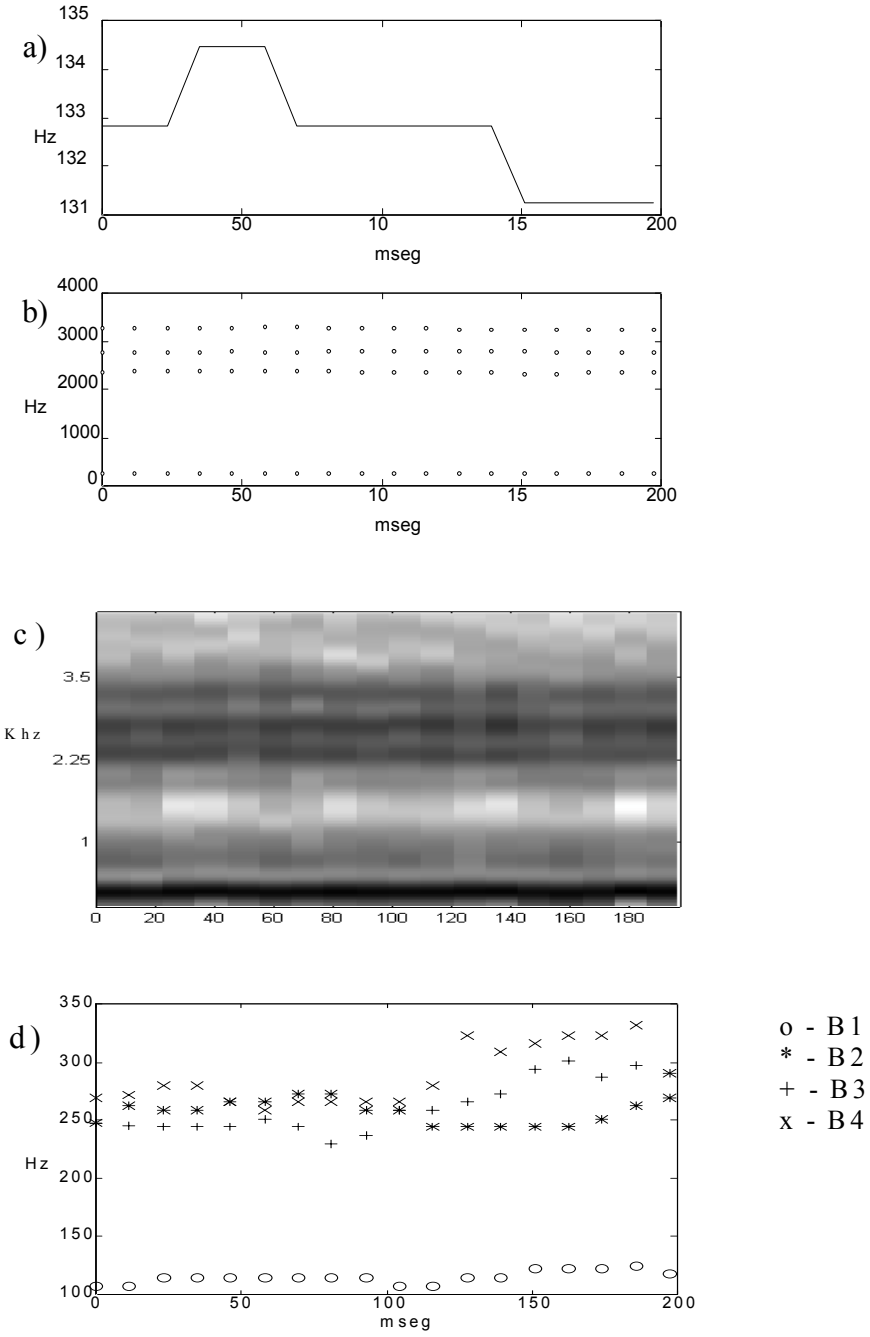


Figura 8.2 - Parâmetros extraídos automaticamente pelo método de análise cepstral para o sinal [i]. a) Variação da frequência fundamental. b) Variação das quatro formantes. c) Espectrograma. d) Larguras de banda.

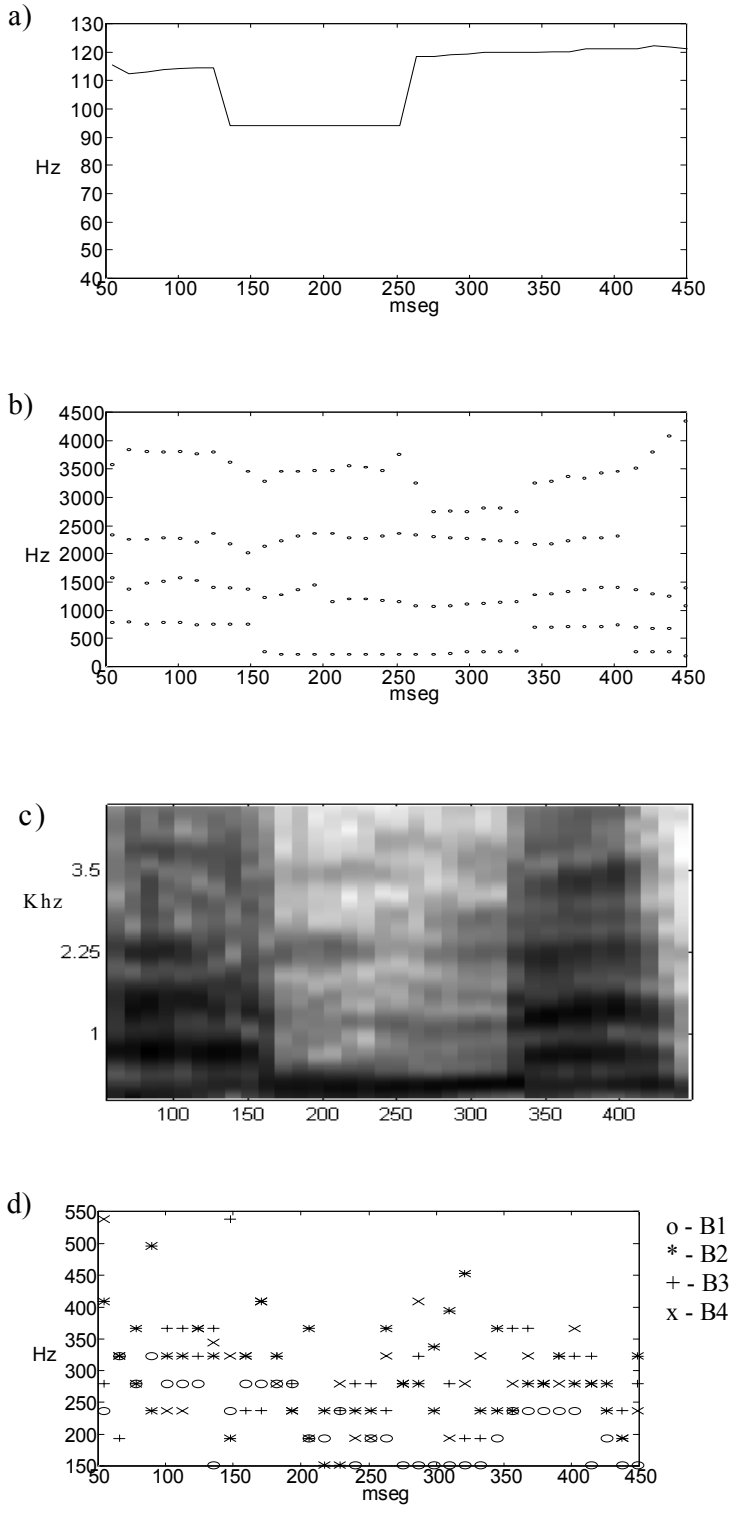


Figura 8.3 - Parâmetros extraídos automaticamente pelo método de análise cepstral para o sinal "ama". a) Variação da frequência fundamental. b) Variação das quatro formantes. c) Espectrograma. d) Larguras de banda.

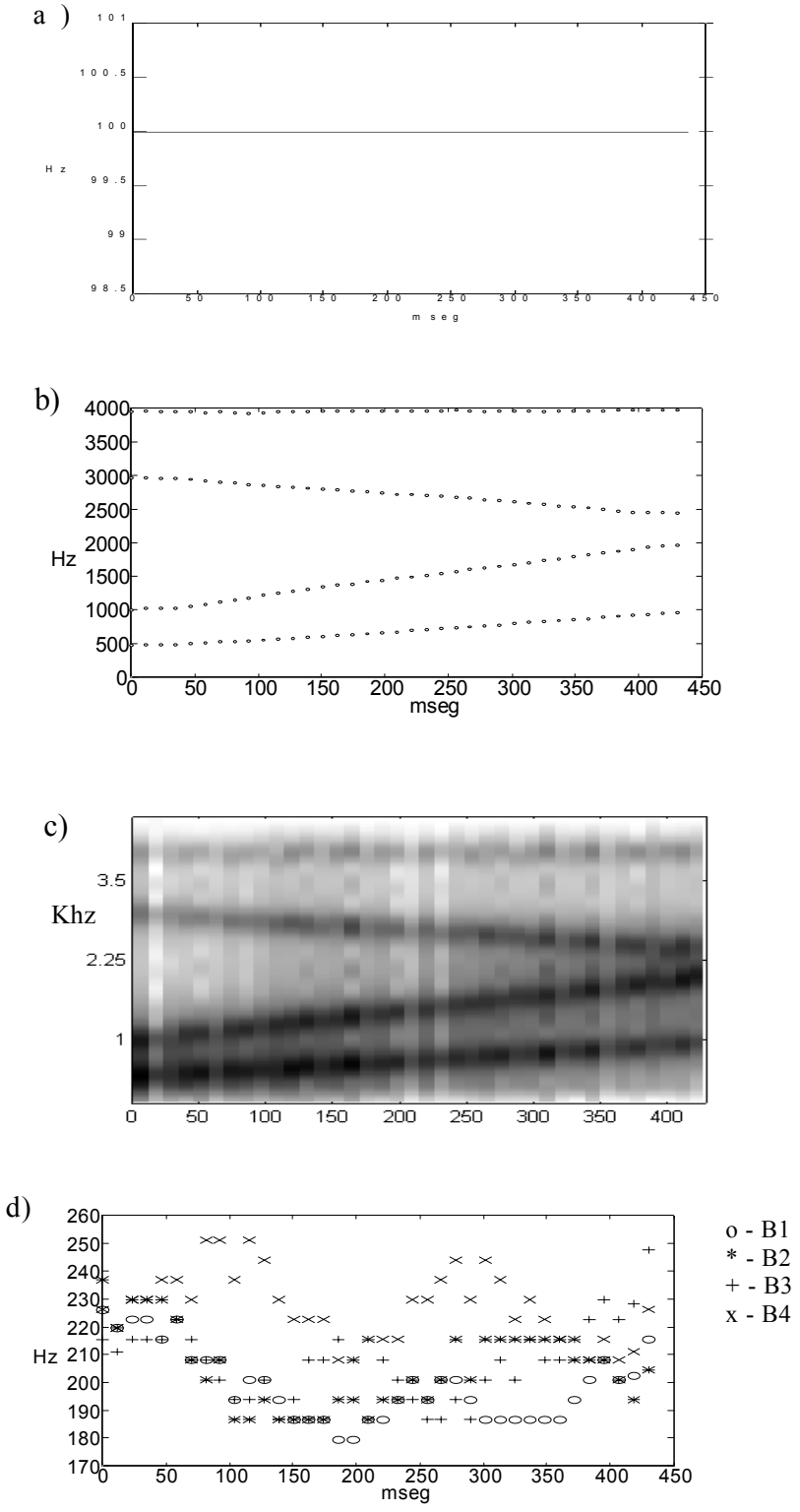


Figura 8.4 - Parâmetros extraídos automaticamente pelo método de análise cepstral para o sinal sintetizado. a) Variação da frequência fundamental. b) Variação das quatro formantes. c) Espectrograma. d) Larguras de banda.

Dos resultados observados nas três figuras anteriores para o método de análise cepstral e de outros testes realizados verifica-se que as frequências formantes extraídas automaticamente por este método seguem fielmente a variação das

formantes no espectrograma. A frequência fundamental determinada apresenta resultados conducentes com os valores obtidos por outros métodos já discutidos para determinação deste parâmetro bem como dos valores usados na síntese do último sinal analisado. As larguras de banda determinadas por este processo apresentam valores pouco credíveis para o sinal sintetizado, contudo, para o sinal [i] os valores destes parâmetros são muito próximos dos determinados pelos métodos a seguir testados. O alisamento espectral realizado por este método conduz a picos com formas parecidas e portanto com larguras de banda de valores idênticos.

Foram realizados testes de funcionalidade do sistema relativamente às variantes:

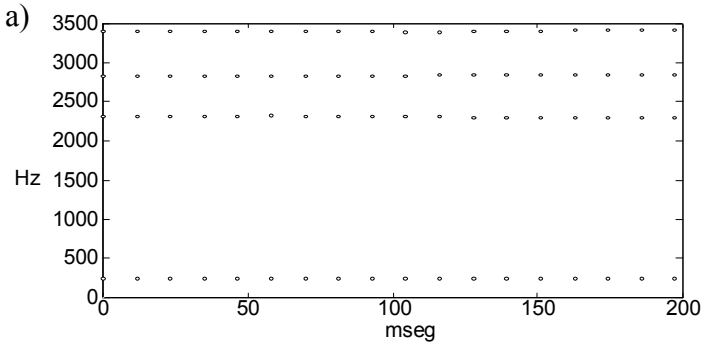
1 - Janela de pesagem temporal - foram testadas as variantes de não utilização de janela de pesagem, utilização da janela de Hanning e utilização da janela de Hamming sendo os melhores resultados obtidos com a janela de Hamming.

2 - Comprimento do segmento - foram testados segmentos de comprimentos de 100, 128, 200, 256, 300, 400, 512 e 1024 amostras sendo os resultados muito parecidos para comprimentos superiores a 256 amostras, tendo-se optado então, por segmentos de 256 amostras (segmentos de 23 ms, devido à sobreposição de segmentos de 50%, os parâmetros estão espaçados de 11.5 ms) permitindo assim um melhor acompanhamento da variação temporal dos parâmetros.

3 - Resolução frequencial - foram testadas FFT's de comprimentos de 128, 256, 400, 512, 1024 e 2048 independentemente do comprimento do segmento analisado devido à utilização da técnica de "zero padding" (acrescento de zeros à direita do segmento até igualar o comprimento da FFT) conseguindo-se bons resultados para comprimentos iguais ou superiores a 512. Assim, usaram-se FFT's de comprimento 512 (resolução frequencial de 21.5 Hz) por ser de cálculo mais rápido.

8.4.2 Método de Predição Linear - Matriz Autocorrelação

As figuras 8.5, 8.6 e 8.7 apresentam os parâmetros: formantes e larguras de banda extraídos automaticamente por este métodos bem como os espectrogramas para os sinais [i], "ama" e sinal sintetizado respectivamente.



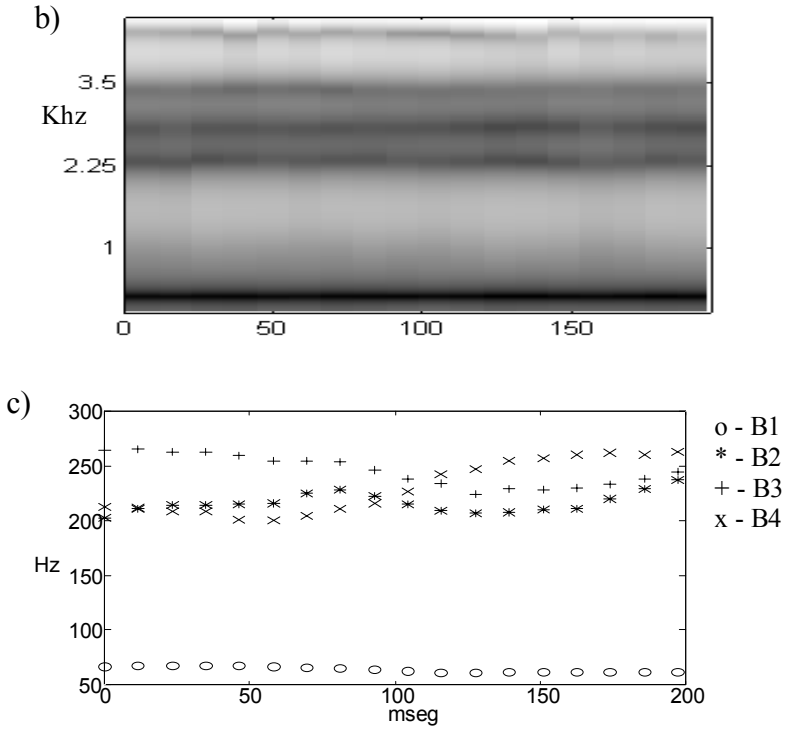
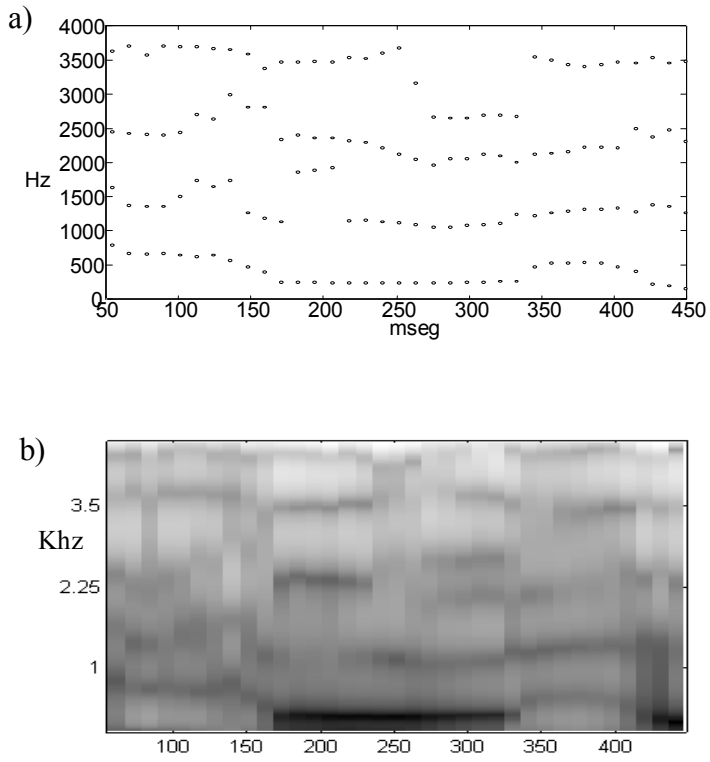


Figura 8.5 - Parâmetros extraídos automaticamente pelo método de LPC - matriz autocorrelação para o sinal [i]. a) Variação das quatro formantes. b) Espectrograma. c) Larguras de banda.



