Faculdade de Engenharia da Universidade do Porto

FEUP

Descaracterização Perceptiva da Assinatura Vocal

Bártolo de Melo Feiteira Maia

Dissertação realizada no âmbito do
Mestrado Integrado em Engenharia Electrotécnica e de Computadores
Major Telecomunicações

Orientador: Prof. Dr. Aníbal João de Sousa Ferreira

Janeiro de 2010

Resumo

Neste trabalho foi avaliado o impacto na percepção auditiva de diferentes métodos de análise, síntese e extracção de características de sinais de voz, após a análise e posterior ressíntese dos mesmos, com alterações deliberadas e selectivas das características extraídas.

Utilizaram-se vários algoritmos baseados no método *Linear Predictive Coding* (LPC) para testar o efeito da quantização do resíduo da análise LPC e da subtituição do resíduo por um sinal aleatório. Também se realizaram testes em que apenas se alterou o resíduo das *frames* vozeadas e outros nos quais se modificou apenas o resíduo das *frames* não-vozeadas. Em experiências baseadas em análises *Mel-Frequency Cepstral Coefficients* (MFCC) e *Perceptual Linear Prediction* (PLP) com utilização de ruído branco, testou-se a influência do número de coeficientes usados na análise.

A avaliação da percepção auditiva dos vários testes focou-se na inteligibilidade dos sinais de fala na saída e na qualidade dos mesmos relativamente aos sinais originais. Complementarmente, foi feita uma avaliação mais pormenorizada das alterações verificadas em classes fonéticas específicas.

Foi testado em cinco vogais da língua portuguesa um método que, após a análise, ressintetiza o sinal de fala, utilizando sinusóides para reproduzir a estrutura harmónica detectada. Alterou-se a magnitude ou fase de determinadas sinusóides da estrutura harmónica e identificaram-se as modificações ocorridas em termos de percepção.

Os resultados obtidos permitem concluir que os fonemas plosivos e aspirados são os que mais facilmente sofrem alterações e que quando um sinal aleatório é introduzido na síntese, apenas as fricativas não-vozeadas premanecem inalteradas. A quantização do resíduo resultante da análise LPC, mesmo com poucos bits, obtém melhores resultados do que os restantes métodos.

Na avaliação das vogais a influência da alteração da fase não é perceptível. Para além disso, verificou-se que a redução da amplitude dos parciais mais baixos da estrutura harmónica origina uma atenuação considerável na intensidade das vogais. Algumas vogais têm regiões de frequências que as identificam das restantes e se a magnitude das sinusóides nessas regiões for consideravelmente alterada registam-se alterações acentuadas na sonoridade do sinal, nomeadamente quanto à identidade da vogal.

Abstract

The purpose of this dissertation is to evaluate the impact that several analysis, synthesis and feature extraction methods have on the perception of voice signals after the features involved in their synthesis are deliberately and selectively altered.

Various algorithms based on the Linear Predictive Coding (LPC) technique were used to test the perceptual effect of quantization on the LPC residue and the changes that occur when the same residue is substituted by random noise. In some of the tests only voiced frames were altered whereas in others only the unvoiced suffered modifications. The importance of the number of coefficients in Mel-Frequency *Cepstral* (MFC) and Perceptual Linear Prediction (PLP) analyses and using random noise in the synthesis was also investigated.

The evaluation of the impact on auditory perception due to the various analysis-synthesis methods studied in this dissertation was focused on the intelligibility of the synthesized signals and on their quality in comparison to the original signals. In addition more comprehensive tests were conducted concerning the perceptual changes in some phonetic classes.

A method that replaces the harmonic structure, detected during the analyses in voice signals was applied to five Portuguese vowels. The magnitude or the phase of some of the sinusoids of the harmonic structure were altered and the resulting perceptual changes were evaluated.

Results obtained during the various experiences indicate that the plosives and /h/ phonemes change more easily and when random noise is introduced in the synthesis, only unvoiced fricatives remain essentially unaltered. Even when only few bits are used in the quantization of the LPC residue, the synthesized signal has better quality than when other methods are used.

The practical effect of phase modification in vowels is imperceptible. On the other hand, magnitude reduction in low frequency sinusoids of the harmonic structure produces a noticeable drop in a vowels' sound volume. Some specific frequency regions allow to differentiate vowels from each other and when the amplitude of the sinusoids in those frequency regions is considerably modified, the auditory perception of the vowel is significantly altered, particularly concerning vowel identity.

Agradecimentos

Quero agradecer ao Nuno Machado da Silva e à Mariana Osswald por terem concordado ser avaliadores dos inúmeros ficheiros de voz que foi necessário classificar nesta dissertação, sem o seu contributo este trabalho não teria tido a objectividade pretendida. Muito obrigado aos dois.

Índice

Resumo ..... **iii**

Abstract ..... **v**

Agradecimentos ..... **vii**

Índice ..... **ix**

Lista de figuras ..... **xiii**

Lista de tabelas ..... **xv**

Abreviaturas e Símbolos ..... **xvii**

Capítulo 1 ..... **1**

Introdução 1
1.1 - Enquadramento 1
1.2 - Caracterização do problema 1
1.3 - Objectivos 2
1.4 - Estrutura 2

Capítulo 2 ..... **5**

Processo fonatório e audição 5
2.1 – Aparelho fonético 5
2.2 – Tipos de excitação dos sinais de fala 7
2.3 – Aparelho auditivo 10
2.3.1 – Anatomia do ouvido 10
2.3.2 – Funcionamento do ouvido 11
2.3.2.1 - Percepção de intensidade 11
2.3.2.2 - Efeito de máscara ou mascaramento 12

Capítulo 3 ..... **15**

Critérios utilizados na classificação fonética 15
3.1 – Tipos de fonemas 16
3.1.1 – Vogais 17
3.1.2 – Consoantes 17
3.2 – Diferenças entre os fonemas do inglês americano e do português europeu 18

3.3 - Formantes típicas dos fonemas 23
3.3.1 - Formantes das vogais 24
3.3.2 - Formantes das fricativas não-vozeadas 29
3.3.3 - Formantes das fricativas vozeadas 30
3.3.4 - Formantes das plosivas 31

Capítulo 4 ..... **35**

Métodos de extracção de características do sinal de voz 35
4.1 - Introdução à extracção de características do sinal de voz 35
4.2 - Método LPC 36
4.2.1 - Introdução teórica ao método LPC 36
4.2.2 - Introdução prática ao método LPC utilizado nas experiências 39
4.2.3 - Experiências baseadas no método LPC 40
4.2.3.1 - Quantização do resíduo 40
4.2.3.2 - Excitação por ruído branco 41
4.2.3.3 - Importância do *pitch* na percepção 41
4.3 - Método MFCC 41
4.3.1 - Introdução teórica ao método MFCC 41
4.3.2 - Introdução prática ao método MFCC utilizado nas experiências 43
4.3.3 - Experiências baseadas no método MFCC 45
4.3.3.1 - Influência do número de coeficientes na percepção 45
4.4 - Método PLP 45
4.4.1 - Introdução teórica ao método PLP 45
4.4.2 - Introdução prática ao método PLP utilizado nas experiências 46
4.4.3 - Experiências baseadas no método PLP 48
4.4.3.1 - Influência do número de coeficientes na percepção 48
4.5 - "Método das sinusóides" 48
4.5.1 - Introdução teórica ao "método das sinusóides" 48
4.5.1.1 - Estimação da frequência 51
4.5.1.2 - Estimação da fase 51
4.5.1.3 - Estimação da magnitude 52
4.5.2 - Introdução prática ao "método das sinusóides" 52
4.5.3 - Experiências baseadas no "método das sinusóides" 53
4.5.3.1 - Alteração da magnitude 53
4.5.3.2 - Alteração da fase 54

Capítulo 5 ..... **55**

Discussão de resultados 55
5.1 - Introdução à metodologia usada durante a fase de testes 55
5.1.1 - Metodologia usada na avaliação das consoantes 55
5.1.2 - Metodologia usada na avaliação das vogais 57
5.2 - Resultados obtidos 57
5.2.1 - Avaliação das consoantes 58
5.2.1.1 - Inteligibilidade 58
5.2.1.2 - Qualidade das ressínteses comparativamente aos sinais originais 63
5.2.1.3 - Anotações às alterações de fonemas específicos 67
5.2.2 - Avaliação das vogais 69
5.2.2.1 - Alteração da magnitude 69
5.2.2.2 - Alteração da fase 70
5.3 - Discussão dos resultados obtidos 70
5.3.1 - Consoantes 70
5.3.2 - Vogais 73