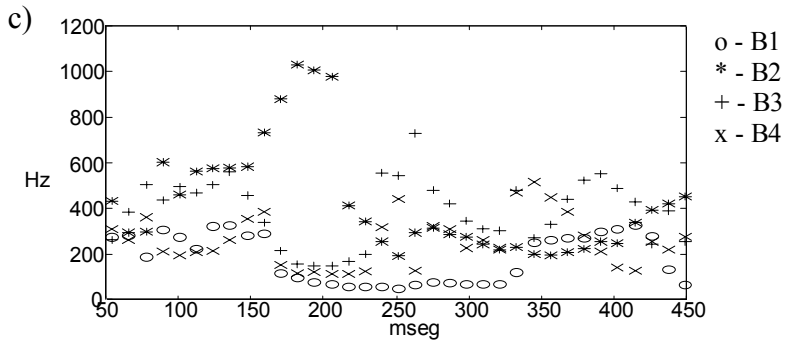


Figura 8.6 - Parâmetros extraídos automaticamente pelo método de LPC - matriz autocorrelação para o sinal "ama". a) Variação das quatro formantes. b) Espectrograma. c) Larguras de banda.

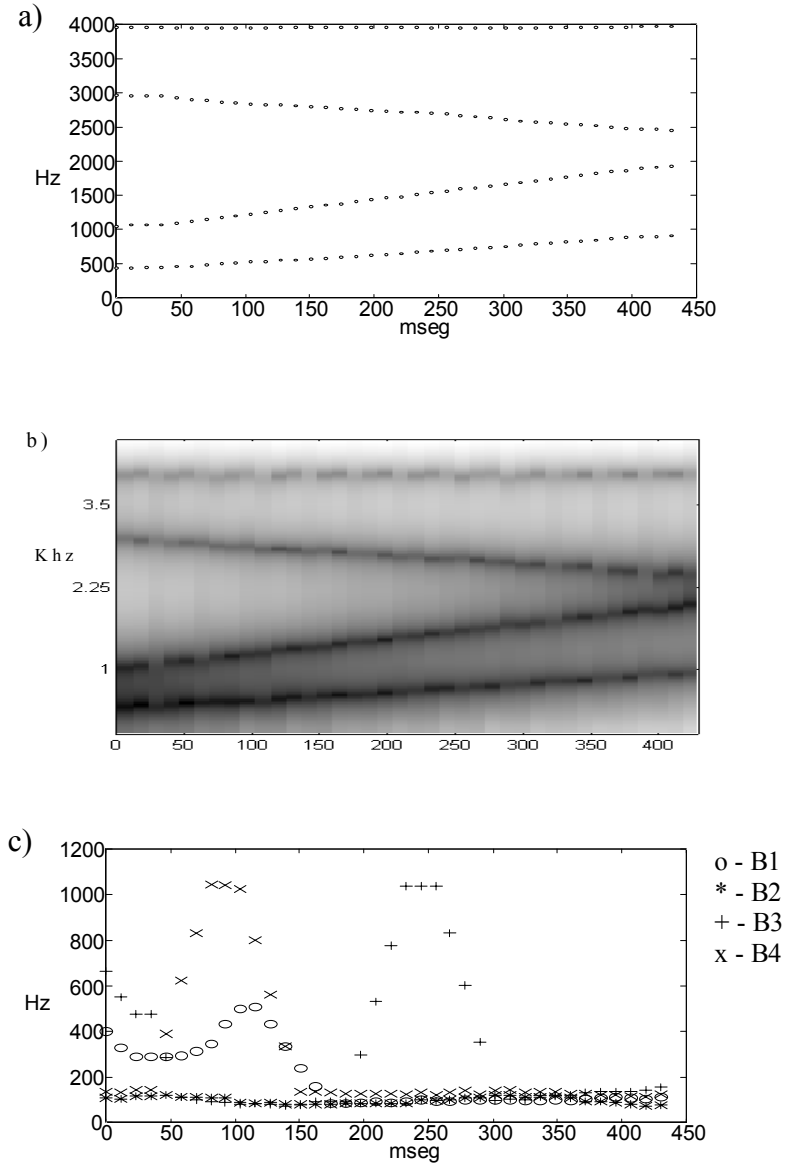


Figura 8.7 - Parâmetros extraídos automaticamente pelo método de LPC - matriz autocorrelação para o sinal sintetizado. a) Variação das quatro formantes. b) Espectrograma. c) Larguras de banda.

As frequências formantes extraídas por este método seguem fielmente as variações experimentadas no espectrograma. As larguras de banda determinadas são de facto proporcionais às largura das formantes visíveis nos espectrogramas, contudo, os seus valores afastam-se ligeiramente na parte inicial do sinal dos valores usados no sinal sintetizado.

Foram realizados testes de funcionalidade do sistema relativamente às variantes:

1 - Janela de pesagem temporal - foram testadas as variantes de não utilização de janela de pesagem, utilização da janela de Hanning e utilização da janela de Hamming sendo os melhores resultados obtidos com as funções das janelas de Hamming e Hanning indiferentemente.

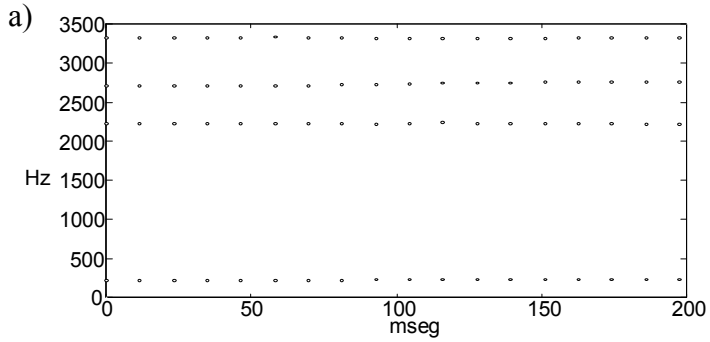
2 - Comprimento do segmento - foram testados segmentos de comprimentos de 100, 200, 256, 300, 400 e 512 amostras sendo os resultados muito parecidos para comprimentos superiores a 256 amostras, tendo-se optado então, por segmentos de 256 amostras (segmentos de 23 ms, devido à sobreposição de segmentos de 50%, os parâmetros estão espaçados de 11.5 ms) permitindo assim um melhor acompanhamento da variação temporal dos parâmetros.

3 - Número de pólos - foram testados modelos com 10, 11, 12, 13, 14 e 15 pólos sendo os melhores resultados obtidos para o modelo com 13 pólos.

A resolução frequencial usada na determinação da função de transferência do trato vocal para a representação do espectrograma foi de 21.5 Hz

8.4.3 Método de Predição Linear - Matriz Covariância

As figuras 8.8, 8.9 e 8.10 apresentam os parâmetros: formantes e larguras de banda extraídos automaticamente por este métodos para os sinais [i], "ama" e sinal sintetizado respectivamente. Não se apresentam os espectrogramas relativos a cada sinal obtidos por este método de análise por serem muito semelhantes aos apresentados para o método da matriz autocorrelação.



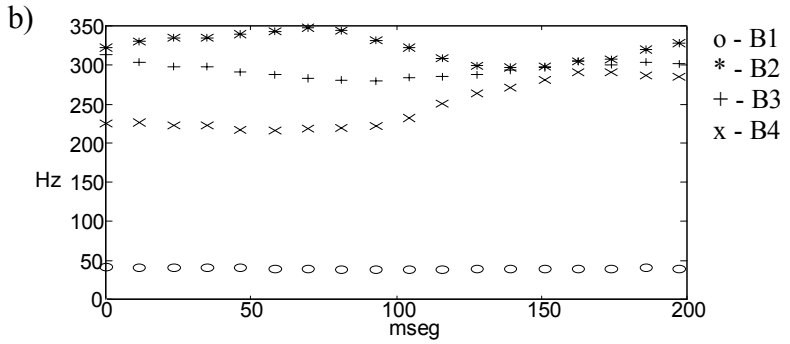


Figura 8.8 - Parâmetros extraídos automaticamente pelo método de LPC - matriz covariância para o sinal [i]. a) Variação das quatro formantes. b) Larguras de banda.

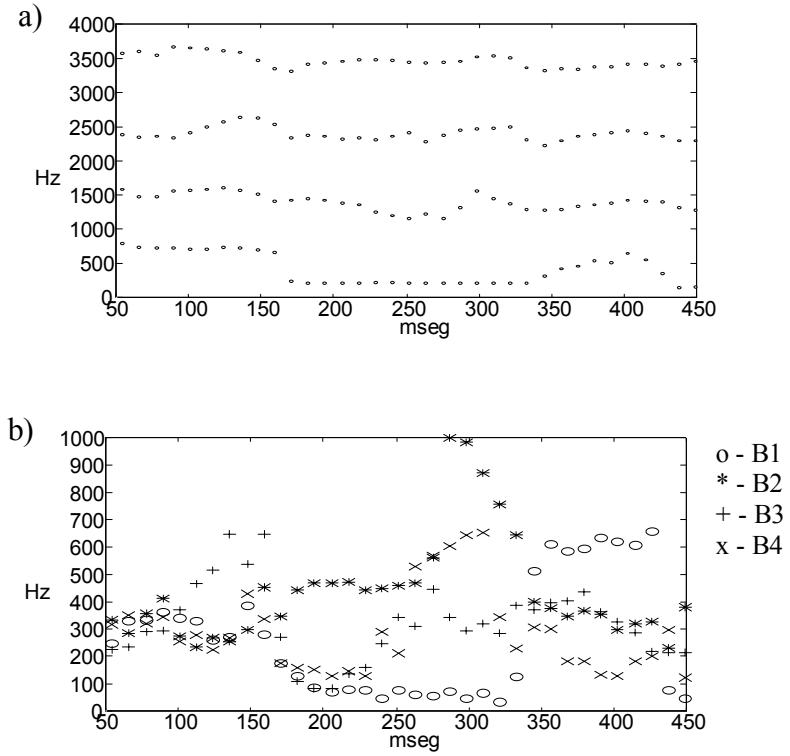


Figura 8.9 - Parâmetros extraídos automaticamente pelo método de LPC - matriz covariância para o sinal "ama". a) Variação das quatro formantes. b) Larguras de banda.

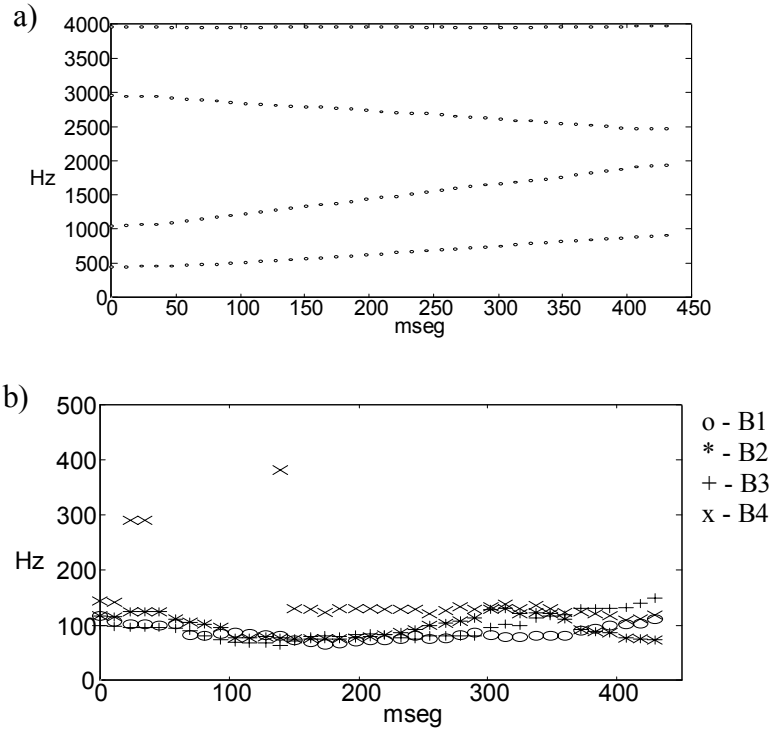


Figura 8.10 - Parâmetros extraídos automaticamente pelo método de LPC - matriz covariância para o sinal sintetizado. a) Variação das quatro formantes. b) Larguras de banda.

Tal como para o método da matriz autocorrelação os valores estimados das frequências formantes seguem fielmente as formantes visíveis nos espectrogramas. As larguras de banda determinadas por este método são mais próximas dos valores reais, como se verifica no caso do sinal sintetizado. Para o sinal [i] há uma concordância por todos os métodos experimentados na variação dos valores das larguras de banda.

Foram realizados testes de funcionalidade do sistema relativamente às variantes:

1 - Janela de pesagem temporal - foram testadas as variantes de não utilização de janela de pesagem, utilização da janela de Hanning e utilização da janela de Hamming sendo os melhores resultados obtidos sem função de janela de pesagem temporal.

2 - Comprimento do segmento - foram testados segmentos de comprimentos de 100, 200, 256, 300, 400, 512 e 1024 amostras sendo os resultados muito parecidos para comprimentos superiores a 256 amostras, tendo-se optado então, por segmentos de 256 amostras (segmentos de 23 ms, devido à sobreposição de segmentos de 50%, os parâmetros estão espaçados de 11.5 ms) permitindo assim um melhor acompanhamento da variação temporal dos parâmetros.

3 - Número de pólos - foram testados modelos com 10, 11, 12, 13, 14 e 15 pólos sendo os melhores resultados obtidos para o modelo com 13 pólos.

A resolução frequencial usada na determinação da função de transferência do trato vocal para a representação do espectrograma foi de 21.5 Hz

8.4.4 Método de Análise Síncrona com o Período Fundamental

As figuras 8.11, 8.12 e 8.13 apresentam os parâmetros: formantes e larguras de banda extraídos automaticamente por este método para os sinais [i], "ama" e sinal sintetizado respectivamente. Não se apresentam os espectrogramas relativos a cada sinal obtidos por este método de análise por serem muito semelhantes aos apresentados para o método da matriz autocorrelação.

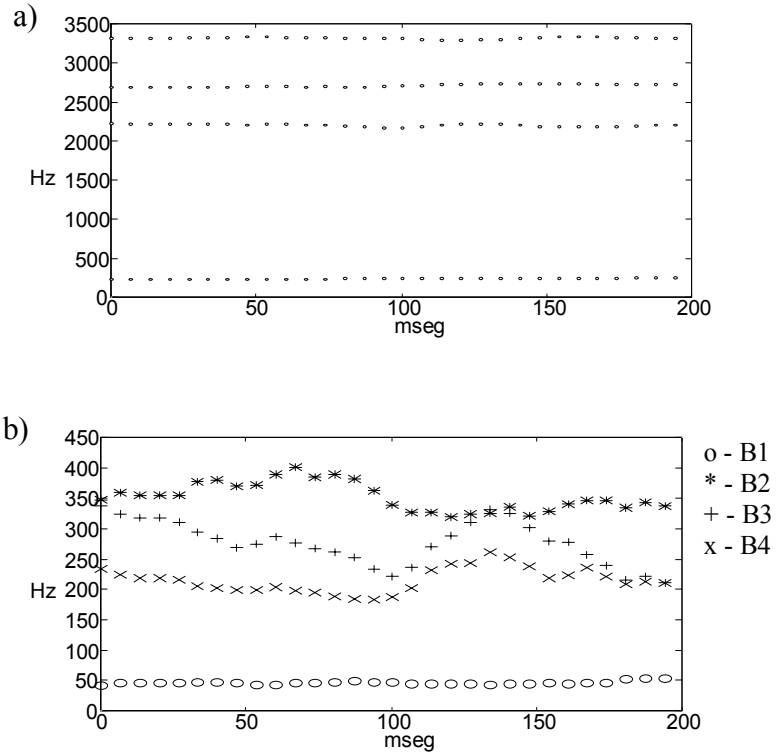
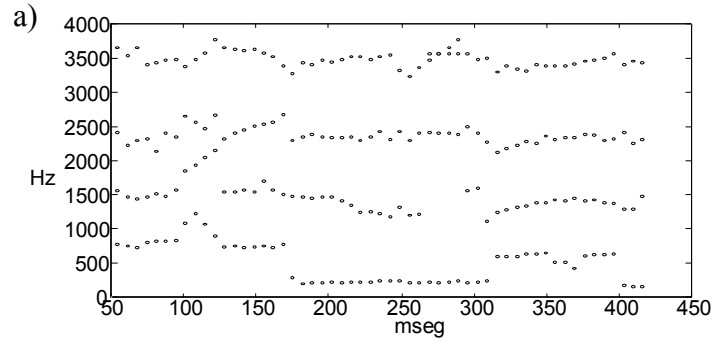


Figura 8.11 - Parâmetros extraídos automaticamente pelo método de análise síncrona (LPC - matriz covariância) para o sinal [i]. a) Variação das quatro formantes. b) Larguras de banda.



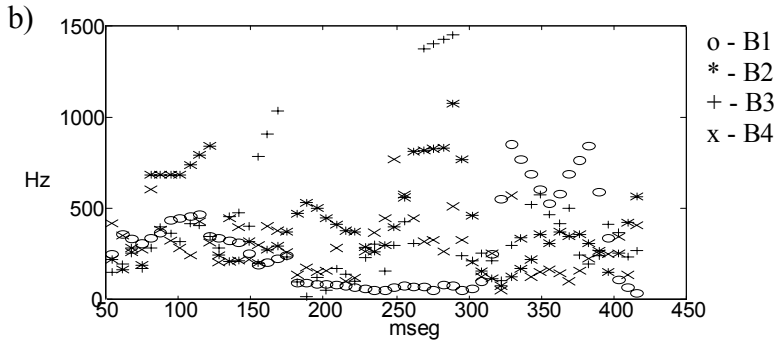


Figura 8.12 - Parâmetros extraídos automaticamente pelo método de análise síncrona (LPC - matriz covariância) para o sinal "ama". a) Variação das quatro formantes. b) Larguras de banda.

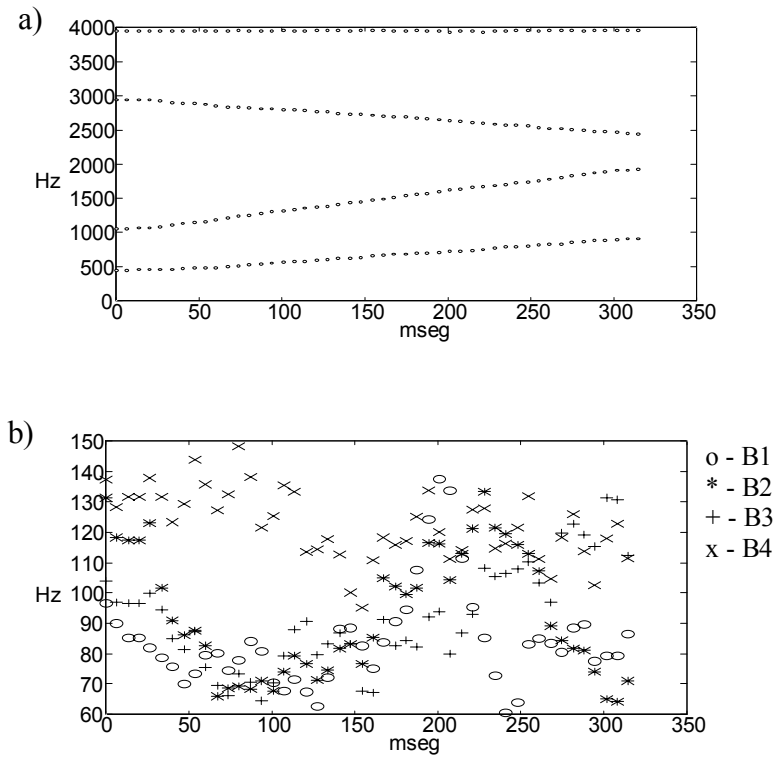


Figura 8.13 - Parâmetros extraídos automaticamente pelo método de análise síncrona (LPC - matriz covariância) para o sinal sintetizado. a) Variação das quatro formantes. b) Larguras de banda.

Os valores das frequências formantes estimadas por este métodos também seguem fielmente os respectivos valores das formantes no espectrograma. As larguras de banda são estimadas neste método com valores um pouco diferentes dos valores obtidos para os outros métodos, contudo, para o sinal [i] a variação destes valores continua a ter a mesma forma.

Foram realizados testes de funcionalidade do sistema relativamente às variantes:

1 - Janela de pesagem temporal - foram testadas as variantes de não utilização de janela de pesagem, utilização da janela de Hanning e utilização da janela de Hamming sendo os melhores resultados obtidos sem função de janela de pesagem temporal.

2 - Início do segmento relativamente ao início do impulso glotal - foram testados inícios do segmento relativamente a ini, início do impulso glotal, de ini-20, ini-10, ini, ini+10, ini+20 e ini+40 tendo os melhores resultados sido obtidos para a posição de início do segmento em ini-10 e ini. Assim, foi usada a posição ini. Não se atribuiu muita importância a este resultado por simultaneamente haver uma variação significativa do comprimento do segmento a analisar com influência na estabilidade e precisão dos resultados obtidos pelo modelo.

3 - Número de pólos - foram testados modelos com 10, 11, 12, 13, 14 e 15 pólos sendo os melhores resultados obtidos para o modelo com 13 pólos.

A resolução frequencial usada na determinação da função de transferência do trato vocal para a representação do espectrograma foi de 21.5 Hz

8.4.5 Comparação de Resultados dos Diferentes Métodos de Análise

Com base nos resultados aqui apresentados e em muitos outros testes realizados com os diferentes métodos de análise desenvolvidos neste trabalho é com convicção que se pode afirmar que a extracção automática das frequências formantes é realizada com sucesso por todos os métodos, isto é, as suas variações ao longo do tempo seguem as variações mostradas pelo espectrograma relativo a cada método. Os espectrogramas obtidos pelos diferentes métodos têm formas muito parecidas apesar de haver ligeiras diferenças de aspecto especialmente entre os espectrogramas baseados nas funções de transferência determinadas pelos métodos de predição linear e os espectrogramas baseados nos espectros alisados por análise cepstral.

No que diz respeito à determinação automática das larguras de banda, é visível que nem sempre há uma convergência de valores ou formas da sua variação pelos diferentes métodos. O sinal correspondente à locução continuada de um [i] proporcionou aos diferentes métodos uma determinação das larguras de banda com formas idênticas e valores próximos, contudo isso não aconteceu para outros sinais menos estáveis. Para o sinal com valores de larguras de banda conhecidos previamente, o sinal sintetizado, as melhores estimativas para as larguras de banda foram obtidas pelo método de predição linear matriz covariância com valores muito próximos dos sintetizados. Relativamente ao sinal "ama", os valores das larguras de banda estimados pelos 4 métodos são em algumas partes do sinal próximos e noutras partes diferentes, contudo os valores reais destes parâmetros para este sinal são desconhecidos não se podendo portanto concluir com base neste sinal acerca do melhor método na estimação das larguras de banda.

O parâmetro frequência fundamental foi para estes exemplos correctamente determinado pelo método do cepstro.

A utilização de qualquer dos métodos desenvolvidos para determinação automática de parâmetros neste trabalho não oferece qualquer dificuldade ao utilizador já que basta chamar o nome do programa *falacont* para os três primeiros métodos ou *falasinc* para

o método de análise síncrona sendo todo o processamento transparente para o utilizador.

O tempo de processamento não foi um aspecto tomado em consideração no desenvolvimento deste trabalho. Neste estado de desenvolvimento, o método do cepstro é o mais rápido. Os métodos de predição linear sem análise síncrona têm tempos de computação muito próximos. A análise síncrona com o período fundamental é relativamente mais lenta devido ao facto de processar segmentos mais curtos e portanto mais segmentos para o mesmo sinal. Não deve ser dada qualquer importância aos tempos de processamento relativos entre os métodos experimentados, já que os métodos de predição linear têm potencialidades de funcionamento em tempo muito mais reduzido recorrendo a outras técnicas mais rápidas para a multiplicação matricial e determinação das raízes de polinómios ([Snell 93], [Reddy 84], [McCandless 74], [Denoël 85]) em vez da implementação em Matlab. Os métodos de predição linear são passíveis de implementação em hardware podendo funcionar em tempo real.

CAPÍTULO 9

CONCLUSÕES E DESENVOLVIMENTOS FUTUROS

9. CONCLUSÕES E DESENVOLVIMENTOS FUTUROS

Neste capitulo que encerra a dissertação apresentam-se, de forma resumida os principais passos abordados e suas conclusões dentro do trabalho desenvolvido. Pretende-se também deixar um alerta para as linhas de investigação que ficam em aberto neste trabalho e fazer um apontamento de algumas aplicações ligadas ao tema desenvolvido.

9.1 Conclusões

O trabalho desta dissertação concentrou-se no estudo e desenvolvimento de ferramentas de análise dos sinais de fala para extracção automática dos parâmetros dos modelos usados para estes sinais com o objectivo de realizar a respectiva síntese.

A motivação para o tema deste trabalho provém da experiência acumulada no desenvolvimento de um sistema de conversão texto fala para o português apresentado no capítulo 5, em que se sente a necessidade de sistemas de análise de fala que apresentem de forma objectiva os valores dos parâmetros, frequências formantes e larguras de banda, presentes num som de fala, bem como as suas variações nas transições entre sons, no sentido de facilmente se estabelecerem regras de concatenação de fonemas baseadas em unidades acústicas estabelecidas. Os resultados atingidos, globalmente considerados, permitem concretizar o objectivo enunciado.

Num trabalho desta natureza é indispensável um conhecimento profundo do objecto de trabalho, a fala. No capítulo 2 é discutido o processo de produção de fala com o aparelho fonador humano assim como se procura caracterizar do ponto de vista acústico os sons existentes na língua portuguesa para uma melhor percepção dos modelos envolvidos em cada tipo de sinais de fala. É reconhecida uma elevada importância em alguns traços distintivos acústicos para a caracterização dos sons da fala como sejam sons vocalizados, sons nasalizados e fricativas. Com base nas características destes diferentes tipos de sinais estabeleceram-se 3 classes de sinais com modelos diferentes. Indicaram-se ainda, os conjuntos de parâmetros mais relevantes nestes modelos.

Assim, considera-se um modelo genérico com uma fonte de excitação e um filtro que modela o trato vocal. A fonte de excitação consiste num sinal periódico de impulsos glotais para os sinais vocalizados, um sinal de ruído branco para os sinais não vocalizados e um sinal composto pela sobreposição dos dois anteriores para sinais de excitação mista. O modelo do trato vocal toma também configurações diferentes consoante o tipo de sinal. Para os sinais vocalizados tem a forma de um filtro só com ressonâncias, para os sinais não vocalizados consistirá num filtro com uma ressonância e uma anti-ressonância e, para os sinais de excitação mista e sons nasais poderá ter a configuração de um filtro com uma anti-ressonância e várias ressonâncias.

Estes modelos são desenvolvidos no capítulo 4, onde também se consideram os efeitos do impulso glotal e da radiação nos lábios. Ainda neste capítulo são apresentadas algumas técnicas de análise usadas neste trabalho como sejam a análise cepstral, predição linear, análise síncrona com o período fundamental e análise por síntese. De uma maneira geral as três primeiras técnicas produzem resultados bons, embora sejam individualmente óptimas em determinadas situações.

No 6º capítulo são discutidas algumas ferramentas usadas no sistema de extracção automática dos parâmetros como a média deslizante, a energia deslizante e a taxa de passagem por zero, estas com particular importância no processo de classificação de segmentos de fala num dos tipos considerados (vocalizada, não vocalizada e excitação mista) ou reconhecendo as partes do sinal correspondentes a silêncio. Discutem-se neste capítulo processos para segmentar/classificar os sinais de fala. Propõe-se um processo de classificar semisegmentos de sinais de fala baseado na taxa de passagem por zero da derivada do sinal e na energia deslizante recorrendo a um domínio de decisão criado com dois vectores correspondentes ao número de vezes que a energia de uma parte do sinal está abaixo do nível considerado máximo para o silêncio e o número de vezes que a taxa de passagem por zero da derivada da mesma parte do sinal está acima do nível considerado mínimo para o silêncio, classificando assim o semisegmento em vocalizado, não vocalizado, de excitação mista ou silêncio.

Na escolha da dimensão dos segmentos, não foi tomada em atenção a eventual necessidade de redução da taxa de transmissão. Contudo, este aspecto não está posto de lado, sabendo-se que quanto maiores forem os segmentos menor será a taxa de transmissão, no entanto também a qualidade tende a degradar-se. Seria necessário estudar a melhor situação de compromisso entre estes dois factores.

São ainda discutidos neste capítulo os processos de estabelecimento das matrizes autocorrelação e covariância usadas na predição linear para determinação dos coeficientes de predição linear do segmento do sinal de fala, bem como as potencialidades da análise cepstral e a utilização de ferramentas computacionais como a FFT (Fast Fourier Transform) e CZT (Chirp Z Transform) para realizar o "zoom" de uma parte do espectro alisado pelo método de análise cepstral.

Foram ainda criados um sintetizador de formantes, computacional, para testar os parâmetros extraídos automaticamente pelo processo de análise e um espectrógrafo que desenha os envelopes espectrais do trato vocal, determinados pelo processo de análise, sob a forma de espectrogramas.

No capítulo 7 são discutidas algumas técnicas para determinação automática da frequência fundamental de sinais de fala vocalizada, mais especificamente técnicas baseadas na análise cepstral, com que foram obtidos bons resultados, predição linear e um algoritmo com processamento exclusivamente no domínio temporal com o qual se obtêm 6 estimativas da frequência fundamental. A combinação destas 6 estimativas e o conhecimento do parâmetro para o segmento anterior resulta numa estimativa muito fiel da frequência fundamental do sinal adquirido nas condições mais adversas.

É ainda neste capítulo que se desenvolvem os algoritmos para extracção automáticas dos parâmetros dos sinais de fala vocalizada e não vocalizada, pelos diversos métodos.

Para o modelo de fala vocalizada foram desenvolvidos 4 métodos. O primeiro baseado na análise cepstral, recorre a esta ferramenta para separar as características do trato vocal das características da fonte excitadora, determinando a frequência fundamental e realizando o alisamento espectral. Com base no espectro alisados é implementado um algoritmo de detecção dos formantes nos picos do espectro recorrendo à expansão de uma região de frequências deste para distinguir dois formantes quando estes se encontram muito próximos. As larguras de banda são determinadas directamente no espectro alisado a -3 dB do pico correspondente a cada formante.

Os segundo e terceiro métodos recorrem às técnicas de predição linear pelas matrizes autocorrelação e covariância para determinar os coeficientes de predição por multiplicação matricial. Os pólos são obtidos por determinação das raízes do polinómio com os coeficientes referidos. Cada par de pólos conjugados é considerado como uma formante sendo depois seleccionadas as 4 formantes mais evidentes.

O quarto método, de análise síncrona com o período fundamental recorre a uma função para detecção de sincronismo com o impulso glotal sendo neste caso cada segmento do sinal considerado desde o início de um impulso até ao início do próximo impulso. Os parâmetros de cada segmento são extraídos também pela mesma técnica de predição linear pelo método da matriz covariância.

As sequências de parâmetros extraídos ao longo do tempo são finalmente sujeitas a um alisamento não linear para corrigir eventuais pontos nitidamente fora de uma linha seguida pelos pontos vizinhos.

Qualquer destes métodos determina com fidelidade as frequências formantes dos sinais, contudo, a determinação das larguras de banda é um pouco menos feliz para o método de análise cepstral. O método de predição linear pela matriz covariância é o que determina valores para as larguras de banda mais próximas dos usados num sinal sintetizado, reunindo por isso e pela sua maior estabilidade para sinais não vocalizados, a preferencia dos métodos experimentados. O método de análise síncrona com o período fundamental nem sempre resulta num sistema estável, possuindo alguns pólos fora do circulo unitário, devido ao número de amostras analisadas ser reduzido.

Foi também desenvolvido um algoritmo para determinação das frequências do zero e do pólo do modelo para fala não vocalizada que apresenta um funcionamento capaz.

No capítulo 3 foram discutidas as condições ideais para a recolha do sinal e as condições em que se realizou essa recolha neste trabalho, podendo concluir-se que são de facto importantes os cuidados a ter relativamente ao aumento da relação sinal/ruído, nomeadamente através do uso de aparelhagem e cablagem de qualidade e utilização de sala insonorizada para a recolha do sinal. Reconhece-se ainda a importância do uso de um filtro para cortar as frequências abaixo de sensivelmente 70 Hz, eliminando a componente DC. Este aspecto é relevante no processo de classificação dos segmentos de sinal.

Por último, refere-se o desenvolvimento de um sistema de conversão texto-fala para o português baseado no conversor multilíngua MULTIVOX apresentado no capítulo 5. Este sistema sintetiza fala, a partir de um texto escrito, com qualidade inteligível reconhecendo-se-lhe condições para produzir fala natural com a continuação do desenvolvimento deste sistema.

9.2 Desenvolvimentos Futuros

Ao longo deste trabalho várias questões foram deixadas em aberto. Em alguns casos, a procura de respostas para um problema levantou novas interrogações, novos caminhos para serem percorridos. Contudo, não se vai olhar para trás à procura do que ficou por responder, mas sim dar uma visão global de um sistema completo de análise dos sinais de fala.

Assim, no seguimento do tema desenvolvido nesta dissertação considera-se que um sistema completo de análise consiste numa primeira fase de segmentação e

classificação de cada segmento nos diferentes tipos de sinal e então a análise realizada com base no modelo considerado para o tipo de segmento em causa.

A primeira fase foi discutida neste trabalho tendo sido também apontado um método para classificação/segmentação do sinais de fala reconhecendo-se a necessidade de mais desenvolvimento nesta matéria usando o processo proposto ou outro processo para atingir uma correcta classificação de todos os segmentos procurando evitar que segmentos de sinal tentem ser modelados com modelos que não correspondem ao tipo de sinal em causa.