Capítulo 6 ..... **75**

Conclusões e trabalho futuro 75
6.1 - Conclusões 75
6.2 - Trabalho futuro 76

ANEXO A **77**

ANEXO B **79**

ANEXO C **81**

ANEXO D **83**

ANEXO E **85**

ANEXO F **87**

Referências **93**

Lista de figuras

Figura 2.1 – Secção sagital média do aparelho vocal [1]. 6

Figura 2.2 - Secções da laringe. Adaptado de [1]. 7

Figura 2.3 - Representação de um ciclo vibratório das pregas vocais [1]. 8

Figura 2.4 – A forma de onda de cima representada o sinal do fluxo aerodinâmico ao longo de um ciclo vibratório das pregas vocais e a forma de onda de baixo representa a derivada desse sinal. 9

Figura 2.5 - Representação do ouvido humano. Adaptado de [4]. 10

Figura 2.6 - Representação do interior da cóclea. Adaptado de [4] 11

Figura 2.7 – Gráfico Intensidade-Frequência sobreposto com diversas curvas de idêntica sonoridade. Adaptado de [1]. 12

Figura 2.8 – Correspondência entre as escalas de frequências Hertz e Bark (à esquerda) e largura de banda da escala Bark (à direita) [5]. 13

Figura 2.9 – Modelo da curva de mascaramento na escala Bark. Ilustra-se a utilização desta curva para calcular o limiar de mascaramento à frequência z_B, devido a um tom puro mascarante à frequência z_C [5]. 14

Figura 3.1 - Classificação de fonemas ingleses – ARPAbet [7]. 17

Figura 3.2 – Mapeamento das vogais orais em função do ponto de articulação e do grau de obstrução. Quando os símbolos aparecem em pares, o símbolo da direita representa a vogal arredondada. Adaptado de [8]. 23

Figura 3.3 – Diagrama que mostra a localização e o grau de obstrução provocado pela língua para as diferentes vogais do inglês americano. Adaptado de [7]. 24

Figura 3.4 – Representa para as 12 principais vogais do inglês americano, um esquema da localização dos articuladores, na coluna (a), um gráfico com a resposta no domínio dos tempos, na coluna (b), e um gráfico com a resposta no domínio das frequências, na coluna (c) [7]. 25

Figura 3.5 – Representa as frequências médias e a amplitude média relativa das três primeiras formantes de 10 das principais vogais do inglês americano [7]. 28

Figura 3.6 – Largura de banda em relação às frequências médias das três primeiras formantes de 10 das principais vogais do inglês americano [7]..... 29

Figura 3.7 – Representação das quatro fricativas não-vozeadas e da consoante aspirada /h/ do inglês americano, um esquema da localização dos articuladores durante a sua produção, na coluna (a), um gráfico com a resposta no domínio dos tempos, na coluna (b), e um gráfico com a resposta no domínio das frequências, na coluna (c) [7]..... 30

Figura 3.8 – Representação das quatro fricativas vozeadas do inglês americano, um esquema da localização dos articuladores durante a sua produção, na coluna (a), um gráfico com a resposta no domínio dos tempos, na coluna (b), e um gráfico com a resposta no domínio das frequências, na coluna (c) [7]. 31

Figura 3.9 – Representação das três plosivas não-vozeadas e das três plosivas vozeadas do inglês americano, um esquema da localização dos articuladores durante a sua produção, na coluna (a), um gráfico com a resposta no domínio dos tempos, na coluna (b), e um gráfico com a resposta no domínio das frequências, na coluna (c) [7]..... 32

Figura 4.1– Modelo genérico de tempo discreto da produção de fala. Segundo Rabiner and Schafer (1978) [7]..... 37

Figura 4.2– Diagrama de blocos da análise LPC..... 40

Figura 4.3– Diagrama de blocos da síntese LPC..... 40

Figura 4.4– Forma canónica para um sistema para desconvolução homomórfica [9]. 42

Figura 4.5– A escala mel. Segundo Stevens e Volkman (1940) [7]..... 43

Figura 4.6– Diagrama de blocos da análise MFCC. 44

Figura 4.7– Diagrama de blocos da síntese MFCC. 45

Figura 4.8– Diagrama de blocos da análise PLP..... 47

Figura 4.9– Diagrama de blocos da síntese PLP..... 47

Figura 4.10 – Resposta em frequência normalizada da "janela de seno"..... 49

Figura 4.11 – Resposta em frequência dos primeiros quatro canais do banco de filtros ODFT..... 50

Figura 4.12 – Relação entre as magnitudes dos canais ODFT $\ell - 1$, ℓ e $\ell + 1$ quando o sinal de entrada é uma sinusóide com frequência dada por $2\pi N\ell + \Delta\ell$. 50

Figura 4.13 – Diagrama de blocos do "método das sinusóides"..... 53

Figura 5.1 – Espectrograma das vogais /a/, /ɛ/, /i/, /ɔ/ e /u/ do sinal original do orador masculino..... 73

Lista de tabelas

Tabela 3.1 - Alfabeto fonético internacional (IPA ou AFI) e ARPAbet para o inglês americano [1]. 19

Tabela 3.2 – Exemplos de correspondências entre os "alfabetos fonéticos" API (AFI) e SAMPA e os grafemas utilizados no alfabeto português. Na primeira coluna estão representados todos os fonemas do IPA (AFI) para o português europeu, na segunda coluna os mesmos fonemas estão representados no "alfabeto fonético" SAMPA, na terceira os grafemas do português que podem representar esses fonemas e na última coluna são indicados alguns exemplos de palavras portuguesas para cada um dos fonemas [14]. 20

Tabela 3.3 – Exemplos de correspondências entre símbolos gráficos e sons na ortografia do português europeu padrão. Na primeira coluna estão representados todos os grafemas simples, na segunda coluna as suas correspondências fonéticas de acordo com o alfabeto Fonético Internacional e na terceira coluna são indicados alguns exemplos de palavras portuguesas para cada um dos fonemas. Em cada palavra o fonema que se pretende exemplificar está a escrito a negrito. 21

Tabela 3.4 – Exemplos de correspondências entre símbolos gráficos e sons na ortografia do português europeu padrão. Na primeira coluna estão representados as sequências de grafemas e grafemas compostos, na segunda coluna as suas correspondências fonéticas de acordo com o alfabeto Fonético Internacional e na terceira coluna são indicados alguns exemplos de palavras portuguesas para cada um dos fonemas. Em cada palavra o fonema que se pretende exemplificar está a escrito a negrito. 22

Tabela 5.1 - Escala absoluta para a degradação subjectiva de áudio codificado. Adaptado de [5]. 57

Tabela 5.2 - Resumo dos resultados da primeira fase da avaliação das consoantes**.** 59

Tabela 5.3 – Resumo estatístico do estudo da inteligibilidade. 62

Tabela 5.4 – Classificações da segunda fase da avaliação das consoantes. 63

Tabela 5.5 – Valores da média e variância da classificação atribuída aos diferentes métodos.. 67

Tabela 5.6 – Subconjunto das palavras utilizadas na avaliação das alterações fonéticas. 68

Tabela 5.7 – Observações das experiências com magnitude alterada para o orador feminino, tendo em consideração os parciais alterados e as modificações ocorridas nas vogais. 70

Tabela 5.8 – Média das classificações dos métodos avaliados. 71

Abreviaturas e Símbolos

Lista de abreviaturas

AFI	Alfabeto Fonético Internacional
ARPA	*United States Advanced Research Projects Agency*
DCT	*Discrete Cosine Transform*
DFT	*Discrete Fourier Transform*
FEUP	Faculdade de Engenharia da Universidade do Porto
FFT	*Fast Fourier Transform*
IDCT	*Inverse Discrete Cosine Transform*
IDFT	*Inverse Discrete Fourier Transform*
IFFT	*Inverse Fast Fourier Transform*
IODFT	*Inverse Odd Discrete Fourier Transform*
IPA	v*International Phonetic Alphabet*
LP	*Linear Prediction*
LPC	*Linear Predictive Coding*
MFC	*Mel Frequency Cepstral*
MFCC	*Mel Frequency Cepstral Coefficients*
Modelo AR	Modelo autorregressivo
ODFT	*Odd Discrete Fourier Transform*
PLP	*Perceptual Linear Prediction*
SAMPA	*Speech Assessment Methods Phonetic Alphabet*

Lista de símbolos

a_k	coeficiente LPC
cm	centímetro
cm^2	centímetro quadrado
Hz	Hertz
ms	milissegundo

dB	deciBel
dB SPL	decibel Sound Presure Level
kHz	kiloHertz
log	logaritmo
rad/s	radianos por segundo
µPa	micro Pascal
W/m^2	Watt por metro quadrado

Capítulo 1

Introdução

Neste capítulo é introduzido o tema abordado nesta dissertação e as diferentes dificuldades que apresenta. São também indicados os diferentes métodos que se vão estudar e os objectivos que se pretendem atingir aquando da conclusão da dissertação.

1.1 - Enquadramento

A comunicação oral é a forma de comunicação mais prática e mais usual entre humanos. A comunicação entre humanos à distância, por meio de equipamentos digitais e a comunicação Homem – Máquina são cada vez mais usuais. Existem já variadíssimos métodos de extracção de características de sinais de fala utilizados em sistemas de reconhecimento e também de ressíntese de fala. Existem também inúmeros métodos de compressão dessa informação, para que o seu transporte e transmissão sejam o mais eficiente possível, bem como algoritmos que avaliam a qualidade dos diferentes métodos, tendo em conta a proximidade entre as amostras do sinal de saída e as amostras do sinal de entrada.

Apesar desses algoritmos serem usados em diversas aplicações, a uma grande proximidade entre amostras de dois sinais diferentes pode não corresponder uma proximidade tão elevada quanto isso em termos da percepção auditiva dos mesmos.

Avaliações da percepção auditiva são avaliações que não podem ser realizadas por algoritmos matemáticos. Como as avaliações têm que ser feitas por pessoas, que têm características diferentes e interpretam a mesma realidade de forma distinta, os próprios resultados obtidos com essas avaliações estão sujeitos a conclusões que podem variar consoante quem estiver a fazer a análise dos dados.

Há, portanto, a necessidade de estabelecer um conjunto de regras na metodologia usada na obtenção das avaliações, bem como estabelecer critérios rigorosos que reduzam ao máximo a interpretação que cada um faz dos diferentes níveis de classificação.

1.2 - Caracterização do problema

O que se pretende realizar nesta dissertação é uma avaliação da inteligibilidade e da qualidade dos sinais de fala após a ressíntese, com o intuito de classificar os diferentes

métodos de extracção de características. A avaliação da qualidade da percepção auditiva é uma avaliação subjectiva dependente da interpretação que cada um faz das diferentes classificações a atribuir, bem como de outras características que variam de avaliador para avaliador. A idade, o conhecimento da linguagem dos sinais de voz, as características anatómicas de cada um e também outros factores como a profissão ou até a região de onde se é natural, são aspectos que tornam uma avaliação da percepção de sinais áudio uma avaliação bastante subjectiva.

Para atenuar a subjectividade inerente ao problema as classificações às diferentes ressínteses foram feitas por três pessoas e houve um cuidado especial na metodologia usada para que reduzisse ao máximo as interpretações pessoais dos avaliadores.

1.3 - Objectivos

O objectivo deste trabalho é avaliar a relevância perceptiva de parâmetros usualmente utilizados na análise de sinais de voz/áudio e concluir sobre o impacto na qualidade e identidade do sinal em resultado da alteração deliberada desses parâmetros.

O impacto na percepção auditiva é estudado de duas maneiras diferentes. A primeira visa estudar os efeitos que diferentes configurações dos métodos *Linear Predictive Coding* (LPC), *Mel-Frequency Cepstral Coefficients* (MFCC) e *Perceptual Linear Prediction* (PLP) produzem na inteligibilidade e qualidade do sinal de voz após a ressíntese e também caracterizar o tipo de modificações que são perceptíveis nalgumas classes fonéticas específicas. A segunda parte tem como objectivo identificar as alterações percebidas por ouvintes humanos em relação a um conjunto de vogais portuguesas quando a amplitude ou a fase de determinadas sinusóides que as constituem são alteradas.

No final, apesar da subjectividade inerente à percepção auditiva de cada indivíduo, pretende-se conseguir ter uma avaliação o mais objectiva e bem fundamentada possível sobre cada um dos métodos e experiências testadas, bem como uma caracterização detalhada das características dos diferentes fonemas e a importância das mesmas no processo de análise e síntese de sinais de fala.

1.4 - Estrutura

O relatório desta dissertação está dividido em seis capítulos, que descrevemos de seguida.

No capítulo 1 introduz-se o problema abordado nesta dissertação e os objectivos da mesma.

No capítulo 2 são explicados o processo fonatório e a audição. Descrevem-se os órgãos envolvidos na produção da fala e na audição e as funções que desempenham.

No capítulo 3 é abordada a temática da fonética. São explicadas as diferentes características articulatórias e acústicas de classificar os fonemas, havendo um especial enfoque para os fonemas da língua inglesa e portuguesa alvo de estudo nesta dissertação.

No capítulo 4 é descrito o estado da arte. Os diferentes métodos utilizados nesta dissertação são caracterizados e as experiências efectuadas com cada um deles expostas.

No capítulo 5 é descrita a metodologia seguida na obtenção das diferentes avaliações e os resultados de cada experiência são discutidos.

No capítulo 6 apresentam-se as conclusões e o trabalho futuro a realizar.

Capítulo 2

Processo fonatório e audição

A fala é a mais importante forma de comunicação entre humanos. A produção da fala tem por base o desejo do orador em transmitir uma mensagem a um ou vários ouvintes. Para que tal aconteça uma série de processos neurológicos e musculares são desencadeados para produzir a onda sonora que transmite a mensagem. Do lado do ouvinte a onda sonora é captada e transformada pelo sistema auditivo em sinais neurológicos que são enviados para o cérebro para que a mensagem possa ser interpretada.

Para que possa haver comunicação entre o orador e o ouvinte, o primeiro tem que converter as suas ideias para uma língua que ambos percebam, utilizando palavras e formando frases, que respeitem as regras gramaticais da língua escolhida. O orador pode também utilizar entoação ou acentuar palavras para enfatizar determinados segmentos da mensagem.

2.1 – Aparelho fonético

Este capítulo centra-se no processo fonatório. Na grande maioria das línguas existentes a produção do sinal de fala inicia-se com o diafragma a comprimir os pulmões, obrigando os alvéolos pulmonares e brônquios a expelirem ar através da traqueia, que faz a interligação entre o pulmão esquerdo e o direito, para a laringe. O fluxo de ar passa depois pela cavidade faríngea para a cavidade oral e/ou nasal, acabando por sair pela boca e/ou narinas. A secção sagital média do aparelho vocal representada na figura 2.1 mostra todos os órgãos e articuladores envolvidos na produção da fala.

Do ponto de vista da engenharia, a produção da fala é vista como uma operação de filtragem acústica. O filtro principal usado nesse sistema acústico representa, na produção da fala, o tracto vocal (cavidade laríngea e cavidade oral) e o tracto nasal (cavidade nasal). O filtro acústico é normalmente excitado por um sinal que simula o efeito do fluxo de ar vindo dos pulmões a atravessar todos os articuladores até ser libertado pelos lábios. Esse modelo acústico tem por carga uma impedância de radiação, que representa o efeito dos lábios na produção da fala.

A separação entre o tracto vocal e o nasal é feita pelo palato, que é constituído pelo palato duro na parte anterior e pelo palato mole ou véu palatino na parte posterior. O palato mole termina na úvula. O comprimento médio do tracto vocal num homem adulto é cerca de 17 cm, enquanto que numa mulher adulta é cerca de 14 cm e numa criança cerca de 10 cm. Ao longo do tracto vocal a secção do mesmo pode variar entre 20 cm^2 e 0 cm^2. Esta variação da secção do tracto vocal é conseguida devido à mobilidade das pregas vocais, do palato mole ou véu palatino, da língua, da mandíbula e dentes e dos lábios que, com alterações das suas posições, conseguem alterar as dimensões do tracto vocal e assim alterar as propriedades acústicas do som emitido. O tracto nasal tem cerca de 12 cm num homem adulto e a ligação ao tracto vocal é feita por uma abertura controlada pelo palato mole. Quando o véu palatino está para baixo há ligação entre o tracto vocal e o nasal e a ligação entre ambos pode atingir os 5 cm^2 num homem adulto. Quando está levantado e encostado à cavidade faríngea a ligação fica completamente fechada e o fluxo de ar atravessa apenas o trato vocal. O facto dos dois tractos estarem isolados é crucial para o som que é libertado, não só porque o fluxo de ar deixa de passar pelo tracto nasal, mas também porque altera significativamente as propriedades do som radiado pelos lábios.