A segunda fase do sistema, extracção dos parâmetros do modelo com recurso à análise de cada segmento foi aqui desenvolvido especialmente para os segmentos vocalizados e não vocalizados, no entanto, os sons nasais e fricativas vocalizadas não são correctamente adaptados a estes modelos reconhecendo-se a necessidade de estudo e desenvolvimento de modelos e processos para analisar estes tipos de sinais.

As ferramentas aqui desenvolvidas têm utilidade não só nos temas de análise e síntese de fala como são também de relevante interesse para o reconhecimento da fala e do falante. As matérias aqui desenvolvidas têm interesses crescentes nas áreas de reabilitação, comunicações, industria e na "interface" de comunicação com o computador.

Não se poderia terminar este trabalho sem deixar de referir as importantes e nobres funções dos sistemas de conversão texto-fala na ajuda a pessoas incapacitadas, quer da visão quer da própria fala. É hoje imperioso o desenvolvimento e aperfeiçoamento de programas que usando estes conversores facilitem a comunicação destas classes de deficientes quer com o computador quer com o exterior, um pouco à semelhança do que acontece com os chamados "Screen Reader's" para invisuais, ou o caso específico do VOXAID para ajuda a deficientes da fala usando o conversor MULTIVOX. Estes sistemas permitirão, sem dúvida, melhorar a desfavorecida qualidade de vida destas pessoas abrindo-lhes novos mundos e criando a possibilidade de se tornarem mais independentes atenuando, de certa forma, o seu "handicap".

BIBLIOGRAFIA

BIBLIOGRAFIA

[Alkin 94] "Digital Signal Processing, A Laboratory Approach Using PC-DSP", Second Edition Prentice-Hall 1994.

[Bracewell 86] "The Fourrier Transform And Its Applications" -R. N. Bracewell, McGRAW-HILL International Editions - Electrical & Electronic Engineering Series 1986.

[Burrus 94] "Computer-Based Exercises for Signal Processing Using Matlab", S. Burrus, J. McClellan, A. Oppenheim, T. Parks, R. Shafer, H. Schuessler, Prentice-Hall International Editions 1994.

[Chomsky 68] "The Sound Pattern of English", Chomsky, Halle, The Hague, Mouton 1968.

[Cost209 88] "Man-Machine Communication by Means of Speech Signals", European Research Project - Cost 209 March 1988.

[Denoël 85] "Linear Prediction of Speech with a Least Absolute Error Criterion", E. Denoël, J. F. Solvay, IEEE Transactions on Acoustics Speech, and Signal Processing, vol. ASSP-33, Nº6, December 1985.

[Dunn 61] "Methods of Measuring Vowel Formant Bandwidths", H. K. Dunn, The Journal of Acoustical Society of America, vol.33 December 1961.

[Flanagan 56] "Automatic Extraction of Formant Frequencies from Continuous Speech", J. L. Flanagan, The Journal of Acoustical Society of America, vol.28 January 1956.

[Flanagan 64] "Digital Equalizer and De-Equalizer for Speech", J. L. Flanagan, D. Meinhart, P. Cummiskey, The Journal of Acoustical Society of America, vol.36 1964.

[Freitas 93] Apontamentos da Cadeira de Sistemas de Instrumentação II do curso de Licenciatura em Engenharia Electrotécnica e Computadores da Faculdade de Engenharia da Universidade do Porto, Diamantino Freitas 1993.

[Gold 69] "Parallel Processing Techniques for Estimating Pitch Periods os Speech in the Time Domain", B. Gold, L. Rabiner, The Journal of Acoustical Society of America, vol.46 Nº 2 (part 2) 1969.

[Gósy 91] "Temporal Factors in Speech", Mária Gósy, A Collection of

Papers 1991.

[Griffin 84] "Signal Estimation From Modified Short-Time Fourrier Transform", D. Griffin, J. S. Lim, IEEE Transactions on Acoustics Speech, and Signal Processing, vol. ASSP-32, Nº2, April 1984.

[Jakobson 63] "Preliminaries to Speech Analysis", Jakobson, Fant, Halle, Cambridge, Mass., MIT Press 1963.

[Jayant 84] "Digital Coding of Waveforms, Principles and Applications to Speech and Video", N. S. Jayant, P. Noll, Prentice-Hall Signal Processing Series 1984.

[Kabal 89] "Join Optimization of Linear Predictors in Speech Coders", P. Kabal, R. Ramachandran, IEEE Transactions on Acoustics Speech, and Signal Processing, vol. 37, Nº5, May 1989.

[Kay-Fu 89] "Automatic Speech Recognition", Kai-Fu Lee, Kluwer Academic Publishers 1989.

[Keller 94] "Fundamentals of Speech Synthesis and Speech Recognition - Basic Concepts, State of the Art and Future Challenges", Eric Keller - JONH WILEY & SONS 1994

[Krishnamurthy 86] "Two-Channel Speech Analysis", A. Krishnamurthy, D. Childers, IEEE Transactions on Acoustics Speech, and Signal Processing, vol. ASSP-34, Nº4, August 1986.

[López 93] "Estudio de Técnicas de Processado Linguístico y Acústico Para Sistemas de Conversión Texto-Voz en Espanhol Baseado en Concatenatión de Unidades", E. López G. Tesis Doctoral - Universidad Politécnica de Madrid 1993.

[Makhoul 73] "Spectral Analysis of Speech by Linear Prediction", J. Makhoul, Transactions on Audio and Electroacoustics vol. AU-21, Nº 3 June 1973.

[Makhoul 75] "Linear Prediction: A Tutorial Review", J. Makhoul, Proceedings of the IEEE, April 1975.

[Malmberg 54] "A Fonética", Bertil Malmberg 1954, tradução de Oliveira Figueiredo, Edição Livros do Brasil - Lisboa.

[Markel 82] "Linear Prediction of Speech", J. D. Markel, A. H. Gray, Springer-Verlag Berlin Heidelberg New York 1982.

[Marques 89] "Frequency Varying Sinusoidal Modelling of Speech", J. Marques, L. B. Almeida, IEEE Transactions on Acoustics Speech, and Signal Processing, vol. ASSP-37, Nº 5, May 1989.

[Marques 90] "Modelamento Sinusoidal da Fala - Aplicação à Codificação a Ritmos Médios e Baixos", J. Marques, Tese de Doutoramento, Instituto Superior Técnico, Universidade Técnica de Lisboa, Fevereiro 1990.

[Martins 92] "Ouvir Falar - Introdução à Fonética do Português", M. R. Delgado Martins, segunda edição, Caminho Colecção Universitária série Linguística 1992.

[Mateus 90] "Fonética, Fonologia e Morfologia do Português", M. H. Mira Mateus, A. Andrade, M. do Céu Viana, A. Villalva, Universidade Aberta Lisboa 1990.

[Mathews 61] "Pitch Synchrous Analysis of Voiced Sounds", M. Mathews, J. Miller, E. David, Journal of Acoustical Society of America, vol.33 Feb. 1961.

[McAulay 84] "Maximum Likelihood Spectral Estimation and its Applications to Narrow-Band Speech Coding", R. McAulay, IEEE Transactions on Acoustics Speech, and Signal Processing, vol. ASSP-32, Nº2, April 1984.

[McAulay 86] "Speech Analysis/Synthesis Based on a Sinusoidal Representation", R. McAulay, T. Quatieri, IEEE Transactions on Acoustics Speech, and Signal Processing, vol. ASSP-34, Nº4, August 1986.

[McCandless 74] "An Algorithm for Automatic Formant Extraction Using Linear Prediction Spectra", S. McCandless, IEEE Transactions on Acoustics Speech, and Signal Processing, vol. ASSP-22, Nº2, April 1974.

[Medan 91] "Super Resolution Pitch Determination of Speech Signal", Y. Medan, E. Yair, D. Chazan, IEEE Transations on Signal Processing, vol. 39, Nº 1, January 1991.

[Miyoshi 87] "Analysis of Speech Signals of Short Pitch Period by a Sample Selective Linear Prediction", Y. Miyoshi, Yamato, Mizoguchi, Yanagida, Kakusho, IEEE Transactions on Acoustics Speech, and Signal Processing, vol. ASSP-35, Nº 9, September 1987.

[Morgan 84] "Talking Chips, IC Speech Synthesis", N. Morgan, National Semiconductors Technology Series 1984.

[Morgan D. P. 91] "Neural Networks and Speech Processing", D. P. Morgan, C. L. Scofield, Kluwer Academic Publishers 1991.

[Niederjohn 85] "A Zero CrossingConsistency Method for Formant Tracking of Voiced Speech in High Noise Levels", R. J. Niederjohn, M. Lahat - IEEE Transactions on Acoustics Speech, and

Signal Processing, vol. ASSP-33, Nº 2, April 1985.

[Noll 67] "Cepstrum Pitch Determination", A. M. Noll, The Journal of Acoustical Society of America, vol.41 Nº 2 1967.

[Olaszy 90] "Phonetic aspects of the MULTIVOX text-to-speech system", Olaszy, G. Gordos, G. Németh, G Proc. of the ESCA Workshop on Speech Syntesis, Autrans 1990, pag. 277-280.

[Olaszy 92] "The MULTIVOX Multilingual text-to-speech converter", G. Olaszy, G. Gordos, G. Németh, Talking Machines 1992.

[Oppenheim 68] "Homomorphic Analysis of Speech", A. Oppenheim, R. W. Schafer, IEEE Transations on Audio Electroacoustic vol. AU-16 June 1968.

[Papoulis 84] "Signal Analysis", A. Papoulis, McGRAW-HILL International Editions - Electrical & Electronic Engineering Series 1984.

[Perckette 95] "Formant-BasedAudio Synthesis Using Nonlinear Distortion", M. Perckette, Journal Audio Eng. Soc., vol. 43, Nº 1/2 January/February 1995.

[Pinheiro 93] "Técnicas de Codificação da Voz", J. J. Pinheiro, Sandra Machado, Trabalho realizado no âmbito da cadeira de Sistemas de Instrumentação I da licenciatura em Engenharia Electrotécnica e Computadores da F.E.U.P. 1993.

[Pinson 63] "Pitch-Synchronous Time-Domain Estimation of Formant Frequencies and Bandwidths", Elliot. Pinson, The Journal of Acoustical Society of America, vol.35 August 1963.

[Proakis 92] "Advenced Digital Signal Processing", J. G. Proakis, C. M. Rader, F. Ling, C. L. Nikias, Macmillan Publishing Company 1992.

[Rabiner 78] "Digital Processing of Speech Signal", L. R. Rabiner, R. W. Schafer, Prentice-Hall Signal Processing Series 1978.

[Ramachandran 89] "Pitch Prediction Filters in Speech Coding", R. Ramachandran, P. Kabal, IEEE Transactions on Acoustics Speech, and Signal Processing, vol. 37, Nº 4, April 1989.

[Reddy 84] "High-Resolution Formant Extraction from Linear-Predictio Phase Spectra", N. S. Reddy, M. Swamy, IEEE Transactions on Acoustics Speech, and Signal Processing, vol. ASSP-32 Nº 6 December 1984.

[Rosenberg 71] "Effect of Glottal Pulse Shape on the Quality of Natural Vowels,", A. E. Rosenberg, J. Acoust. Soc. Am, Vol. 49,

Nº.2, pp. 583-590, February 1971

[Rowden 92] "Speech Processing", C. Rowden, McGRAW-HILL Book Company The Essex Series in Telecommunications and Information Systems 1992.

[Schafer 70] "System for Automatic Formant Analysis of Voiced Speech", R. W. Schafer, L. R. Rabiner, The Journal of Acoustic Society of America, vol. 47 Nº 2 1970.

[Schafer 75] "Digital Representations of Speech Signals", R. W. Schafer, L. R. Rabiner, Proc. IEEE vol. 63, April 1975.

[Silveira 86] "Estudos de Fonologia Portuguesa", Regina C. P. da Silveira, Cortez Editora 1986.

[Snell 93] "Formant Location From LPC Analysis Data", R. C. Snell, F. Milinazzo, IEEE Transations on Speech and Audio Processing, vol. 1 Nº 2, April 1993.

[Sum 89] "Unstable Covariance LPC Solutions from Nonstationary Speech Waveforms", Chun Sum, P. H. Milenkovic, IEEE Transactions on Acoustics Speech, and Signal Processing, vol. ASSP-37, Nº 5, May 1989.

[Verhelst 86] "A New Model for the Short-Time Complex Cepstrum of Voiced Speech", W. Verhelst, O. Steenhaut, IEEE Transactions on Acoustics Speech, and Signal Processing, vol. ASSP-34, Nº 1, February 1986.

[Viana 73] "Estudos de Fonética Portuguesa", G. Viana, Imprensa Nacional - Casa da Moeda 1973.

ANEXO A

Anexo A1

Representação interna adicional de caracteres do MULTIVOX para o português.

Grafema	Repres Inter..	Código Hex. em 850	Código Dec. em 850	Código Hex. em 860	Código Dec. em 860
á	a´	A0	160	A0	160
é	e´	82	130	82	130
í	i´	A1	161	A1	161
ó	o´	A2	162	A2	162
ú	u´	A3	163	A3	163
à	a´	85	133	85	133
è	e´	8A	138	8A	138
ì	i´	8D	141	8D	141
ò	o´	95	149	95	149
ù	u´	97	151	97	151
â	a^	83	131	83	131
ê	e^	88	136	88	136
î	o^	8C	140	-	-
ô	o^	93	147	93	147
û	u´	96	150	-	-
ã	a#	C6	198	84	132
õ	o#	E4	228	94	148
Ã	a#	C7	199	8E	142
Õ	o#	E5	229	99	153
Á	a´	B5	181	86	134
É	e´	90	144	90	144
Í	i´	D6	214	8B	139
Ó	o´	E0	224	9F	159
Ú	u´	E9	233	96	150
À	a´	B7	183	91	145
È	e´	D4	212	92	146
Ê	eˆ	D2	210	89	137
Â	aˆ	B6	182	8F	143
Ô	oˆ	E2	226	8C	140
ô	oˆ	93	147	93	147
ç	c'	87	135	87	135
Ç	c'	80	128	80	128

Anexo A2

Regras de correcção da transcrição fonética implementadas no módulo "Preparação prosódica ao nível de códigos fonéticos" no conversor texto-fala MULTIVOX para o português.

As chavetas {} indicam a ocorrência de qualquer um dos códigos do seu interior. Vog indica a ocorrência de uma vogal. Cons indica a ocorrência de uma consoante.

As regras são apresentadas pela ordem de execução dos seus algoritmos.

1ª - Se acontece a sequência de códigos fonéticos 4, {2, 8, 10}, 34 então troca 4 por 5.

2ª - Se acontece a sequência de códigos fonéticos 4, 32, Cons então troca 4 por 3.

3ª - Se acontece a sequência de códigos fonéticos 9, 32, Cons então troca 9 por 2.

4ª - Se acontece a sequência de códigos fonéticos Vog, 27, Vog então troca 27 por 26.

5ª - Se acontece a sequência de códigos fonéticos 27, 27 então fica apenas 27.

6ª - Se acontece a sequência de códigos fonéticos {6, 8, 10}, {19, 20}, Cons então ficará 40 Cons.

7ª - Se acontece a sequência de códigos fonéticos{2, 9}, {19, 20}, Cons então ficará 36 Cons.

8ª - Se acontece a sequência de códigos fonéticos{7, 22}, {19, 20}, Cons então ficará 38 Cons.

9ª - Se acontece a sequência de códigos fonéticos{3, 4}, {19, 20}, Cons então ficará 37 Cons.

10ª - Se acontece a sequência de códigos fonéticos{5, 29}, {19, 20}, Cons então ficará 39 Cons.

11ª - Se acontece a sequência de códigos fonéticos {não pausa}, 27, Cons então troca 27 por 30.

12ª - Se acontece a sequência de códigos fonéticos {de 36 a 40}, 34, {não pausa} então troca 34 por 33.

13ª - Se acontece a sequência de códigos fonéticos 30, 1, Vog então ficará 26, 1, 88, Vog.

14ª - Procura início e fim de palavra.

Se a palavra contém mais de dois códigos então

Procura silaba tónica.

Troca todos os códigos 4 à esquerda da silaba tónica por códigos 5.

15ª - Procura início e fim da palavra.

Se a palavra termina com a sequência de códigos 6, 14, 40, 19 ("mente"), então

Se o resto da palavra não contém nenhum dos códigos {2, 3, 10, 17} então

Procura a penúltima sílaba

Se a vogal dessa sílaba for {9, 4, 6} então troca respectivamente por {2, 3, 10}.

16ª - Se acontece a sequência de códigos fonéticos {6, 8, 10}, 16, 27, Vog então troca pela sequência 8, 22, 30, Vog.

17ª - Procura início e fim de palavra.

Se a palavra ainda não tem a marca da sílaba tónica e não contém nenhum dos códigos {2, 3, 10, 17} então

Procura penúltima sílaba

Se a vogal dessa sílaba for {9, 4, 6} então troca respectivamente por {2, 3, 10}.

Anexo A3

Tabela de regras de conversão grafema-fonema implementadas no módulo "Regras de conversão grafema-código de fonema na forma tabular".

Por não ser visível o "espaço" existente no início das primeiras regras, foi substituido, apenas o espaço no início das regras, pelo caracter "□".