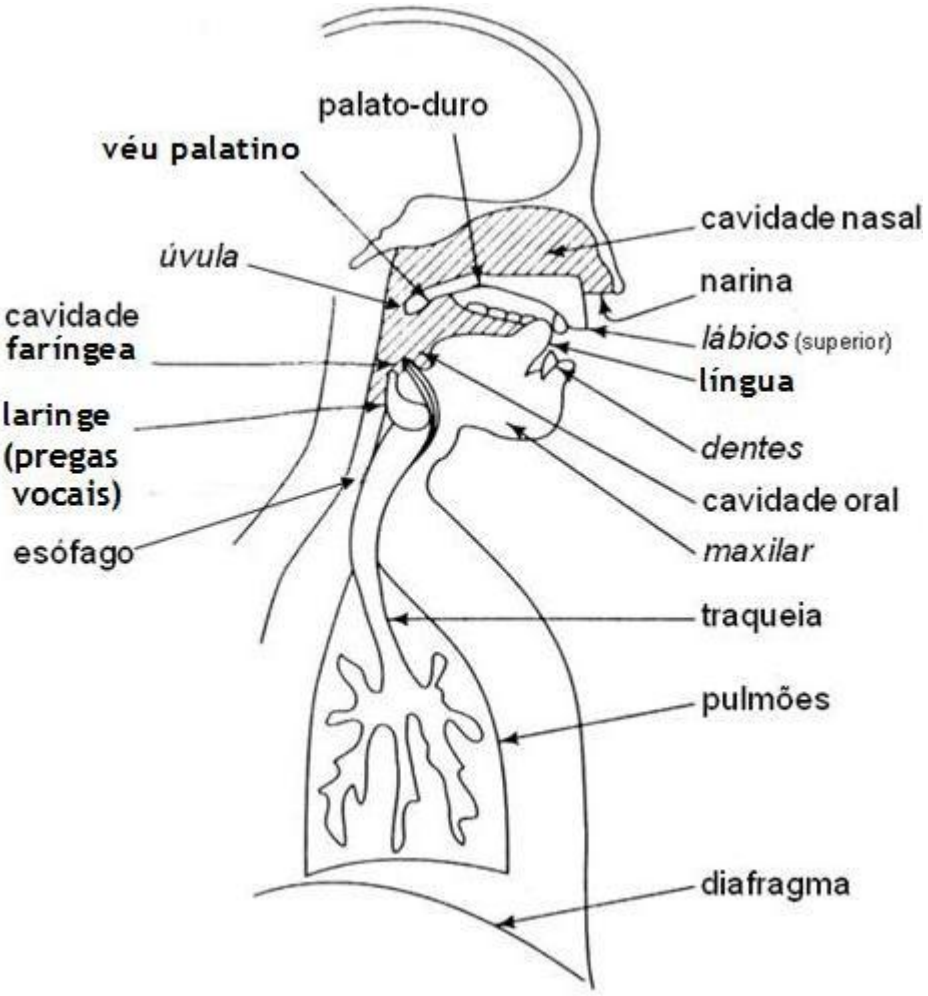


Figura 2.1 – Secção sagital média do aparelho vocal [1].

2.2 – Tipos de excitação dos sinais de fala

O tipo de excitação é uma das mais importantes características sonoras dos sinais de fala. Existem seis tipos de excitação: vozeado, não vozeado, misto, plosivo, sussurro e silêncio, sendo que as duas primeiras são as mais relevantes. O silêncio só é considerado como um tipo de excitação, porque ao analisar um sinal de voz surgem segmentos de silêncio, que correspondem a pausas no discurso, que necessitam de ser classificadas quanto ao tipo de excitação.

O vozeamento dos sinais de fala acontece quando o fluxo de ar vindo dos pulmões passa pela laringe e as pregas vocais interrompem esse fluxo de uma forma quase periódica, excitando assim o tracto vocal.

Na parte interior da laringe encontram-se as pregas vocais (ou cordas vocais), que são constituídas por ligamentos e músculos e ao espaço entre as duas pregas vocais (esquerda e direita) dá-se o nome de glote (figura 2.2).

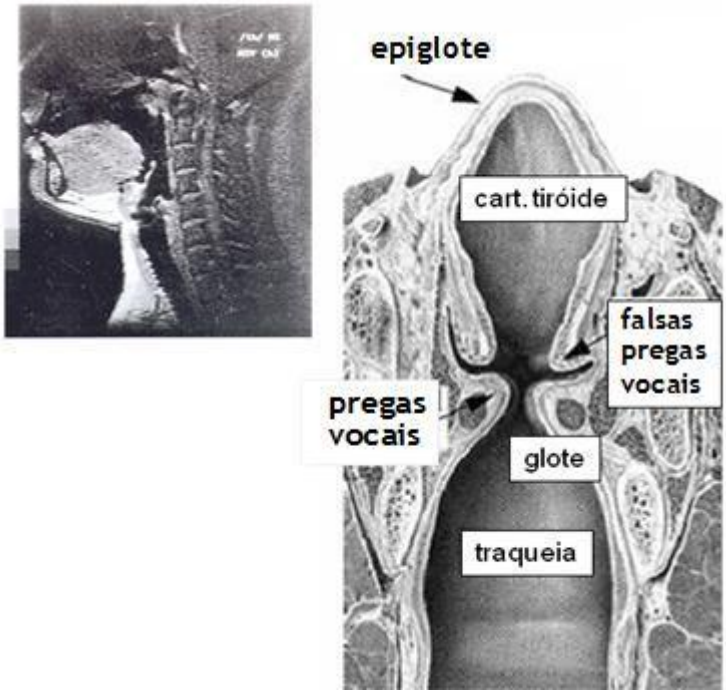


Figura 2.2 - Secções da laringe. Adaptado de [1].

As pregas vocais são responsáveis pelo vozeamento dos sinais de fala, ao abrir e fechar rapidamente a passagem do ar vindo dos pulmões. A junção das pregas vocais cria pressão do ar subglotal que vai aumentando até vencer a resistência das pregas vocais que se começam a separar uma da outra. Quando as pregas vocais se afastam, o fluxo de ar passa pela glote, o que origina um decréscimo da pressão de ar subglotal. A glote continua a abrir-se até atingir o seu máximo, quando a tensão natural das pregas vocais é igual à força de separação causada pela pressão de ar e a partir desse momento a glote começa-se a fechar. A força elástica das pregas vocais aumenta a velocidade de fecho da glote e quando a glote está suficientemente fechada verifica-se um efeito de sucção causado pela força de Bernoulli, que fecha a glote de forma abrupta. A pressão do ar subglotal volta a aumentar e o ciclo repete-se. Na figura 2.3 está representada a evolução do movimento das pregas vocais ao longo de

um ciclo vibratório e na figura 2.4 estão representadas as formas de onda do sinal do fluxo aerodinâmico que passa pela glote e da sua derivada também ao longo de um ciclo vibratório das pregas vocais. Como é possível observar na figura 2.4 a fase de abertura da glote é mais demorada e, como seria de esperar, o fluxo aerodinâmico é máximo quando a glote está mais aberta. A fase de fechamento da glote é mais curta, cerca de metade da fase de abertura, devido ao efeito de sucção causado pela força de Bernoulli o que provoca uma rápida redução do fluxo de ar que passa pela glote. Ao tempo entre duas sucessivas aberturas da glote chama-se período fundamental (*To*) e à frequência a que decorre essa abertura dá-se o nome de frequência fundamental ($Fo = 1/To$). A fase de abertura e fechamento da glote têm aproximadamente a mesma duração.

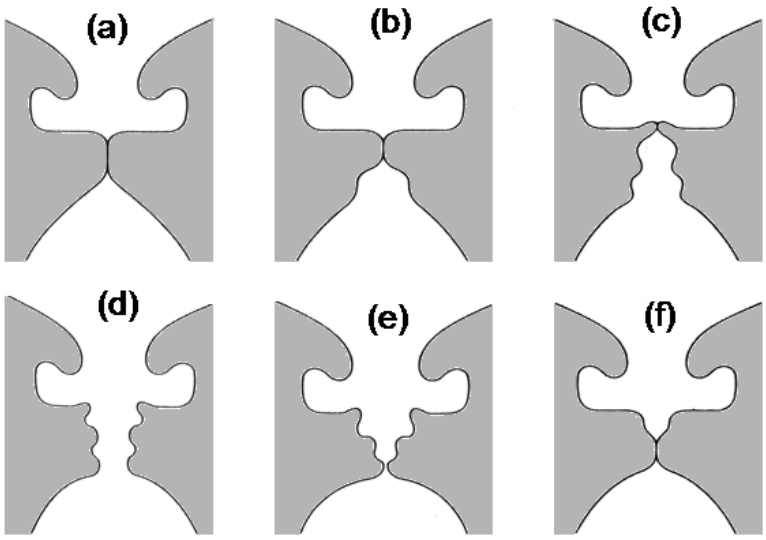


Figura 2.3 - Representação de um ciclo vibratório das pregas vocais [1].

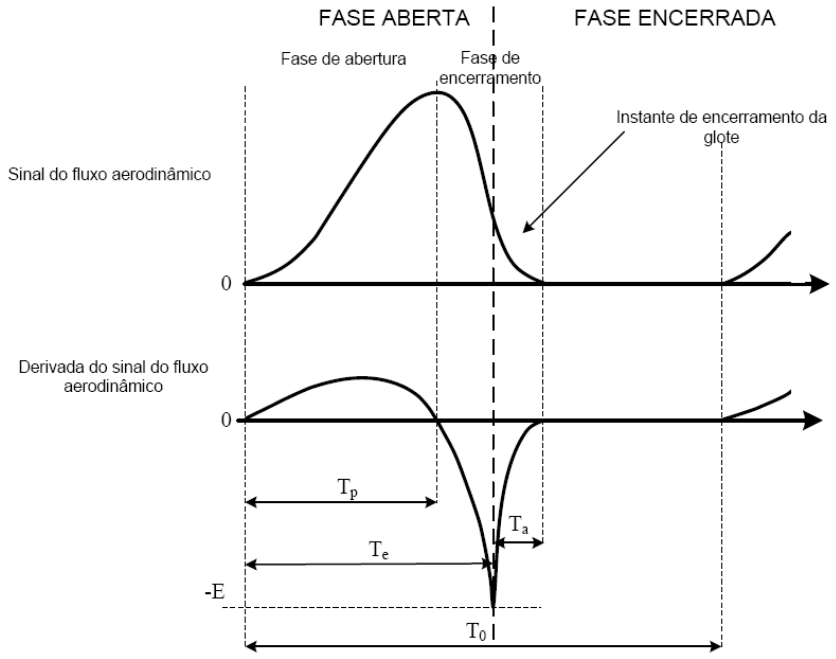


Figura 2.4 – A forma de onda de cima representada o sinal do fluxo aerodinâmico ao longo de um ciclo vibratório das pregas vocais e a forma de onda de baixo representa a derivada desse sinal.