□emerge^ncia \=1 7 19 6 34 28 240 40 27 7 9 0 0 0
□fotografias \=1 25 5 14 5 15 34 9 25 207 7 9 30 0
□jornalistas \=1 28 5 34 20 9 32 207 7 30 14 9 30 0
□clarinetes \=1 16 32 9 34 7 20 208 8 14 6 30 0
□fotografia \=1 25 5 14 5 15 34 9 25 207 7 9 0
□jornalista \=1 28 5 34 20 9 32 207 7 30 14 9 0
□acarretas \=1 9 16 9 33 210 10 14 9 30 0 0
□cheiretes \=1 30 8 22 34 208 8 14 6 30 0 0
□cinquenta \=1 27 38 35 240 40 14 9 0 0 0 0
□clarinete \=1 16 32 9 34 7 20 208 8 14 6 0
□colchetes \=1 16 4 32 30 208 8 14 6 30 0 0
□decide-te \=1 13 6 27 27 207 7 13 6 1 14 6
□esquec'as \=1 22 30 16 208 8 27 27 9 30 0 0
□paralelos \=1 12 9 34 9 32 210 10 32 5 30 0
□por favor \=1 12 205 34 25 9 24 204 4 34 0 0
□recochete \=1 33 6 16 5 30 208 8 14 6 0 0
□restolhos \=1 33 6 30 14 204 4 18 5 30 0 0
□acarreta \=1 9 16 9 33 210 10 14 9 0 0
□aguentas \=1 9 15 29 240 40 14 9 30 0 0
□cheirete \=1 30 8 22 34 208 8 14 6 0 0
□colchete \=1 16 4 32 30 208 8 14 6 0 0
□decretas \=1 13 6 16 34 210 10 14 9 30 0
□decretos \=1 13 6 16 34 210 10 14 5 30 0
□empenhas \=1 40 12 208 8 21 9 30 0 0 0
□esquec'a \=1 22 30 16 208 8 27 27 9 0 0
□hardware \=1 2 34 13 29 210 10 34 0 0 0
□mulheres \=1 19 5 18 210 10 34 6 30 0 0
□multivox \=1 19 5 32 14 7 24 203 3 16 27
□paralelo \=1 12 9 34 9 32 210 10 32 5 0
□podera#o \=1 12 5 13 6 34 217 0 0 0 0
□portugues\=1 12 5 34 14 5 15 208 8 0 0
□restolho \=1 33 6 30 14 204 4 18 5 0 0
□senhoras \=1 27 6 21 204 4 34 9 30 0 0
□software \=1 27 3 25 14 29 210 10 34 0 0
□aguenta \=1 9 15 29 240 40 14 9 0 0
□alcacer \=1 2 32 16 202 2 27 27 10 34
□amanha# \=1 2 19 9 21 236 36 0 0 0
□aquelas \=1 9 16 210 10 32 9 30 88 0
□caetano \=1 16 10 7 14 209 9 20 5 0
□camelos \=1 16 9 19 208 8 32 5 30 0
□decreta \=1 13 6 16 34 210 10 14 9 0

decreto \=1 13 6 16 34 210 10 14 5 0
empenha \=1 40 12 208 8 21 9 0 0 0
encolho \=1 40 16 204 4 18 5 0 0 0
espetas \=1 22 30 12 210 10 14 9 30 0
estamos \=1 30 14 209 9 19 5 30 0 0
estas a \=1 7 30 14 202 2 30 26 9 88
os meus \=85 5 30 19 208 29 30 0 0 0
pessoas \=1 12 6 27 27 204 5 9 30 0
quantas \=49 35 236 36 14 9 30 88 0 0
quantos \=49 35 236 36 14 5 30 88 0 0
quelhas \=1 16 210 10 18 9 30 0 0 0
recolho \=1 33 6 16 204 4 18 5 0 0
retomas \=1 33 6 14 203 3 19 9 30 0
sa'bado \=1 27 27 202 2 11 9 13 5 0
senhora \=1 27 6 21 204 4 34 9 0 0
trolhas \=1 14 34 208 8 18 9 30 0 0
va' la' \=64 60 24 202 2 61 32 202 2 0
acordo \=1 9 16 204 4 34 13 5 0
aquela \=1 9 16 210 10 32 9 88 0
aquele \=1 9 16 208 8 32 6 88 0
camelo \=1 16 9 19 208 8 32 5 0
coisas \=1 16 204 4 22 26 9 30 0
coisos \=1 16 204 4 22 26 5 30 0
corras \=1 16 204 4 33 9 30 0 0
credos \=1 16 34 210 10 13 5 30 0
espeta \=1 22 30 12 210 10 14 9 0
forc'a \=1 25 204 4 34 27 27 9 0
houver \=1 4 29 24 210 10 34 0 0
mesmos \=1 19 208 8 30 19 5 30 0
moc'as \=1 19 204 4 27 27 9 30 0
morras \=1 19 204 4 33 9 30 0 0
muitos \=1 219 39 38 14 5 30 88 0
mulher \=1 19 5 18 210 10 34 0 0
noites \=1 20 204 4 22 14 6 30 0
outras \=1 204 29 14 34 9 30 0 0
outros \=1 204 29 14 34 5 30 0 0
poetas \=1 12 29 210 10 14 9 30 0
possuo \=1 12 5 27 27 205 5 0 0
poucos \=1 12 204 4 29 16 5 30 0
quanta \=49 35 236 36 14 9 88 0 0
quanto \=49 35 236 36 14 5 88 0 0
que a' \=1 16 7 202 2 0 0 0 0
que e' \=1 16 7 210 10 0 0 0 0
quelha \=1 16 210 10 18 9 0 0 0
retoma \=1 33 6 14 203 3 19 9 0
trolha \=1 14 34 208 8 18 9 0 0
velhas \=1 24 210 10 18 9 30 0 0

velhos \=1 24 210 10 18 5 30 0 0
veloso \=1 24 6 32 204 4 26 5 0
viagem \=1 24 22 202 2 28 36 38 0
abril \=1 9 11 34 207 7 32 0
adeus \=1 9 13 208 8 29 30 0
ainda \=89 9 238 38 13 9 0 0
baixo \=1 11 202 2 22 30 5 0
cepos \=1 27 208 8 12 5 30 0
ceras \=1 27 208 8 34 9 30 0
coisa \=1 16 204 4 22 26 9 0
coiso \=1 16 204 4 22 26 5 0
corra \=1 16 204 4 33 9 0 0
corvo \=1 16 204 4 34 24 5 0
cozas \=1 16 204 4 26 9 30 0
credo \=1 16 34 210 10 13 5 0
creta \=1 16 34 210 10 14 9 0
e ~~\\\=1 254 7 88 0 0 0 0
ecran \=1 10 16 34 236 36 0 0
estas \=1 210 10 30 14 9 30 88
falar \=1 25 9 32 202 2 34 0
fizer \=1 25 7 26 210 10 34 0
foges \=1 25 203 3 22 28 6 30
fomos \=1 25 204 4 19 5 30 0
foram \=1 25 204 4 34 36 0 0
foste \=1 25 204 4 30 14 6 0
jorge \=1 28 203 3 34 28 6 0
maior \=1 19 2 22 203 3 34 0
mas ~\\=1 1 254 19 209 9 30 0
mesmo \=1 19 208 8 30 19 5 0
metas \=1 19 210 10 14 9 30 0
moc'a \=1 19 204 4 27 27 9 0
morra \=1 19 204 4 33 9 0 0
muito \=1 219 39 38 14 5 88 0
nomes \=1 20 204 4 19 6 30 0
o meu \=85 5 19 208 29 0 0 0
o que \=49 5 16 206 88 0 0 0
o ~~\\\=1 5 88 0 0 0 0 0
outra \=1 204 29 14 34 9 0 0
outro \=1 204 29 14 34 5 0 0
parti \=1 12 9 34 14 207 7 0
pelas \=1 12 208 8 32 9 30 0
podem \=1 12 203 3 13 236 38 0
poeta \=1 12 29 210 10 14 9 0
por ~\\=1 1 12 5 34 88 0 0
porta \=1 12 203 3 34 14 9 0
porto \=1 12 204 4 34 14 5 0
posto \=1 12 204 4 30 14 5 0

pouco \=1 12 204 4 29 16 5 0
quais \=49 35 202 2 22 30 88 0
se eu \=1 254 27 7 208 8 5 0
setas \=1 27 210 10 14 9 30 0
somos \=1 27 204 4 19 5 30 0
start \=1 27 14 202 2 34 14 0
stock \=1 27 27 14 203 3 16 0
todos \=1 14 204 4 13 5 30 0
tolos \=1 14 208 8 32 5 30 0
tomas \=1 14 203 3 19 9 30 0
velha \=1 24 210 10 18 9 0 0
velho \=1 24 210 10 18 5 0 0
alex \=1 2 32 210 10 16 27
alho \=1 202 2 18 5 0 0
ames \=1 209 9 19 6 30 0
bois \=1 11 204 4 22 30 0
cada \=1 16 209 9 13 9 0
cepo \=1 27 208 8 12 5 0
cera \=1 27 208 8 34 9 0
como \=49 16 204 4 19 5 88
coza \=1 16 204 4 26 9 0
dual \=45 13 5 202 2 32 0
eles \=1 208 8 32 6 30 0
esta \=1 210 10 30 14 9 88
este \=1 208 8 30 14 6 88
foge \=1 25 203 3 22 28 6
hoje \=1 204 4 22 28 6 0
hora \=1 203 3 34 9 0 0
lixo \=1 32 207 7 30 5 0
mais \=1 19 202 2 22 30 88
meta \=1 19 210 10 14 9 0
na#o \=1 20 217 88 0 0 0
nome \=1 20 204 4 19 6 0
onde \=49 237 37 13 6 88 0
os ~\\=1 5 30 0 0 0 0
ou ~\\=1 1 254 204 4 29 88
para \=85 12 209 9 34 9 0
pela \=1 12 208 8 32 9 0
pior \=1 12 22 203 3 34 0
qual \=49 35 202 2 32 88 0
quem \=49 16 236 36 38 88 0
seta \=1 27 210 10 14 9 0
te^m \=1 14 40 238 38 40 0
todo \=1 14 204 4 13 5 0
tolo \=1 14 204 4 32 5 0
toma \=1 14 203 3 19 9 0
um mau=1 39 19 202 2 29 0

□a a~\=89 2 2 0 0 0
□a's \=1 2 30 88 0 0
□ame \=1 209 9 19 6 0
□bem \=1 11 236 38 0 0
□boi \=1 11 204 4 22 0
□cem \=1 27 236 38 0 0
□cim \=1 27 238 38 0 0
□das \=1 13 9 30 88 0
□dos \=1 13 5 30 88 0
□ele \=1 208 8 32 6 0
□est~\=89 30 14 0 0 0
□exist=1 7 26 7 30 14
□exp~\=1 8 22 30 12 0
□foi \=1 25 204 22 0 0
□que \=49 16 6 0 0 0
□a' \=1 202 2 88 0
□ale\=1 9 32 0 0
□alo\=1 9 32 0 0
□ao \=1 2 29 0 0
□da \=1 13 9 88 0
□de \=1 13 6 88 0
□do \=1 13 29 88 0
□e' \=1 210 10 0 0
□em \=1 36 38 0 0
□es \=1 10 30 88 0
□es~\=1 22 30 0 0
□na \=1 20 9 88 0
□no \=1 20 5 88 0
□pro'=1 12 34 203 3
□se \=1 27 6 88 0
□tenh=1 14 209 9 21
□um \=1 39 88 0 0
□venh=1 24 209 9 21
□a \=1 9 88 0
□ali=1 9 32 7
□alu=1 9 32 5
□ce^=1 27 208 8
□hou=1 4 29 0
□por=1 12 5 34
□pro=1 12 34 5
□al=1 2 32
□ce=1 27 6
□e^=1 208 8
□eu=1 8 5
□ho=1 3 0
□o^=1 204 4
□oi=1 4 22

□ou=1 4 29
□e=1 10
□o=1 3
□r=1 33
□=1
#+#=60 0 0
#-#=61 0 0
&+&=80 0 0
&-&=81 0 0
(tch)=31 0 0 0 0
(,,o)=17 0 0 0
(a-)=9 0 0 0
(an)=36 0 0 0
(ch)=30 0 0 0
(e^)=8 0 0 0
(en)=40 0 0 0
(ii)=22 0 0 0
(in)=38 0 0 0
(lh)=18 0 0 0
(nh)=21 0 0 0
(on)=37 0 0 0
(oo)=3 0 0 0
(qu)=35 0 0 0
(rr)=33 0 0 0
(un)=39 0 0 0
(uu)=29 0 0 0
(@)=6 0 0
(a)=2 0 0
(b)=11 0 0
(d)=13 0 0
(e)=10 0 0
(f)=25 0 0
(g)=15 0 0
(i)=7 0 0
(j)=28 0 0
(k)=16 0 0
(l)=32 0 0
(m)=19 0 0
(n)=20 0 0
(o)=4 0 0
(p)=12 0 0
(r)=34 0 0
(s)=27 0 0
(t)=14 0 0
(u)=5 0 0
(v)=24 0 0
(z)=26 0 0
+=52 0 0
-=45 0 0
, a ~\=1 1 254 9 9 0

,~\=1 1 254
-~\=88 1 0
:~~\\=1 1 1 1 254
;~~\\=1 1 1 1 254
@+@=64 0 0
@-@=65 0 0
alhar \=9 18 202 2 34 0 0
alhas \=202 2 18 9 30 0 0
alhos \=202 2 18 5 30 0 0
anhas \=209 9 21 9 30 0 0
anhes \=209 9 21 6 30 0 0
anhos \=209 9 21 5 30 0 0
adas \=202 2 13 9 30 0
ados \=202 2 13 5 30 0
agem \=202 2 28 36 38 0
alem \=202 2 32 36 38 0
amas \=209 9 19 9 30 0
ames \=209 9 19 6 30 0
amos \=209 9 19 5 30 0
anas \=209 9 20 9 30 0
anha \=209 9 21 9 0 0
anhe \=209 9 21 6 0 0
anho \=209 9 21 5 0 0
anos \=209 9 20 5 30 0
ada \=202 2 13 9 0
ado \=202 2 13 5 0
air \=9 207 7 34 0
ais \=202 2 22 30 0
ama \=209 9 19 9 0
ame \=209 9 19 6 0
amo \=209 9 19 5 0
ana \=209 9 20 9 0
ani~\=9 20 207 7 0
ano \=209 9 20 5 0
ar-~\=202 2 34 88 1
asa \=2 26 9 0 0
az-~\=202 2 30 88 1
ac \=202 2 16 0
ai \=202 2 22 0
al \=202 2 32 0
alha=2 18 9 0
alho=2 18 5 0
am \=17 0 0 0
ar \=202 2 34 0
at \=202 2 14 0
az \=202 2 30 0
a \=9 0 0
a#e=236 36 38
a#o=217 0 0
ai'=9 207 7

a#=236 36
a'=202 2
a^=209 0
ai=2 22
au=2 29
a=9
b=11
ce^~\=27 208 8 0 0
cei~\=27 27 10 10 22
cer \=27 27 208 8 34
cir \=27 27 207 7 34
ce~\=27 27 6 0
ci~\=27 27 7 0
c'=27 27
ch=30 0
c=16
das \=13 9 30 0 0
d=13
elhas \=208 8 18 9 30 0 0
elhos \=208 8 18 5 30 0 0
enhas \=209 9 21 9 30 0 0
enhes \=208 8 21 6 30 0 0
enhos \=208 8 21 5 30 0 0
ebos \=208 8 11 5 30 0
edos \=208 8 13 5 30 0
efos \=208 8 25 5 30 0
eles \=208 8 32 6 30 0
elha \=208 8 18 9 0 0
elho \=208 8 18 5 0 0
elos \=208 8 32 5 30 0
emas \=208 8 19 9 30 0
emos \=208 19 5 30 0 0
enas \=208 8 20 9 30 0
enha \=209 9 21 9 0 0
enhe \=208 8 21 6 0 0
enho \=208 8 21 5 0 0
enos \=208 8 20 5 30 0
epos \=208 8 12 5 30 0
esas \=208 8 26 9 30 0
etas \=208 8 14 9 30 0
etos \=208 8 14 5 30 0
evos \=208 8 24 5 30 0
exce~\=8 22 30 27 27 6
ezas \=208 8 26 9 30 0
e'm \=236 36 38 0 0
ear \=7 202 2 34 0
ebo \=208 8 11 5 0
edo \=208 8 13 5 0
efo \=208 8 25 5 0
ele \=208 8 32 6 0

```
elo \=208 8 32 5 0
ema \=208 8 19 9 0
ena \=208 8 20 9 0
eno \=208 8 20 5 0
epo \=208 8 12 5 0
er-~\=208 8 34 88 1
esa \=208 8 26 9 0
esa \=8 26 9 0 0
eta \=208 8 14 9 0
eto \=208 8 14 5 0
evo \=208 8 24 5 0
ext~\=10 22 30 14 0
ez-~\=208 8 30 88 1
eza \=208 8 26 9 0
ec \=210 10 16 0
ei \=10 10 22 0
ej~\=10 10 22 28
el \=210 10 32 0
em \=36 36 38 0
er \=208 8 34 0
et \=210 10 14 0
ex \=210 10 16 30
exp\=8 22 30 0
ex~\=8 22 30 0
ez \=208 8 30 0
e \=6 0 0
e^n=240 40 0
ech=8 22 30
exe=7 26 6
e'=210 10
e^=208 8
ei=10 22
eo=7 5
eu=8 29
e=6
foto=25 3 14 3
f=25
gem \=28 36 38 0 0
gue^=15 208 8 0
gue=15 6 0
gui=15 7 0
ge=28 6
gi=28 7
g=15
h=0
intas \=238 38 14 9 30 0 0
ie^n\=22 240 40 0 0
ir-~\=207 7 34 88 1
iz-~\=207 7 30 88 1
im \=238 38 0 0
```