Habitualmente a frequência fundamental para homens adultos está compreendida entre os 50 e os 250 Hz e para mulheres adultas está compreendida entre 120 e 350 Hz. A frequência fundamental varia consoante o comprimento, largura e a tensão das pregas vocais, a porção membranosa das mesmas, a cartilagem tiroideia e a largura da cavidade laríngea, o que origina que a frequência fundamental da voz vozeada de uma pessoa varia também com a idade [2].

Durante a produção dos sinais de fala não vozeados as pregas vocais não vibram. Estes sinais não vozeados são gerados posicionando os diferentes articuladores nas posições desejadas e forçando o fluxo de ar vindo dos pulmões a atravessar o tracto vocal, provocando desse modo turbulência. As diferenças entre os diferentes sons não vozeados dependem do tipo de obstrução no tracto vocal. Essas obstruções variam consoante o posicionamento dos diferentes articuladores que, alterando os tamanhos e as localizações das mesmas, modificam as características frequenciais dos sinais de fala.

Aos sinais de fala que têm simultaneamente características de sinais vozeados e de sinais não vozeados chamam-se sinais de fala mistos.

Aos sons constituídos por uma primeira fase de silêncio seguido por uma fase vozeada, não vozeada ou mista dá-se o nome de sons plosivos. Estes sinais de fala são gerados fechando completamente os lábios durante a fase de silêncio, retendo desse modo o fluxo de ar vindo dos pulmões e aumentando a pressão do ar junto dos lábios. O fluxo de ar é libertado abruptamente durante a segunda fase dos sinais plosivos, formando a fase vozeada, não vozeada ou mista.

A frequência fundamental dos sinais vozeados apresenta uma fase transitória quando precedidos por segmentos não vozeados ou de silêncio, pois devido à inércia das pregas vocais a frequência fundamental não é atingida instantaneamente.

Às frequências favorecidas pelos tractos nasal e vocal dá-se o nome de frequências formantes, ou simplesmente formantes. Por vezes, a cavidade nasal, desfavorece a passagem de certas frequências criando anti-ressonâncias, também conhecidas por anti-formantes [3]. Certos fonemas, como será demonstrado mais abaixo, possuem formantes e/ou anti-formantes características que os diferenciam de todos os outros fonemas e que são de grande importância na análise de sinais de fala.

2.3 – Aparelho auditivo

2.3.1 – Anatomia do ouvido

O ouvido humano está dividido em três sub-regiões o ouvido externo (orelha e canal auditivo), o ouvido médio (tímpano e os ossículos martelo, bigorna e estribo) e o ouvido interno (cóclea e nervos auditivos). Na figura 2.5 estão ilustrados os diferentes elementos que constituem o ouvido humano.

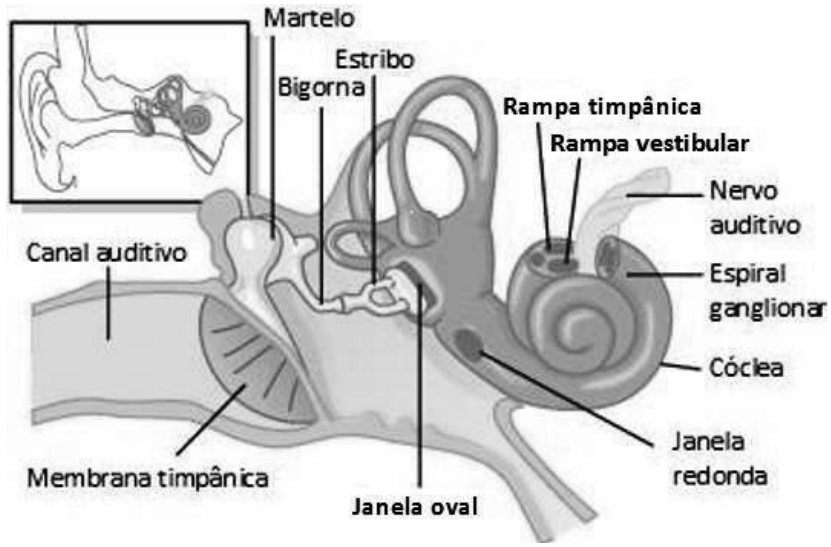


Figura 2.5 - Representação do ouvido humano. Adaptado de [4].

A onda sonora é captada pela orelha e transmitida pelo canal auditivo até ao tímpano, pondo-o em oscilação. O tímpano transmite a energia mecânica através de três ossículos (martelo, bigorna e estribo) a uma membrana, designada de janela oval que comunica as oscilações a um meio aquoso. A relação de impedâncias entre os dois meios é obtida através do efeito de alavanca proporcionado pelos ossículos e a relação de áreas entre o tímpano e a janela oval. Os ossículos também protegem o ouvido pois, na presença de intensidades sonoras demasiado elevadas, os pequenos músculos que controlam os ossículos conseguem reduzir o ganho de transmissão da energia acústica [5]. A cóclea é uma estrutura rígida (óssea) em forma de caracol e possui três canais paralelos e enrolados, chamados rampa vestibular, ducto coclear e rampa timpânica. Estes canais estão preenchidos com líquido e separados entre si por membranas elásticas. A rampa vestibular começa na janela oval e tem ligação com a rampa timpânica no outro extremo da cóclea. O outro extremo da rampa

timpânica é uma membrana, a janela redonda, que a separa do ouvido médio. É na cóclea que a energia mecânica é convertida nos impulsos nervosos que são posteriormente enviados para o cérebro. Essa conversão é realizada por milhares de células ciliadas distribuídas ao longo de uma membrana flexível, a membrana basilar, que separa a rampa timpânica do ducto coclear. A membrana basilar tem cerca de 35mm de comprimento e a sua rigidez e grossura variam ao longo do seu comprimento, sendo a extremidade junto à janela oval a mais fina e a mais rígida e a extremidade oposta mais grossa e flexível. As diferenças das propriedades mecânicas ao longo da membrana basilar fazem com que a zona estimulada esteja dependente da frequência do som recebido. A membrana basilar efectua, portanto, uma análise espectral à onda sonora captada. As variações de pressão criadas por um sinal sonoro e comunicadas ao interior da cóclea através da janela oval, escapam-se para a janela redonda escolhendo o ponto da membrana basilar de menor impedância (figura 2.6). As células ciliadas presentes na zona da membrana basilar mais estimulada geram os impulsos nervosos a uma cadência superior e esses impulsos são depois transmitidos através do nervo auditivo ao cérebro.

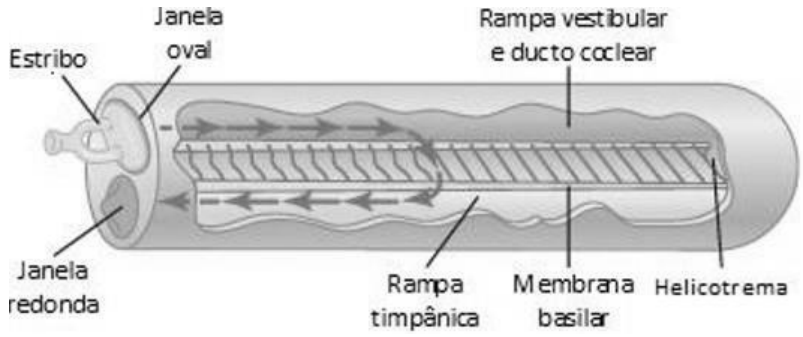


Figura 2.6 - Representação do interior da cóclea. Adaptado de [4]

2.3.2 – Funcionamento do ouvido

2.3.2.1 - Percepção de intensidade

As características principais para a percepção de um som são a sua frequência e intensidade. A intensidade sonora mede-se habitualmente em dB SPL (deciBel Sound Presure Level), que corresponde ao logaritmo da relação entre a pressão acústica e o valor de referência 20μPa (ou 10^{-12} W/m^2).

$$L_{SPL} = 20\log(p/p_0)\ dB \tag{2.1}$$

onde L_{SPL} é a intensidade em SPL de um estimulo, p é a pressão acústica do estimulo em Pascals e p_0 é o nível de referência standard (20μPa).

À intensidade sonora mínima de um som a uma determinada frequência e que o torna perceptível ao ouvido de um humano dá-se o nome de limiar absoluto de audição. O limiar da dor define a intensidade sonora a partir da qual um som começa a causar dor ao ouvinte. Sons com intensidades superiores ao limiar da dor, para além de causarem sofrimento, podem provocar perdas auditivas permanentes.

Na figura 2.7 estão representadas várias curvas de idêntica sonoridade ou *equal-loudness contours*, bem como as curvas representativas do limiar de audibilidade e limiar de dor. As curvas de idêntica sonoridade são obtidas ajustando a intensidade sonora de um sinal com uma determinada frequência, até que, perceptivamente, este possua a mesma intensidade de um tom puro (i.e., um sinal de uma só frequência, ou frequência pura) de 1000 Hz, regulado para uma dada pressão acústica em dB SPL [5].

O limiar absoluto de audição caracteriza a quantidade de energia necessária num tom para que possa ser detectado por um ouvinte num ambiente silencioso [6] e é um caso especial de uma curva de idêntica sonoridade. O limiar de audibilidade pode ser aproximado pela equação (2.2) [6].

$$L = 3.64\left(\frac{f}{1000}\right)^{-0.8} - 6.5e^{-0.6\left(\frac{f}{1000}-3.3\right)^2} + 10^{-3}\left(\frac{f}{1000}\right)^4 \ (dB\ SPL) \qquad (2.2)$$

A sonoridade ou "*loudness*" tem como unidade de medição o Phon.

Observando a figura 2.7 é possível retirar algumas conclusões a propósito da sensibilidade auditiva humana. O ouvido humano é mais apurado para frequências entre os 2 e os 5 kHz (região preta). Para frequências inferiores a 100 Hz e superiores a 10 kHz, o ouvido humano perde rapidamente sensibilidade, sendo naturalmente surdo a sinais de frequência inferior a 20 Hz (infra-sons) e a sinais de frequência superior a 20 kHz (ultra-sons) [5]. A região cinzento-claro assinala a região é que o ouvido humana perde significativamente sensibilidade a frequências inferiores a 100 Hz. Por fim, a região cinzento-escuro indica a gama típica de frequências e intensidades da fala.

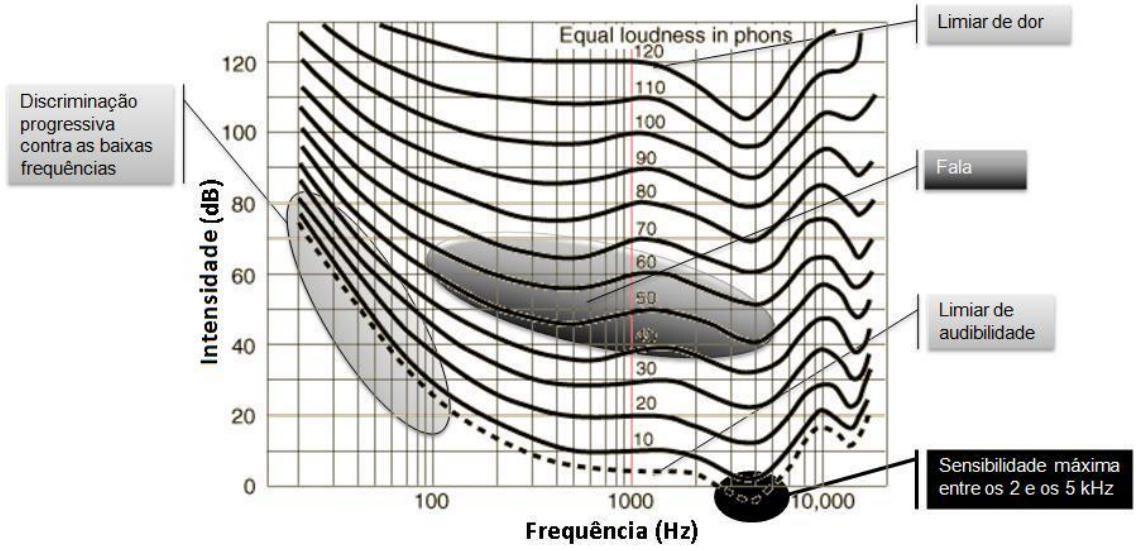


Figura 2.7 – Gráfico Intensidade-Frequência sobreposto com diversas curvas de idêntica sonoridade. Adaptado de [1].

2.3.2.2 - Efeito de máscara ou mascaramento

O efeito de máscara ou mascaramento consiste na influência que uma dada componente de som (mascarante) exerce na audibilidade de uma outra componente de som (mascarada) na vizinhança da primeira. O mascaramento depende da intensidade, da frequência e do local e tempo da ocorrência das duas componentes de som. O mascaramento pode ser parcial,

quando o som mascarante reduz a audibilidade do som mascarado ou total, quando é impossível ouvir o som mascarado [5].

O termo banda crítica está ligado a um estudo, realizado por Harvey Fletcher em 1940, sobre a capacidade de uma banda estreita de ruído mascarar um tom puro. O tom puro é posto no limiar do mascaramento total. Qualquer pequeno aumento da intensidade do tom puro faz com que ele passe a ser audível e para que este volte a estar completamente mascarado é necessário que a largura de banda do ruído seja aumentada. Mas a partir de um determinado valor é impossível compensar o aumento da intensidade do tom puro com um aumento da largura da banda do ruído. Para essa largura de banda em que o tom puro ainda se encontra totalmente mascarado dá-se o nome de banda crítica.

A largura das bandas críticas depende da frequência. A largura de banda das bandas críticas é de aproximadamente 100 Hz até aos 500 Hz, crescendo a partir dessa frequência para cerca de 20 por cento da frequência central [5]. Devido à sua importância criou-se uma nova escala de frequências para representar o espectro audível para os humanos, em que cada unidade representa uma banda crítica e tem como unidade de medição o Bark. A gama de frequências audíveis (até aos 20 kHz) é composta por cerca de 25 Bark (figura 2.8).

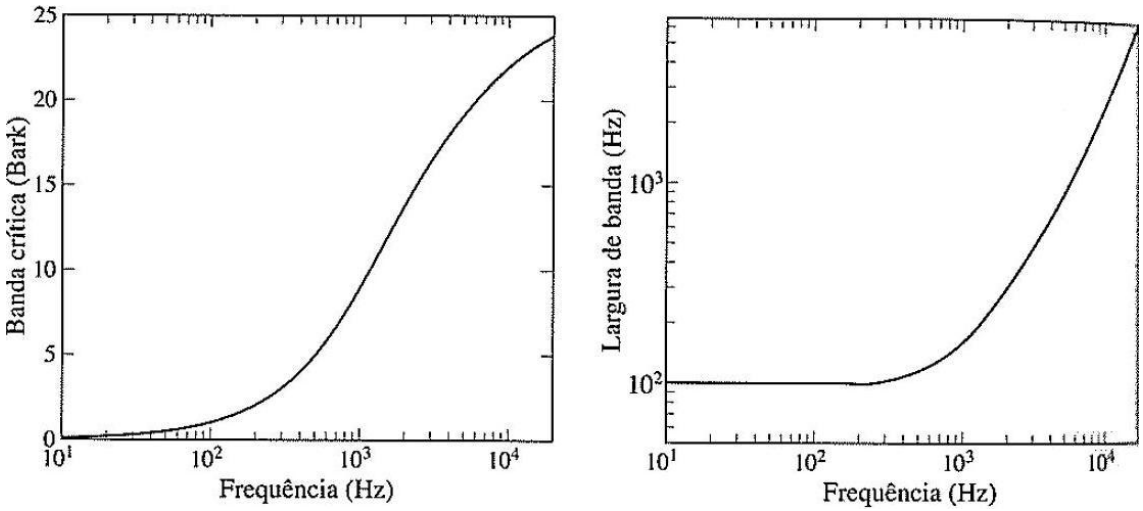


Figura 2.8 – Correspondência entre as escalas de frequências Hertz e Bark (à esquerda) e largura de banda da escala Bark (à direita) [5].

A curva representada na figura 3.9 designa-se por curva de mascaramento e traduz o limiar do mascaramento total (*Threshold of Masking*) devido ao tom puro mascarante. Qualquer sinal na vizinhança do tom puro mascarante e com amplitude abaixo da curva de mascaramento será inaudível. A forma das curvas de mascaramento depende de diversos factores, incluindo a intensidade e frequência, apesar disso na escala Bark a forma da curva pode ser aproximada pelo modelo apresentado na figura 2.9.