im \=38 0 0 0
ir \=207 7 34 0
is \=207 7 30 0
iz \=207 7 30 0
i \=207 7 0
i'=207 7
i^=207 7
ia=7 9
ie=22 10
iu=7 29
i=7
jo~~\\=28 5 0 0 0 0
j=28
k=16
lei=32 8 22
leo=32 7 29
lh=18 0
l=32
melos \=19 210 10 32 5 30 0
mente \=19 240 40 14 6 0 0
melo \=19 210 10 32 5 0
m=19
nh=21 0
n=20
oit~ \\=204 4 22 14 0 0 0
onhas \=204 4 21 9 30 0 0
onhes \=204 4 21 6 30 0 0
onhos \=204 4 21 5 30 0 0
orros \=204 4 33 5 30 0 0
obos \=204 4 11 5 30 0
oc'o \=204 4 27 27 5 0
olhi \=5 18 207 7 0 0
omas \=204 4 19 9 30 0
onas \=204 4 20 9 30 0
onha \=204 4 21 9 0 0
onhe \=204 4 21 6 0 0
onho \=204 4 21 5 0 0
ores \=204 4 34 6 30 0
orro \=204 4 33 5 0 0
osso \=204 4 27 27 5 0
oas \=204 4 9 30 0
obo \=204 4 11 5 0
oje \=208 8 22 28 6
oma \=204 4 19 9 0
omi \=5 19 207 7 0
ona \=204 4 20 9 0
ons \=237 37 39 30 0
osi \=5 26 207 7 0
ozi \=5 26 207 7 0
oc \=203 3 16 0

```
ol \=203 3 32 0
om \=237 37 39 0
or \=204 4 34 0
os \=5 30 0 0
ot \=203 3 14 0
ou \=4 29 0 0
ox \=203 3 16 30
oz \=203 3 30 0
o \=5 0 0
o#e=237 38 0
o#=37 0
o'=203 3
o^=204 4
oe=5 8
oi=4 22
ou=204 29
o=4
p=12
que o \=16 7 5 0 0 0 0
quel \=16 210 10 32 0 0
quem \=16 36 36 38 0 0
quer \=16 10 34 0 0 0
quais=35 202 22 30 0
quela=16 10 32 9 0
quele=16 8 32 6 0
qual=35 2 32 0
que^=16 208 8 0
quei=16 210 10 22
que=16 6 0
qui=16 7 0
qu=35 0
rr=33 0
r=34
s-~\=30 88 1 0
s \=30 0 0
sp=30 12
st=30 14
s=27
t=14
ual \=5 202 2 32 0
uns \=239 39 30 0 0
um \=39 0 0 0
uz \=205 5 30 0
u'=205 5
u^=205 5
ua=5 9
ui=5 22
u=5
v=24
w=5
```

xei=30 8 22
x~\=16 27 0
y~\=207 7 0
z=26

ANEXO B

Anexo B1

Código da função fmedia() em Matlab:

```
function media=fmedia(sinal1,janela,espacamento)

% Função que determina a média deslizante do sinal com comprimento da
%janela e espaçamento como parâmetros.

n=length(sinal1);
media=zeros(fix(n/espacamento),1);
for i=janela/2:espacamento:n-janela/2,
  media(i/espacamento+1)=sum(abs(sinal1(i-
janela/2+1:i+janela/2).*hanning(janela)))/janela;
end;
for j=1:round(janela/2/espacamento),
  media(j)=sum(abs(sinal1(1:j+janela/2).*hanning(j+janela/2)))/(j+janela/2);
end;
v=(i/espacamento+2);
f=(fix(n/espacamento));
for j=v:f,
  media(j)=sum(abs(sinal1((j-1)*espacamento-janela/2+1:n).*hanning(n-((j-
1)*espacamento-janela/2))))/(n+1-((j-1)*espacamento-janela/2+1));
end;
end;
```

Anexo B2

Código das funções fenergia() e fenerg2() em Matlab:

```
function energia=fenergia(sinal1,comp_janela,espacamento)

% Função que determina a energia deslizante do sinal com janela de
%pesagem. A função tem como parâmetros de entrada o comprimento da
%janela e o espaçamento.
% Sendo energia o sinal de saída, sinal1 o sinal de entrada, comp_janela o
%comprimento da janela e espaçamento o espaçamento entre amostras em
%que é determinada a energia.
% O sinal energia terá o nº de elementos do sinal de entrada/espaçamento.

N=comp_janela
n=length(sinal1);
energia=zeros(length(sinal1)/espacamento,1);
han=hanning(N);
for i=N/2:espacamento:n-N/2,
  sinal2=sinal1(i-N/2+1:i+N/2).*han; % *** função de pesagem ***
energia((i)/espacamento+1)=sum((sinal2).^2)/N;
end;
end;

**************************************************************

function energia=fenerg2(sinal,espacamento)

% Função que determina a energia de 1 amostra do sinal espaçadas de
%espacamento.

i=1:espacamento:length(sinal);
energia=sinal(i).^2;
end;
```

Anexo B3

Código das funções fderivad() e fzero2() em Matlab:

```
function der=fderivad(sinal)

% Função que determina a derivada de um sinal através de delta Y
%considerando delta x constante. der=fderivad(sinal).

fim=length(sinal)-1;
der=zeros(fim,1);
i=1:fim;
der(i)=sinal(i+1)-sinal(i);
end;

**************************************************************************************
***


function M = fzeros2(sinal1,N,espacamento)

% Função M=fzeros(sinal1,N,espacamento) que determina a taxa de
%passagem por zero do sinal com janela (rectangular) de comprimento N e
%espaçamento dos elementos de saída a escolher, espacamento.

n=length(sinal1);
sinal2=zeros(N,1);
M=zeros(n/espacamento,1);
fim=(n-N)/espacamento;
j=1:N-1;
for i=1:fim,
  sinal2=sinal1(i*espacamento+1:i*espacamento+N);
  taxa=(abs(sign(sinal2(j+1))-sign(sinal2(j))));
  M(i)=sum(taxa);
end;
end;
```

Anexo B4

Código do programa classif1.m em Matlab:

```
% Script que classifica o sinal em semi-segmentos de 100 amostras em
%vocalizados, não vocalizado, misto, zona comum ( não vocalizado ou
%misto), silêncio e não definido, baseado na taxa de passagem por zero,
%energia e dimensão de decisão.

[filename,sinal,Fs]=lesinal;
if Fs>20000,
  R=2;
  sinal=decimate(sinal,R);            % Decimação do sinal
  Fs=Fs/R;
end;
espacamento=10;
sinal=detrend(sinal);
energia=fenerg2(sinal,espacamento); % Energia do sinal
energia=fmedia(energia,15,1);       % Alisamento
LSES=max(energia(1:100));
der=fderivad(sinal);
zero1=fzeros2(der,50,espacamento);      % Taxa de passagem por zero da
                                        %derivada
zero=fmedia(zero1,10,1);    % Alisamento
LIZ=min(zero(1:100)); % Limite Inferior da Taxa de Passagem por Zero

car=zeros(length(sinal)/(espacamento*10),2);
s=zeros(length(sinal)/(espacamento*10),1);
decisao=zeros(length(sinal)/(espacamento*10),1);
% decisao= 0->nao definido; 1-> silencio; 2-> não vocalizado; 2,5-> zona
%comum; 3-> excitação mista; 4-> vocalizado;
for i=0:10:length(zero)-10,
  z=0;e=0;
  for j=1:10,
    if zero(i+j)>LIZ, z=z+1; end;
    if energia(i+j)<LSES, e=e+1; end;
  end;
  car(i/10+1,1)=e;   % Energia
  car(i/10+1,2)=z;   % Taxa de Passsagem por zero
end;
c=car';
```

```
s=sum(c);
s=s';
for i=1:length(decisao);
  if (s(i)>=12 & (car(i,1)>8 | car(i,2)>8)), decisao(i)=1; end;
  if (car(i,1)<=8 & car(i,2)<=3), decisao(i)=4; end;
  if (car(i,1)<3 & car(i,2)>3), decisao(i)=2; end;
  if (car(i,1)<=6 & car(i,1)>=3 & car(i,2)>3 & car(i,2)<=8), decisao(i)=3; end;
  if (car(i,1)<3 & car(i,2)>3 & car(i,2)<=8), decisao(i)=2.5; end;
end;
end;
```

Anexo B5

Código da função flifter() em Matlab:

```
function l=flifter(jan_anal,Fs,tau1,deltatau);

% Função que fornece o vector que multiplicará o cepstro para lifteragem
%das qufrências realizaando o alisamento espectral em função de tau1,
%deltatau, jan_anal e Fs.

for i=1:round(Fs*tau1),
  l(i)=1;
end;
for i=round(Fs*tau1)+1:round((tau1+deltatau)*Fs),
  l(i)=0.5*(1+cos(pi*(i/Fs-tau1)/deltatau));
end;
for i=round((tau1+deltatau)*Fs)+1:jan_anal/2,
  l(i)=0;
end;
l=[l';rot90(rot90(l))'];
end;
```

Anexo B6

Código da função espectro() em Matlab:

```
function espectro(vecSPE,Fs,eixtemp,eixfreq)

% Função espectrógrafo. Desenha o espectrograma da sequência de
%espectros alisados do parâmetro de entrada vecSPE. Fs é a frequência de
%amostragem, eixtemp e eixfreq os vectores para os eixos do tempo e das
%frequências respectivamente.

mini=min(min(vecSPE));
maxi=max(max(vecSPE));
matriz=(vecSPE-mini)*240/(maxi-mini);
fig=figure;
set(gcf,'Color',[1 1 1]);
colormap(gray(240));
colorm=colormap;
cor=1.-colorm;
colormap(cor);
image(eixtemp,eixfreq,rot90(matriz));
set(gca,'Clipping','off','XColor',[0 0 0],'YColor',[0 0 0]);
end;
```

Anexo B7

Código da função fcorprd2() em Matlab:

```
function [F0,FORM,BAN,SPE] = fcorprd2(x,Fs,jan_freq)
% Função fcorprd2() extrai automaticamente os formantes, larguras de banda
%e determina a função de transferência do trato vocal de um segmento de
%fala x com recurso à predição linear pelo método da autocorrelação. Fs é a
%frequência de amostragem e jan_freq a gama de frequências que se
%pretende analisar. O número de pólos p do modelo é alterável. As variáveis
%de saída são Fo - frequência fundamental (não determinada nesta função),
%FORM - vector com os 4 formantes, BAN - vector com as 4 larguras de
%banda, SPE - função de transferência do trato vocal.

p=13;                   % Número de polos
clear ACORR;
[m,l]=size(x);
if l==1, x=x';l=m;end;
x=x.*hanning(l)';       % Janela de Hanning
R=xcorr(x);
R=R(l:l+p);
i=1:p;
for j=1:p,
  ACORR(i,j)=R(abs(i-j)+1);
end;
R=R(2:p+1);
a=inv(ACORR)*R;
for i=1:l,
  k=1:min([(i-1) p]);
  e(i)=x(i)-sum(a(k)'.*x(i-k));% Erro residual
end;
%plot(e);                       % Representação do erro residual
a=[1;-a];
raiz=roots(a);
r=[abs(raiz) angle(raiz)];
num=0;
for i=1:p-1,
  if (r(i,1)==r(i+1,1) & r(i,2)==-r(i+1,2)),              % pólo conjugado
    num=num+1;
    teta=abs(r(i,2));
    r0=r(i,1);
```

```
   FOR(num)=Fs*teta/(2*pi);                    % Formantes  e LB Obtidos
   BAN(num)=-Fs/pi*log(r0);                    % Tradicionalmente
  end;
end;
for i=num+1:4,
  FOR(i)=0;
  BAN(i)=0;
end;

%G=G+sum(a(2:p+1).*R(1:p));     % Ganho da Função de Transferência
h=freqz(1,a,jan_freq/2);
eq=equaliz1(jan_freq,10,Fs);              % "Preemphasys" no Espectro
SPE=20*log10(abs(h))+eq;
SPE=SPE(1:fix(4500*jan_freq/Fs));
es=fft(x,jan_freq);
es=es(1:jan_freq/2)+eq';
mage=20*log10(abs(es(1:4500*jan_freq/Fs)));
media=mean(mage);
%SPE=SPE+media;

F0=100;
                              % Selecção dos 4 formantes reais
picos=fpicos(SPE);
 [FORM,i]=sort(FOR);BAN=BAN(i);              % Ordena os Formantes
for i=5:num,
  gama=80;
  if FORM(length(FORM))>=4200,
    FORM=FORM(1:length(FORM)-1);
    BAN=BAN(1:length(BAN)-1);
  else
    factorFI=jan_freq/Fs;      % Factor de transformação de frequências em
                                 % índices.
    j=1;testa=1;
    while (testa & j<(5+num-i)),
      minimo=max([(FORM(j)-gama)*factorFI 1]);
      maximo=min([(FORM(j)+gama) 4500]);
      maxim=max(picos(minimo:fix(maximo*factorFI)));
      if (maxim==0),
        testa=0;
        FORM=[FORM(1:j-1) FORM(j+1:length(FORM))];
        BAN=[BAN(1:j-1) BAN(j+1:length(BAN))];
      else
```

```
      j=j+1;
    end;
  end;
  if (testa),
    for j=1:(5+num-i),
      minimo=max([(FORM(j)-gama)*factorFI 1]);
      maximo=min([(FORM(j)+gama) 4500]);
      maxj(j)=max(picos(minimo:fix(maximo*factorFI)));
    end;
    mesmo=0;
    for j=2:(5+num-i),
      if maxj(j)==maxj(j-1), mesmo=j; end;
    end;
    if mesmo,
      FORM=[FORM(1:mesmo-2) (FORM(mesmo-1)+FORM(mesmo))/2
FORM(mesmo+1:length(FORM))];
      BAN=[BAN(1:mesmo-2) (BAN(mesmo-1)+BAN(mesmo))/2
BAN(mesmo+1:length(BAN))];
    else
      [mini,indmini]=min(maxj);
      FORM=[FORM(1:indmini-1) FORM(indmini+1:length(FORM))];
      BAN=[BAN(1:indmini-1) BAN(indmini+1:length(BAN))];
    end;
  end;
 end;
end;
end;
```

Anexo B8

Código da função fcovpr2() em Matlab:

```
function [F0,FORM,BAN,SPE] = fcovpr2(x,Fs,jan_freq)
% Função fcovpr2() para extração automática dos formantes, larguras de
%banda e determinação da função de transferência do trato vocal de um
%segmento de fala x com recurso à predição linear pelo método da matriz
%covariância. Fs é a frequência de amostragem e jan_freq a gama de
%frequência que se pretende analisar. O número de pólos p do modelo é
%alterável. As variáveis de saída são Fo - frequência fundamental (não
%determinada nesta função), FORM - vector com os 4 formantes, BAN -
%vector com as 4 larguras de banda, SPE - função de transferência do trato
%vocal.

p=13;                   % Número de pólos
clear COV;
[m,l]=size(x);
if l==1, x=x';l=m;end;
AR=1e-2;
BR=[1 -1];
x=filter(AR,BR,x);    % "Preemphsys"
COV=zeros(p,p);
F=zeros(1,p);
x=[x zeros(1,p)];
for i=1:p,
  F(i)=sum(x(p+1:l).*x(p+1-i:l-i));
end;
for j=1:p,
  for i=1:p,
    COV(j,i)=sum(x(p+1-j:l-j).*x(p+1-i:l-i));
  end;
end;
a=(inv(COV))*F';
for i=1:l,
  k=1:min([(i-1) p]);
  e(i)=x(i)-sum(a(k)'.*x(i-k));
end;
%plot(e);pause;                    % Apresentação do erro residual
a=[1;-a];
raiz=roots(a);
r=[abs(raiz) angle(raiz)];
```

```
num=0;
for i=1:p-1,
  if (r(i,1)==r(i+1,1) & r(i,2)==-r(i+1,2)),          % pólo conjugado
    num=num+1;
    teta=abs(r(i,2));
    r0=r(i,1);
    FOR(num)=Fs*teta/(2*pi);                          % Formantes  e LB Obtidos
    BAN(num)=-Fs/pi*log(r0);                          % Tradicionalmente
  end;
end;
for i=num+1:4,
  FOR(i)=0;
  BAN(i)=0;
end;
%G=COV0(1)+sum(a(2:p+1).*COV0(1:p)');      % Ganho da Funç. de Transf.
h=freqz(1,a,jan_freq/2);
eq=equaliz1(jan_freq,10,Fs);
SPE=20*log10(abs(h))+eq;
SPE=SPE(1:fix(4500*jan_freq/Fs));
F0=100;
picos=fpicos(SPE);
[FORM,i]=sort(FOR);BAN=BAN(i);                        % Ordena os Formantes
for i=5:num,
  gama=80;
  clear maxj;
  if FORM(length(FORM))>=4100,
    FORM=FORM(1:length(FORM)-1);
    BAN=BAN(1:length(BAN)-1);
  else
    factorFI=jan_freq/Fs;      % Factor de transformação de frequências em
                               %índices.
    j=1;testa=1;
    while (testa & j<=(5+num-i)),
      minimo=max([(FORM(j)-gama)*factorFI 1]);
      maximo=min([(FORM(j)+gama) 4500]);
      maxim=max(picos(minimo:fix(maximo*factorFI)));
      if (maxim==0),
        testa=0;
        FORM=[FORM(1:j-1) FORM(j+1:length(FORM))];
        BAN=[BAN(1:j-1) BAN(j+1:length(BAN))];
      else
        j=j+1;
```

```
    end;
  end;
  if (testa),
    for j=1:(5+num-i),
      minimo=max([(FORM(j)-gama)*factorFI 1]);
      maximo=min([(FORM(j)+gama) 4500]);
      maxj(j)=max(picos(minimo:fix(maximo*factorFI)));
    end;
    mesmo=0;
    for j=2:(5+num-i),
      if maxj(j)==maxj(j-1), mesmo=j; end;
    end;
    if mesmo,
      FORM=[FORM(1:mesmo-2) (FORM(mesmo-1)+FORM(mesmo))/2
FORM(mesmo+1:length(FORM))];
      BAN=[BAN(1:mesmo-2) (BAN(mesmo-1)+BAN(mesmo))/2
BAN(mesmo+1:length(BAN))];
    else
      [mini,indmini]=min(maxj);
      FORM=[FORM(1:indmini-1) FORM(indmini+1:length(FORM))];
      BAN=[BAN(1:indmini-1) BAN(indmini+1:length(BAN))];
    end;
  end;
 end;
end;
end;
```