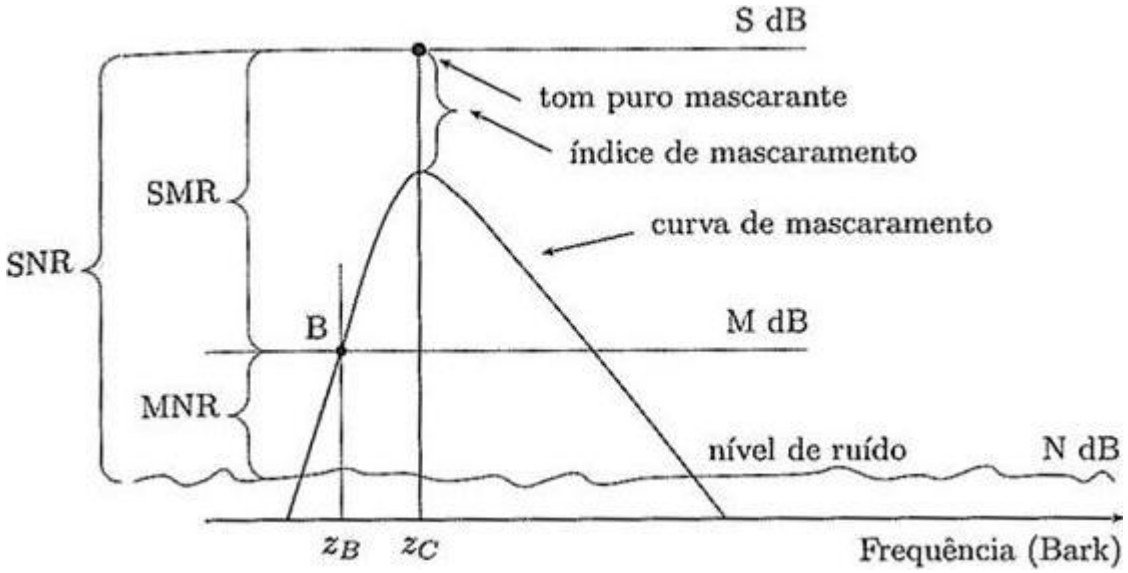


Figura 2.9 – Modelo da curva de mascaramento na escala Bark. Ilustra-se a utilização desta curva para calcular o limiar de mascaramento à frequência z_B, devido a um tom puro mascarante à frequência z_C [5].

A equação (2.3) relaciona o valor *z*, em Bark, com a frequência *f* em Hz.

$$z = 13 \arctan\left(\frac{76f}{10^5}\right) + 3.5\ arctan\left(\frac{f}{7500}\right)^2 \tag{2.3}$$

Na figura 2.9 está representado o limiar de mascaramento devido a um tom puro de intensidade S dB e à frequência z_C Bark. Um modelo possível para a curva de mascaramento é o proposto por Schroeder, Atal e Hall [5]:

$$CM_{dB} = 15.81 + 7.5(\mathrm{z} + 0.474) - 17.5\sqrt{1 + (\mathrm{z} + 0.474)^2}. \tag{2.4}$$

Como é possível observar na figura 2.9 o efeito de máscara devido ao tom puro é assimétrico, influenciando mais as frequências superiores à sua.

Apesar da maioria dos estudos se debruçar sobre os efeitos de mascaramento dentro da mesma banda crítica, os efeitos do mesmo fazem-se sentir a frequências fora dessa banda crítica.

O efeito de máscara ou mascaramento pode ser classificado quanto à relação temporal da ocorrência das componentes do som. Este pode ser simultâneo, quando as componentes mascarante e mascarada coexistem temporalmente, pré-mascaramento, quando a componente mascarada ocorre antes da componente mascarante ou pós-mascaramento, quando a componente mascarada ocorre depois da componente mascarante.

O efeito de mascaramento é máximo quando o sinal mascarante e mascarado coincidem temporalmente (mascaramento simultâneo) e o seu efeito decresce rapidamente com o aumento do intervalo temporal entre os dois acontecimentos. Apesar da intensidade do efeito de máscara variar muito com a natureza dos sinais mascarante e mascarado, diversos estudos realizados indicam que os efeitos do pós-mascaramento são mais prolongados que os do pré-mascaramento.

Capítulo 3

Critérios utilizados na classificação fonética

A forma da onda dos sinais de fala reais varia com o tempo, ou seja, os sinais de fala são não-estacionários. Devido às rápidas alterações das propriedades acústicas e espectrais destes sinais num curto espaço de tempo, é necessário subdividir estes sinais de fala em segmentos de curta duração que tenham características semelhantes para que estas possam ser eficazmente analisadas.

Na linguística os fonemas são a unidade elementar da fala e as suas características acústicas e espectrais diferem de língua para língua. Um fone é o som efectivamente produzido na realização de um fonema. Essa realização difere consoante diversos factores, como por exemplo, sexo, idade e região de um indivíduo. Assim, associado a cada fonema está um conjunto de fones com ligeiras variações acústicas, sendo que a essa colecção de fones se dá o nome de alofones [7]. A maioria das palavras é composta por mais do que um fonema e cada fonema difere dos restantes na duração, no tipo de excitação e no posicionamento dos diferentes articuladores durante a sua produção. A transição na mesma palavra de um fonema para um outro é feita de forma contínua, o que implica que a transição entre fonemas das propriedades acústicas e espectrais também varia continuamente. As fases de transição entre fonemas acontecem, porque ao transitar de um fonema para outro é necessário rearranjar o posicionamento dos articuladores de modo a alterar o formato do tracto vocal para produzir o novo fonema e, como esses reajustamentos são feitos por músculos, é impossível realizá-los instantaneamente. Por este motivo o mesmo fonema inserido em duas palavras distintas pode não ter exactamente nem a mesma duração, nem as mesmas características sonoras e espectrais, pois essas mesmas características estão dependentes do fonema anterior e do fonema posterior. Nos sinais de fala reais, quando se fala depressa e os fones são de muito curta duração, por vezes dá-se o caso, do posicionamento final normal dos articuladores ao produzir um determinado fonema não chegar a ser atingido, pois a fase de transição desse fonema com o fonema anterior e com o fonema seguinte estão parcialmente sobrepostas. Após a fase de transição as diferenças entre fonemas iguais de palavras distintas produzidos pelo mesmo indivíduo em circunstâncias parecidas são normalmente reduzidas.

3.1 – Tipos de fonemas

As letras do alfabeto não são a melhor forma de representar um fonema, pois normalmente não existe uma correspondência directa entre uma letra e as características acústicas da mesma quando está inserida numa determinada palavra. Para que não houvesse ambiguidades quanto à sonoridade das palavras foram criados diferentes "alfabetos fonéticos", estes alfabetos consistem em associar um símbolo diferente a cada fonema utilizado na produção da fala. O primeiro "alfabeto fonético" foi criado em 1888 na Europa e recebeu o nome de International Phonetic Alphabet (IPA), em português Alfabeto Fonético Internacional (AFI). A versão completa do alfabeto IPA contém os fonemas de todas as línguas faladas no mundo. O alfabeto IPA não pode ser escrito numa máquina de escrever ou computador e por esse motivo ao longo dos anos foram surgindo novos "alfabetos fonéticos" que não tivessem essas limitações. Entre esses encontra-se o "alfabeto fonético" SAMPA (Speech Assessment Methods Phonetic Alphabet) e o ARPAbet, desenvolvido pela United States Advanced Research Projects Agency (ARPA). Como alguns símbolos usados no ARPA são também letras do alfabeto, quando estas representarem um fonema serão delimitadas com o símbolo "/", deste modo evita-se que haja confusão entre a palavra "e" e o fonema /e/.

Existem diversas formas de classificar fonemas. Esses critérios podem-se agrupar em duas grandes categorias, os que utilizam características acústicas e os que usam características articulatórias para analisar os fonemas. Os critérios que analisam os fonemas quanto ao tipo de excitação, modo de articulação, ponto de articulação e estacionaridade do sinal de fala são critérios articulatórios. Este último é simultaneamente um critério acústico. Os outros critérios acústicos focam a sua análise dos sinais de fala nas suas características no domínio dos tempos ou nas suas características no domínio das frequências.

Na figura 3.1 é possível ver uma classificação em forma de árvore de todos os fonemas representados no ARPAbet do inglês americano. A primeira separação da árvore consiste na utilização do critério da estacionaridade do fonema. No grupo dos fonemas não-contínuos estão os fonemas que para serem gerados necessitam que os articuladores modifiquem a configuração do tracto vocal de forma significativa, pois para a sua produção contribuem mais do que um "estado sonoro", enquanto que os restantes fonemas são produzidos com os articuladores numa posição estática ou com movimentações muito ligeiras durante a passagem do fluxo de ar. Os fonemas não-contínuos diferem dos fonemas contínuos, pois para serem produzidos é obrigatório movimentar um ou vários articuladores para que seja possível proceder às alterações necessárias na configuração do tracto vocal. Estes fonemas são habitualmente mais difíceis de caracterizar e modelar do que os fonemas contínuos devido às transições que apresentam ao longo do seu tempo de produção. O subgrupo dos fonemas contínuos é posteriormente novamente dividido em dois grandes grupos: o grupo das vogais e o grupo das consoantes.