Anexo B9

Código do programa sintese.m em Matlab:

```
% Programa que realiza a síntese de fala tendo como fonte uma matriz
%vecFOR com os formantes F1,F2,F3 e F4, a matriz vecBAN com as
%larguras de banda B1, B2, B3, B4, o ganho vecAMP e o Pitch vecF0 e a
%duração janela. É suposto estes vectores existirem já no sistema matlab. O
%sinal de saída pode ser armazenado num ficheiro wave.

Fs=11025;
DF=1/Fs*janela;                    % Duração de uma frame.
T=1/Fs;
e=zeros(1,janela);
aant=400*pi;
bant=5000*pi;
Tant=1/10000;
tau1=1e-3;    % 1 ms.
deltatau=2e-3;        % 2 ms.
saida=zeros(1,(jmax+1)*janela/2);
indmax=round(4500*janela/Fs);
vecSPE=zeros(jmax,indmax);
H=hanning(janela);
ind=1;
sobrep=50;
resto=[];
for j=1:jmax,
  [sinal,resto]=fgerimp(Fs,F0(j),janela,resto);
  sinal=sinal*vecAMP(j);
  for i=1:4,
    F=vecFOR(j,i);
    B=vecBAN(j,i);
    BB=1-2*exp(-2*pi*B*T)*cos(2*pi*F*T)+exp(-2*pi*B*T); %
    A=[1 -2*exp(-2*pi*B*T)*cos(2*pi*F*T) exp(-2*pi*B*T)];
    sinal=filter(BB,A,sinal);          % Filtro que modela os formantes
  end;
  AR=1e-2;
  BR=[1 -1];
  sinal1=filter(BR,AR,sinal);
  s=sinal1.*H';
  saida(ind:ind+janela-1)=saida(ind:ind+janela-1)+s;
```

```
  ind=ind+janela-sobrep/100*janela;
end;
espac=janela/2;inicio=1;fim=length(saida);i=0;  %Determinação do espectro
                                                %alisado
for k=inicio+janela/2:espac:fim-janela/2,
  i=i+1;
  sinal2=saida(k-janela/2:k-1+janela/2);
  c=real(ifft(20*log10(abs(fft(sinal2))/1)));
  l=flifter(janela,Fs,tau1,deltatau);  % l(nT)
  cl=c.*l';
  X=real(fft(cl));
  X=X(1:indmax);
  vecSPE(i,:)=X;
end;
eixtemp=(1:jmax)*DF*1000;
eixfreq=[4.5 0];
espectro(vecSPE,Fs,eixtemp,eixfreq);

salvar=input('Deseja salvar o sinal de saída ? ','s');
if ((salvar=='s') | (salvar=='S')),
  filename=input('Nome do sinal a guardar ? ','s');
  savewave(filename,saida,Fs);
end;
```

Anexo B10

Código da função fgerimp() em Matlab:

```
function [sinal,restos]=fgerimp(Fs,f0,janela,resto)

% Função geradora de impulsos glotais à frequência fundamental f0 num
%segmento de comprimento janela a uma frequência de amostragem Fs. O
%sinal resto e restos permitem que a função comece a gerar o primeiro
%impulso do segmento no ponto onde terminou o último impulso do
%segmento anterior anterior.
salto =round(Fs/f0);
a=0.90;
B=[0 -a*exp(1)*log(a)];
A=[1 -2*a a^2];
comp_res=length(resto);
sinal=zeros(1,janela-comp_res);
for i=1:salto:(janela-comp_res); sinal(i)=1; end;
sinal=[sinal,zeros(1,i+salto-(janela-comp_res))];
sinal=filter(B,A,sinal);
restos=sinal(janela-comp_res:length(sinal));
sinal=[resto,sinal(1:janela-comp_res)];
end;
```

Anexo B11

Código da função fpitch3() em Matlab:

```
function pitch=fpitch3(sinal1,pitch_ant,comp_janela)
% Função que determina o período fundamental baseada em processamento %paralelo no domínio temporal.
% Alisando o sinal com a média. Determina 6 estimativas para o pitch.
% pitch será o vector de comprimento 6 que contém as seis estimativas para %o pitch com esta função.
% sinal1 é o vector que contém a porção de sinal que se pretende analisar.
% pitch_ant será o valor estimado da frequencia do pitch do segmento %anterior.
% comp_janela será um inteiro que indica o comprimento da janela com que %é feita a pesagem com a média (segundo cálculos efectuados %anteriormenta este valor deverá ser entre 50 e 80).

Fs=11025;      % Frequência de amostragem
N=comp_janela;
n=length(sinal1);
M=[];
espacamento=1;   %              ***** Espaçamento *****
                                %      ******Cálculo da Média ******
for i=N/2:espacamento:n-N/2,
  sinal2=sinal1(i-N/2+1:i+N/2).*hanning(N);
  M=[M;sum(abs(sinal2))/N];
end;
M=detrend(M);
T=espacamento/Fs;     % intervalo de tempo (seg) entre elementos.
n=length(M);
m=zeros(n,6);
                      %      ****** Determinação de m1 e m4 ******
k=1;
while (k<n-1),
  while ((M(k+1) < M(k)) & (k<n-1)), k=k+1; end;
  if (k<n-1), m(k,4)=abs(M(k)); end;
  while ((M(k+1) >= M(k)) & (k<n-1)), k=k+1; end;
  if (k<n-1), m(k,1)=abs(M(k)); end;
end;
                            %      *****  Determinação de m2 e m3  *****
anterior=1;
while (m(anterior,1)==0), anterior=anterior+1; end;
```

```
for k=anterior+1:n,
  if (m(k,1)~=0),
    vale=min(M(anterior:k));
    m(k,2)=M(k)+abs(vale);
    if (m(k,1)<m(anterior,1)), m(k,3)=0; else m(k,3)=m(k,1)-m(anterior,1); end;
    anterior=k;
  end;
end;
                                    %    *****  Determinação de m5 e m6   *****
anterior=1;
while (m(anterior,4)==0), anterior=anterior+1; end;
for k=anterior+1:n,
  if (m(k,4)~=0),
    pico=max(M(anterior:k));
    m(k,5)=abs(M(k))+abs(pico);
    if (m(k,4)<m(anterior,4)), m(k,6)=0; else m(k,6)=m(k,4)-m(anterior,4); end;
    anterior=k;
  end;
end;
                                %     ***** Cálculo do pitch usando m3 e m6 *****
saltomin=round(1/(pitch_ant*1.15)/T);
saltomax=round(1/(pitch_ant*0.85)/T);
for j=3:3:7,
  anterior=1;
  som=0;
  conta=0;
  while ((m(anterior,j)==0) & (anterior<n)), anterior=anterior+1; end;
  k=anterior+saltomin;
  while (k<=n),
    while ((m(k,j)==0) & (k<n)), k=k+1; end;
    if ((k<=anterior+saltomax) & (m(k,j)~=0)),
      conta=conta+1;
      som=som+(k-anterior);
    end;
    anterior=k;
    k=anterior+saltomin;
  end;
  if (som ~= 0), pitch(j)=conta/(som*T); else pitch(j)=NaN; end;
end;
                        %     ***** Cálculo do pitch usando m1 m2 m4 e m5 *****
saida=zeros(n,6);
for l=0:3:3,
```

```
  for j=1:2,
    ppi=0;
    inicio=1;
    fim=1;
    salto=round(1/(pitch_ant*T*3));
    const=log(3*salto);   %   valor de tau
    while (m(inicio,j+l)==0), inicio=inicio+1; end;
    k=inicio+salto;
    marca=inicio;
    for i=marca:k, saida(i,j+l)=m(marca,j+l); end;
    while (k<n),
      if (m(k,j+l)>max(0,(m(marca,j+l)-log(k-marca-
salto+1)/const*m(marca,j+l)))),
        ppi=ppi+1;
        fim=k;
        marca=k;
        for i=k:min(n,k+salto), saida(i,j+l)=m(marca,j+l); end;
        k=k+salto;
      else
        saida(k,j+l)=m(marca,j+l)-log(k-marca-salto+1)/const*m(marca,j+l);
      end;
      k=k+1;
    end;
    pitch(j+l)=1/((fim-inicio)*T)*(ppi);
  end;
end;
end;
```

Anexo B12

Código do programa falacont.m em Matlab:

```
% Sript que faz a análise de fala contínua com recurso às funções
% fformvo2, fcorprd2() e fcovpr2().
jan_anal=256;                          % Janela de Análise
jan_freq=max([512 jan_anal]);
[filename,sinal,Fs]=lesinal;
if Fs>20000,
  R=2;
  sinal=decimate(sinal,R);          % Decimação do sinal
  Fs=Fs/R;
end;
figure;plot(sinal);
inicio=input('Escolha início da análise ');
fim=input('Escolha fim da análise ');
metodo='';
while (metodo~='lc' & metodo~='c' & metodo~='la'),
  metodo=input('Escolha método a usar: lc-LPC(covariancia); la-
LPC(autocorrelação); c-CEPSTRO ? ','s');
end;
espac=jan_anal/2;                  % Espaçamento entre janelas
durframe=espac/Fs;                 % Duração da frame (segmento)
Freqmax=4500;                          % Máxima Frequência de análise
factorFI=jan_freq/Fs;                  % Factor de transformação de
                                       %frequências em índices.
indmax=fix(Freqmax*factorFI);
com_vec=fix((fim-inicio-jan_anal)/espac+1);
vecF0=zeros(com_vec,1);
vecAMP=zeros(com_vec,1);
vecFOR=zeros(com_vec,4);
vecBAN=zeros(com_vec,4);
vecSPE=zeros(com_vec,indmax);
eixtemp=zeros(1,round(com_vec));
i=0;
for k=inicio+jan_anal/2:espac:fim-jan_anal/2,
  i=i+1;
  sinal2=sinal(k-jan_anal/2:k-1+jan_anal/2);
  eixtemp(i)=( k-1+jan_anal/2)*1000/Fs;
  if metodo=='lc', [F0,FORM,BAN,SPE] = fcovpr2(sinal2,Fs,jan_freq);
```

```
   elseif metodo=='la', [F0,FORM,BAN,SPE] = fcorprd2(sinal2,Fs,jan_freq);
   elseif metodo=='c', [F0,FORM,BAN,SPE] = fformvo1(sinal2,Fs,jan_freq);
end;
 vecAMP(i)=sum(sinal2.^2)/jan_anal;
 vecF0(i)=F0;
 vecFOR(i,1:4)=FORM;
 vecBAN(i,1:4)=BAN;
 vecSPE(i,1:indmax)=SPE';
end;

deltaT=1e-3;
deltaF1=100;
deltaF2=150;
deltaF3=200;
deltaF4=250;
deltaB1=100;
deltaB2=150;
deltaB3=200;
deltaB4=250;
for I=1:1,
 vecF0=fcorrec(1./vecF0,deltaT);
 vecF0=1./vecF0;
 vecFOR(:,1)=fcorrec(vecFOR(:,1),deltaF1);
 vecFOR(:,2)=fcorrec(vecFOR(:,2),deltaF2);
 vecFOR(:,3)=fcorrec(vecFOR(:,3),deltaF3);
 vecFOR(:,4)=fcorrec(vecFOR(:,4),deltaF4);
 vecBAN(:,1)=fcorrec(vecBAN(:,1),deltaB1);
 vecBAN(:,2)=fcorrec(vecBAN(:,2),deltaB2);
 vecBAN(:,3)=fcorrec(vecBAN(:,3),deltaB3);
 vecBAN(:,4)=fcorrec(vecBAN(:,4),deltaB4);
end;

s=['Frequência Fundamental '];
figure();
subplot(211);plot(eixtemp,vecF0,'w');title(s);
ylabel('Hz');xlabel('mseg');
subplot(212);
plot(eixtemp,vecFOR(:,1),'w.',eixtemp,vecFOR(:,2),'w.',eixtemp,vecFOR(:,3),'
w.',eixtemp,vecFOR(:,4),'w.');
title('Formantes');
ylabel('Hz');xlabel('mseg');
figure();
```

```
plot(eixtemp,vecBAN(:,1),'wo',eixtemp,vecBAN(:,2),'w*',eixtemp,vecBAN(:,3),'
w+',eixtemp,vecBAN(:,4),'wx');
title('Larguras de Banda');
xlabel('mseg');ylabel('Hz');
eixfreq=[4.5 0];
espectro(vecSPE,Fs,eixtemp,eixfreq);
end;
```

Anexo B13

Código da função fformvo2() em Matlab:

```
function [F0,FORMANTES,BANDAS,X] = fformvoc(sinal2,Fs,jan_freq)

% Função que faz a análise do segmento de sinal sinal2 pelo método do %cepstro.
%Fs - frequência de amostragem; jan_freq - janela de frequências usada, F0-%frequência fundamental; X - espectro alisado; FORMANTES - vector com %os 4 formantes; BANDAS - vector com as 4 larguras de banda.


jan_anal=length(sinal2);            % Comprimento da janela de amostragem
F0max=330;          % Frequência do período mínimo de pitch

F2tresh1=-17.4;                     % Valores de treshould para obtenção
F2tresh2=-48.6;                     % dos formantes.
F3tresh1=-34.8;
F3tresh2=-1000;
F4tresh1=-48.6;
at=3;                       % Obtenção das larguras de banda a -3 dB.
                            % Determinação automática dos Formantes

regF1=[200 1000];
regF2=[550 2500];           % Região dos Formantes.
regF3=[1100 3000];          %
regF4=[2700 4500];          % Análise até 4500 Hz.

Freqmax=4500;                       % Máxima Frequência de análise
factorFI=jan_freq/Fs;               % Factor de transformação de frequências
                                    %em índices.
indmax=fix(Freqmax*factorFI);       % Índice da frequência máxima de análise
c=zeros(jan_freq,1);
l=zeros(jan_freq,1);
H=hamming(jan_anal);                        % Janela de Hamming
sinal2=sinal2.*H;
c=real(ifft(log10(abs(fft(sinal2,jan_freq))/jan_anal)));
%figure;plot((10:jan_anal/2)*1000/Fs,c(10:jan_anal/2),);title('c');
[a,b]=max(c(round(Fs/F0max):jan_freq/2));       % Determinação de F0
b=b+round(Fs/F0max);
```

```
F0=Fs/b;                          % Frequência fundamental

tau1=1e-3;    % 1 ms.
deltatau=3e-3;        % 2 ms.
tau2=3e-3;    % usado na expansão.
l=flifter(jan_freq,Fs,tau1,deltatau);% l(nT)
                                        % e(nT)
e=equaliz1(jan_freq,10,Fs);       % "Preemphase" (Eliminação dos efeitos de
                                  % radiação e do impulso glotal)
cl=c.*l;
X=20*real(fft(cl));
X=X(1:jan_freq/2)+e;
X=X(1:indmax);                    % Restrição do espectro alisado a Freqmax
offset=min(X);
if (offset<0), X=X-offset; else offset=0; end;
eixfreq=(1:indmax)*Freqmax/indmax;     % Eixo das frequências
                                  % Obtenção das amplitudes e
                                  % índices de todos os picos
picos=fpicos(X);
%figure;plot(picos);

                                  % Determinação de F1
[F0amp,indF0]=max(picos(1:round(900*factorFI)));
F1found=0;
F1amp=1;
picosF=picos(regF1(1)*factorFI:regF1(2)*factorFI);
while (~(F1found | F1amp==0)),
  [F1amp,F1ind]=max(picosF);
  if (F1amp>F0amp+F2tresh1 & F1amp>0),
    F1=(F1ind-1)/factorFI+regF1(1);
    picos(F1ind+round(regF1(1)*factorFI)-1)=0;
    F1found=1;
    B1=flbanda(X,round(regF1(1)*factorFI)+F1ind-1,at);      % Largura de
                                                            %Banda a -3 Db.
    B1=(2*B1-1)/factorFI;
  else,
    picosF(F1ind)=0;
  end;
end;
if (~F1found),
                                        % Expansão de F1
  deltaF=900/(jan_freq/2);
```

```
  clexpand=c.*flifter(jan_freq,Fs,tau2,deltatau);
  expand=20*real(czt(clexpand(1:jan_freq/2),jan_freq/2,exp(-
j*2*pi*deltaF/Fs),exp(j*2*pi*0/Fs)));
  picosF=fpicos(expand);
  [F1amp,F1ind]=max(picosF);
  if (F1amp>0),
    F1=deltaF*F1ind;
    B1=(2*flbanda(expand,F1ind,at)-1)*deltaF;
  else,
    F1=regF1(1);
    B1=30;                                  % Valor por defeito.
  end;
F1amp=X(F1*factorFI);
end;

                                        % Determinação de F2
F2tresh=ftresh2(F2tresh1,F2tresh2,factorFI);    % F2 treshould
if F1>regF2(1), FL=regF1(1);
else FL=regF2(1);
end;
F2found=0;
F2amp=1;
picosF=picos(round(FL*factorFI):round(regF2(2)*factorFI));
while (~(F2found | F2amp==0)),
  [F2amp,F2ind]=max(picosF);
  if ((F2amp-F1amp)>F2tresh(F2ind+round(FL*factorFI)-1) & F2amp>0),
    F2found=1;
    F2=(F2ind-1)/factorFI+FL;
    F3tresh=F3tresh1;
    picos(F2ind+round(FL*factorFI)-1)=0;
    B2=flbanda(X,round(FL*factorFI)+F2ind-1,at);        % Largura de Banda
    B2=(2*B2-1)/factorFI;
  else,
    picosF(F2ind)=0;
  end;
end;
if (~F2found),
                                                % Expansão de F2
  iniexp=max([regF2(1)-450,F1-450,200]);
  deltaF=900/(jan_freq/2);
  clexpand=c.*flifter(jan_freq,Fs,tau2,deltatau);
```

```
  expand=20*real(czt(clexpand(1:jan_freq/2),jan_freq/2,exp(-
j*2*pi*deltaF/Fs),exp(j*2*pi*iniexp/Fs)));
picosF=fpicos(expand);
  [F1amp,F1ind]=max(picosF);
  F1=deltaF*(F1ind-1)+iniexp;
  picosF(F1ind)=0;
  B1=(2*flbanda(expand,F1ind,at)-1)*deltaF;
  [F2amp,F2ind]=max(picosF);
  if (F2amp>0),
    F2=deltaF*(F2ind-1)+iniexp;
    B2=(2*flbanda(expand,F2ind,at)-1)*deltaF;
  else,
    F2=F1+200;
    B2=50;                              % Valor por defeito
  end;
  F3tresh=F3tresh2;
end;
if F1>F2,
  troca=F1;
  F1=F2;
  F2=troca;
  troca=B1;
  B1=B2;
  B2=troca;
end;
F2amp=X(F2*factorFI);
                                % Determinação de F3

if F2>regF3(1), FL=regF2(1);
else FL=regF3(1);end;
F3found=0;
F3amp=1;
picosF=picos(round(FL*factorFI):round(regF3(2)*factorFI));
while (~(F3found | F3amp==0)),
  [F3amp,F3ind]=max(picosF);
  if ((F3amp-F2amp)>F3tresh & F3amp>0),
    F3found=1;
    F3=(F3ind-1)/factorFI+FL;
    F4tresh=F4tresh1;
    picos(F3ind+round(FL*factorFI)-1)=0;
    B3=flbanda(X,round(FL*factorFI)+F3ind-1,at);        % Largura de Banda
    B3=(2*B3-1)/factorFI;
```

```
  else
    picosF(F3ind)=0;
  end;
end;
if (~F3found),
                                                % Expansão de F3
  iniexp=max([regF3(1)-450,F2-450]);
  deltaF=900/(jan_freq/2);
  clexpand=c.*flifter(jan_freq,Fs,tau2,deltatau);
  expand=20*real(czt(clexpand(1:jan_freq/2),jan_freq/2,exp(-
j*2*pi*deltaF/Fs),exp(j*2*pi*iniexp/Fs)));
picosF=fpicos(expand);
  [F2amp,F2ind]=max(picosF);
  F2=deltaF*(F2ind-1)+iniexp;
  picosF(F2ind)=0;
  B2=(2*flbanda(expand,F2ind,at)-1)*deltaF;
  [F3amp,F3ind]=max(picosF);
  if (F3amp>0),
    F3=deltaF*(F3ind-1)+iniexp;
    B3=(2*flbanda(expand,F3ind,at)-1)*deltaF;
  else,
    F3=F2+200;
    B3=60;                              % Valor por defeito
  end;
  F4tresh=F3tresh2;
end;
if F2>F3,
  troca=F2;
  F2=F3;
  F3=troca;
  troca=B2;
  B2=B3;
  B3=troca;
  if F1>F2,
    troca=F1;
    F1=F2;
    F2=troca;
    troca=B1;
    B1=B2;
    B2=troca;
  end;
end;
```

```
F3amp=X(F3*factorFI);
                                % Determinação de F4

if F3>regF4(1), FL=regF3(1);
else FL=regF4(1);end;
F4found=0;
F4amp=1;
picosF=picos(round(FL*factorFI):fix(regF4(2)*factorFI));
while (~(F4found | F4amp==0)),
  [F4amp,F4ind]=max(picosF);
  if ((F4amp-F3amp)>F4tresh & F4amp>0),
    F4found=1;
    F4=(F4ind-1)/factorFI+FL;
    B4=flbanda(X,round(FL*factorFI)+F4ind-1,at);        % Largura de Banda
    B4=(2*B4-1)/factorFI;
  else
    picosF(F4ind)=0;
  end;
end;
if (~F4found),
                                          % Expansão de F4
  iniexp=max([regF4(1)-450,F3-450]);
  deltaF=900/(jan_freq/2);
  clexpand=c.*flifter(jan_freq,Fs,tau2,deltatau);
  expand=20*real(czt(clexpand(1:jan_freq/2),jan_freq/2,exp(-
j*2*pi*deltaF/Fs),exp(j*2*pi*iniexp/Fs)));
  picosF=fpicos(expand);
  [F3amp,F3ind]=max(picosF);
  F3=deltaF*(F3ind-1)+iniexp;
  picosF(F3ind)=0;
  B3=(2*flbanda(expand,F3ind,at)-1)*deltaF;
  [F4amp,F4ind]=max(picosF);
  if (F4amp>0),
    F4=deltaF*(F4ind-1)+iniexp;
    B4=(2*flbanda(expand,F4ind,at)-1)*deltaF;
  else,
    F4=F3+200;
    B4=90;
  end;
end;
if F3>F4,
  troca=F3;
```

```
  F3=F4;
  F4=troca;
  troca=B3;
  B3=B4;
  B4=troca;
  if F2>F3,
    troca=F2;
    F2=F3;
    F3=troca;
    troca=B2;
    B2=B3;
    B3=troca;
    if F1>F2,
      troca=F1;
      F1=F2;
      F2=troca;
      troca=B1;
      B1=B2;
      B2=troca;
    end;
  end;
end;

FORMANTES=[F1 F2 F3 F4];
BANDAS=[B1 B2 B3 B4];
end;
```

Anexo B14

Código da função equiliz1() em Matlab:

```
function e=equaliz1(jan_anal,db,Fs)

%Função que implementa o equalizador G(z)R(z) segundo [Flanagan 64].
% db=-amplitude em dB do equalizador para f=0 Hz; jan_anal=comprimento
%da janela; Fs=frequência de amostragem.
e=zeros(1,jan_anal/2);
aant=400*pi;
bant=5000*pi;
Tant=1/10000;
T=1/Fs;
a=aant*Tant/T;
b=bant*Tant/T;        % Dependem do falante
N=round(jan_anal*5000/Fs);
B=(1+exp(-b*T)-exp(-a*T)-exp(-T*(a+b)))/(10^(-db/20));
A=[1 exp(-b*T)-exp(-a*T) -exp(-T*(a+b))];
[H,W] = freqz(A,B,N);
eixfreq=(1:jan_anal/2)*Fs/jan_anal;
mag=abs(H);
e=20*log10(mag);
for i=N+1:jan_anal/2, e(i)=e(N); end;
end;
```

Anexo B15

Código da função ftresh2() em Matlab:

```
function F2tresh=ftresh2(F2tresh1,F2tresh2,factorFI)

% Função que fornece a curva mínima da relação entre as amplitudes de F2
% e F1.

F2tresh=zeros(round(2400*factorFI),1);
for i=round(200*factorFI):round(500*factorFI), F2tresh(i)=F2tresh1; end;
const=F2tresh1-(F2tresh2-F2tresh1)/2;
for i=round(500*factorFI)+1:round(1500*factorFI),
  F2tresh(i)=const+(F2tresh2-F2tresh1)/1000*i/factorFI;
end;
const=F2tresh1-(F2tresh1-F2tresh2)*2400/900;
for i=round(1500*factorFI)+1:round(2700*factorFI),
  F2tresh(i)=const+(F2tresh1-F2tresh2)/900*i/factorFI;
end;
end;
```

Anexo B16

Código da função fpicos() em Matlab:

```
function picos=fpicos(X)

% Função que determina os picos de um vector. Retorna um vector picos
% com a mesma dimensão de X mas só com os picos.

k=1;
n=length(X);
picos=zeros(n,1);
while (k<n-1),
  while ((X(k+1) < X(k)) & (k<n-1)), k=k+1; end;
  while ((X(k+1) >= X(k)) & (k<(n-1))), k=k+1; end;
  if (k<n-1), picos(k,1)=X(k); end;
end;
```

Anexo B17

Código da função flbanda() em Matlab:

```
function i=flbanda(X,ind,at)

% Função que determina i, o número de índices em que o sinal X decaiu at.

imax=min([length(X)-ind ind-1]);
i=0;
niv=X(ind)-at;
while ((X(ind-i)>niv) & (X(ind+i)>niv) & (i<imax)),
 i=i+1;
end;
```

Anexo B18

Código da função fdetsinc() em Matlab:

```
function m = fdetsinc(sinal)
% Função que detecta o sincronismo do período fundamental para a análise
%síncrona com pitch. O vector m de saída contém picos nos índices
%correspondentes ao início do impulso glotal do sinal.

n=length(sinal);
N=50;                   % Comprimento (alterável) da janela da média
espacamento=1;          %    ***** ESPAÇAMENTO *****
M=zeros(1,n);
%                               ****** Média ******
j=0;
M=fmedia(sinal,N,espacamento);
M=detrend(M);
n=length(M);
m=zeros(n,1);

          %     ****** Determinação de m1  ******
k=1;
while (k<n-1),
  while ((M(k+1) < M(k)) & (k<n-1)), k=k+1; end;
  while ((M(k+1) >= M(k)) & (k<n-1)), k=k+1; end;
  if (k<n-1), m(k,1)=M(k); end;
end;
end;
```

Anexo B19

Código do programa falasinc.m em Matlab:

```
% Sript que faz a análise de fala vocalizda contínua com recurso à função
% fcovpr2(). Realiza a análise síncrona com o "pitch".

[filename,sinal,Fs]=lesinal;
if Fs>20000,
  R=2;
  sinal=decimate(sinal,R);            % Decimação do sinal
  Fs=Fs/R;
end;
figure;plot(sinal);
inicio=input('Escolha início da análise ');
fim=input('Escolha fim da análise ');
sinal=sinal(inicio:fim);
m = fdetsinc(sinal);
n=length(m);
comp=0;
for i=1:n, if m(i)>0, comp=comp+1; end; end;    %comp= número de períodos
Freqmax=4500;                                   % Máxima Frequência de análise
jan_freq=512;
factorFI=jan_freq/Fs;% Factor de transformação de frequências em índices.
indmax=fix(Freqmax*factorFI);
com_vec=comp-1;
vecF0=zeros(com_vec,1);
vecAMP=zeros(com_vec,1);
vecFOR=zeros(com_vec,4);
vecBAN=zeros(com_vec,4);
vecSPE=zeros(com_vec,indmax);
eixtemp=zeros(1,round(com_vec));
ini=1;
while m(ini)<=0, ini=ini+1; end;
for i=1:comp-1,
  fim=ini+1;
  while m(fim)<=0, fim=fim+1; end;
  sinal2=sinal(ini:fim);
  eixtemp(i)=fim*1000/Fs;
  [F0,FORM,BAN,SPE] = fcovpr2(sinal2,Fs,jan_freq);
  vecF0(i)=F0;
```

```
  vecFOR(i,1:4)=FORM;
  vecBAN(i,1:4)=BAN;
  vecSPE(i,1:indmax)=SPE';
  vecAMP(i)=sum(sinal2.^2)/jan_anal;
  ini=fim;
end;
deltaT=1e-3;
deltaF1=100;
deltaF2=150;
deltaF3=200;
deltaF4=250;
deltaB1=100;
deltaB2=150;
deltaB3=200;
deltaB4=250;
for i=1:1,
 vecF0=fcorrec(1./vecF0,deltaT);
 vecF0=1./vecF0;
 vecFOR(:,1)=fcorrec(vecFOR(:,1),deltaF1);
 vecFOR(:,2)=fcorrec(vecFOR(:,2),deltaF2);
 vecFOR(:,3)=fcorrec(vecFOR(:,3),deltaF3);
 vecFOR(:,4)=fcorrec(vecFOR(:,4),deltaF4);
 vecBAN(:,1)=fcorrec(vecBAN(:,1),deltaB1);
 vecBAN(:,2)=fcorrec(vecBAN(:,2),deltaB2);
 vecBAN(:,3)=fcorrec(vecBAN(:,3),deltaB3);
 vecBAN(:,4)=fcorrec(vecBAN(:,4),deltaB4);
end;
s=['Frequência Fundamental - ' filename];
figure;subplot(211);plot(eixtemp,vecF0,'w');title(s);
ylabel('Hz');xlabel('mseg');
subplot(212);
plot(eixtemp,vecFOR(:,1),'w.',eixtemp,vecFOR(:,2),'w.',eixtemp,vecFOR(:,3),'
w.',eixtemp,vecFOR(:,4),'w.');
title('Formantes');ylabel('Hz');xlabel('mseg');
figure;plot(eixtemp,vecBAN(:,1),'wo',eixtemp,vecBAN(:,2),'w*',eixtemp,vecBA
N(:,3),'w+',eixtemp,vecBAN(:,4),'wx');
title('Larguras de Banda');xlabel('mseg');ylabel('Hz');
eixfreq=[4.5 0];
espectro(vecSPE,Fs,eixtemp,eixfreq);
end;
```

Anexo B20

Código da função fcorrec() em Matlab:

```
function vec=fcorrec(vecin,delta)

% Função que faz a correcção de 1 e 2 pontos "outlier's".
vec=vecin;
der=fderivad(vec);
comp=length(der);
for i=1:comp-2,
  if (der(i)<-delta),                          % Possível situação A
    if (der(i+1)>delta),                       % Situação A 1 ponto
      vec(i+1)=(vec(i)+vec(i+2))/2;            % Correcção de 1 ponto
    end;
    if ((der(i+2)>delta) & (der(i+1)>-delta)),       % Situação A 2 pontos
      salto=(vec(i+3)-vec(i))/3;
      vec(i+1)=vec(i)+salto;                   % Correcção de 2 pontos
      vec(i+2)=vec(i)+2*salto;
    end;
  end;
end;
der=fderivad(vec);
for i=1:comp-2,
  if (der(i)>delta),                           % Possível situação B
    if (der(i+1)<-delta),                      % Situação B 1 ponto
      vec(i+1)=(vec(i)+vec(i+2))/2;            % Correcção de 1 ponto
    end;
    if ((der(i+2)<-delta) & (der(i+1)<delta)),       % Situação B 2 pontos
      salto=(vec(i+3)-vec(i))/3;
      vec(i+1)=vec(i)+salto;                   % Correcção de 2 pontos
      vec(i+2)=vec(i)+2*salto;
    end;
  end;
end;
i=i+1;                              % Situação de 1 ponto A ou B no final do vector.
if (((der(i)<-delta) & (der(i+1)>delta)) | ((der(i)>delta) & (der(i+1)<-delta))),
  vec(i+1)=(vec(i)+vec(i+2))/2;                % Correcção de 1 ponto
end;
end;
```

Anexo B21

Código do programa falanvoc.m em Matlab:

```
% Script que faz a análise de um segmento de sinal não vocalizado e
% obtém as frequencias do pólo e do zero.
[filename,sinal,Fs]=lesinal;
if Fs>20000,
  R=2;
  sinal=decimate(sinal,R);          % Decimação do sinal
  Fs=Fs/R;
end;
figure;plot(sinal);
inicio=input('Escolha início da análise ');
jan_anal=512;                       %Comprimento do segmento
sinal2=sinal(inicio:inicio+jan_anal-1);
plot(sinal2);
Fpmin=1000;                         % Frequência mínima para o pólo.
Freqmax=4000;                       % Máxima Frequência de análise
factorFI=jan_anal/Fs;% Factor de transformação de frequencias em indices.
indmax=round(Freqmax*factorFI);% Índice da frequencia máxima de análise
c=zeros(jan_anal,1);
l=zeros(jan_anal,1);
H=hamming(jan_anal);                % janela de Hamming
sinal2=sinal2.*H;
c=real(ifft(log10(abs(fft(sinal2))/jan_anal)));
tau1=2e-3;    % 1 ms.
deltatau=2e-3;        % 2 ms.
l=flifter(jan_anal,Fs,tau1,deltatau);         % l(nT)
cl=c.*l;
X=20*abs(fft(cl));
X=X(1:indmax);                      % Restrição do espectro alisado a Freqmax
eixfreq=(1:indmax)*Freqmax/indmax;      % Eixo das frequências
SPE=20*log10(abs(fft(sinal2))/jan_anal);
SPE=SPE(1:indmax);
                    % Obtenção das amplitudes e índices de todos os picos
picos=fpicos(X);
[a b]=max(picos(round(Fpmin*factorFI):indmax));
Fp=(b+round(Fpmin*factorFI)-1)/factorFI;
Fz=(0.0065*Fp+4.5-a+X(1))*(0.014*Fp+28);
end;
```