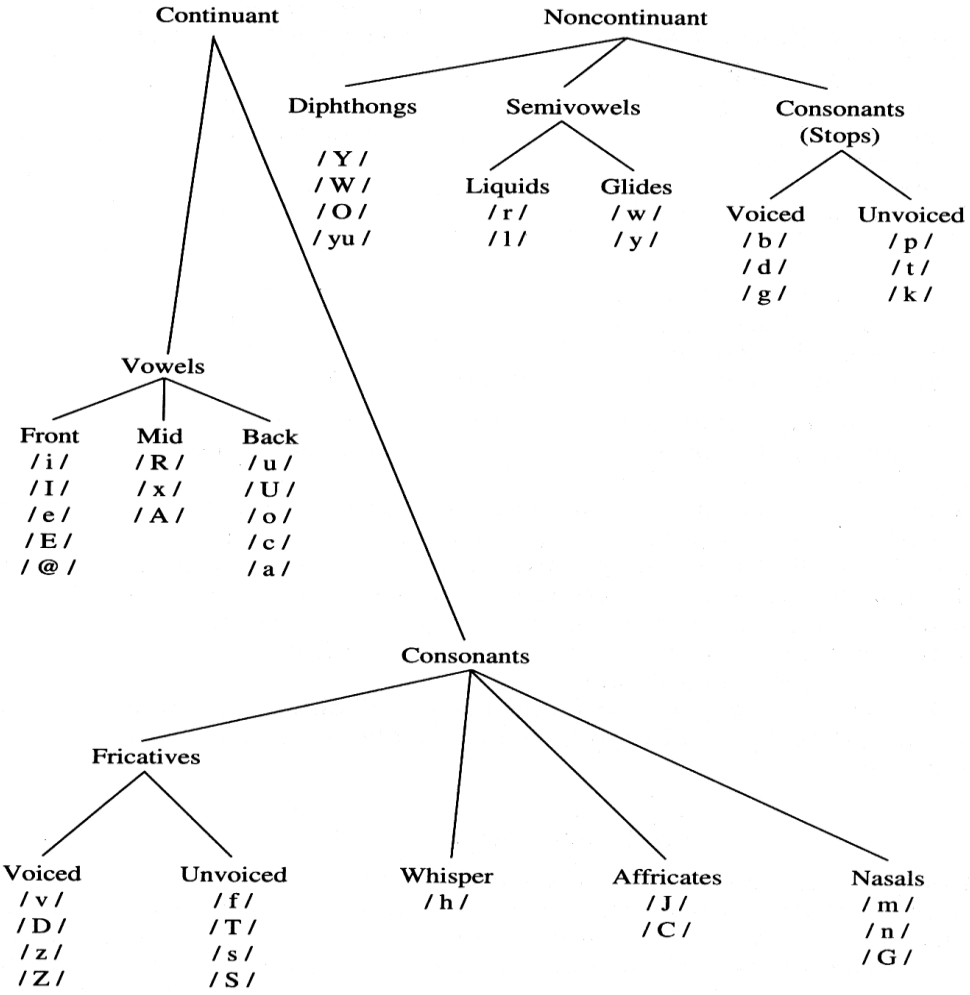


Figura 3.1 - Classificação de fonemas ingleses – ARPAbet [7].

3.1.1 – Vogais

As vogais são fonemas diferentes de todos os outros, pois têm somente uma reduzida obstrução ao longo de todo o tracto vocal, para além de serem todas vozeadas e da onda do seu sinal ter maior amplitude que os restantes. As vogais são posteriormente repartidas em três subgrupos tendo em conta o seu ponto de articulação, que no caso das vogais consiste principalmente na posição da língua, que tanto pode estar à frente (*front vowels*), como atrás (*back vowels*) ou então numa posição intermédia (*mid vowels*).

3.1.2 – Consoantes

As consoantes contínuas têm uma maior obstrução do tracto vocal do que as vogais e por esse motivo apresentam uma menor amplitude na onda do seu sinal. Estas consoantes estão subdivididas em fricativas, aspiradas, africadas e nasais.

As consoantes fricativas são geradas excitando o tracto vocal com um fluxo constante de ar, que se torna turbulento ao passar pela zona parcialmente obstruída. As fricativas são divididas em fricativas não-vozeadas, quando o fluxo de ar é contínuo, ou vozeadas, quando esse fluxo de ar é quase periodicamente interrompido pela vibração das pregas vocais.

Existe apenas uma consoante aspirada, o fonema /h/, que não existe na língua portuguesa. Este fonema quando está presente é no início de uma palavra, por exemplo na palavra inglesa “head” ou na palavra alemã “Hamburg”.

As consoantes africadas são constituídas transitando de uma consoante plosiva para uma fricativa. As consoantes africadas, ao contrário do português do Brasil ou da língua inglesa, não existem no português europeu. Estas consoantes africadas tanto podem ser não-vozeadas, por exemplo o fonema /C/, que consiste na transição da consoante plosiva não-vozeada /t/ para a consoante fricativa não-vozeada /S/, como vozeadas, por exemplo o fonema /J/, que é composto pela passagem da consoante plosiva vozeada /d/ para a consoante fricativa vozeada /Z/.

Os fonemas não-contínuos diferem dos contínuos, pois durante a sua produção é necessário alterar a forma do tracto vocal movendo um ou mais articuladores. Os fonemas não-contínuos estão subdivididos em plosivas, semivogais e ditongos.

As consoantes plosivas têm algumas parecenças com as consoantes nasais, pela forma como são produzidas, pois ambas têm durante a fase inicial a passagem do fluxo de ar pela boca completamente obstruídas, mas no caso das consoantes plosivas o caminho alternativo pelo tracto nasal também está fechado, pois o palato mole ou véu palatino está encostado à cavidade laríngea. Como ambos os trajectos estão tapados, o ar vindo dos pulmões é acumulado junto à obstrução, que tanto pode ser nos lábios como na língua. Na segunda fase da produção do fonema a pressão acumulada é libertada, aquando da desobstrução do tracto vocal. As consoantes plosivas tanto podem ser vozeadas como não-vozeadas dependendo da vibração ou não-vibração das pregas vocálicas durante a segunda fase do fonema.

3.2 – Diferenças entre os fonemas do inglês americano e do português europeu

Os “alfabetos fonéticos” ARPA e IPA (AFI) para o inglês americano podem ser consultados na tabela 3.1. O número de fonemas usados numa determinada língua depende do “alfabeto fonético” utilizado, mas tanto a língua inglesa como a portuguesa são habitualmente representadas por cerca de 40 fonemas. A correspondência entre grafemas e a fonética nem sempre é a mesma entre as duas línguas.

Tabela 3.1 - Alfabeto fonético internacional (IPA ou AFI) e ARPAbet para o inglês americano [1].

ARPAbet	IPA	Exemplo	ARPAbet	IPA	Exemplo
i	i	heed	v	v	vice
I	ɪ	hid	T	θ	thing
e	e	hayed	D	ð	then
E	ɛ	head	s	s	so
@	æ	had	z	z	zebra
a	ɑ	hod	S	ʃ	show
c	ɔ	hawed	Z	ʒ	measure
o	o	hoed	h	h	help
U	ʊ	hood	m	m	mom
u	u	who'd	n	n	noon
R	ɝ	heard	G	ŋ	sing
x	ə	ago	l	l	love
A	ʌ	mud	L	l	cattle†
Y	ɑɪ	hide	M	m	some†
W	ɑʊ	how'd	N	n	son†
O	ɔɪ	boy	F	f	batter‡
X	ɨ	roses	Q	ʔ	§
p	p	pea	w	w	want
b	b	bat	y	j	yard
t	t	tea	r	r	race
d	d	deep	C	tʃ	church
k	k	kick	J	dʒ	just
g	g	go	H	ʍ	when
f	f	five			

Na tabela 3.2 podem ser consultados os "alfabetos fonéticos" IPA (AFI) e SAMPA para o português europeu, os grafemas que podem simbolizar esses fonemas e algumas palavras a servir de exemplo.

Tabela 3.2 – Exemplos de correspondências entre os "alfabetos fonéticos" API (AFI) e SAMPA e os grafemas utilizados no alfabeto português. Na primeira coluna estão representados todos os fonemas do IPA (AFI) para o português europeu, na segunda coluna os mesmos fonemas estão representados no "alfabeto fonético" SAMPA, na terceira os grafemas do português que podem representar esses fonemas e na última coluna são indicados alguns exemplos de palavras portuguesas para cada um dos fonemas [14].

AFI	SAM-PA	Grafemas	Exemplo	AFI	SAM-PA	Grafemas	Exemplo
i	i	i,í,y,e	v*i* [ví]	p	p	p	*p*á [pá]
e	e	e,ê	v*ê* [vé]	b	b	b	*b*em [bẽj]
ɛ	E	e,é	p*é* [pɛ́]	t	t	t	*t*u [tú]
a	a	a,á,à	p*á* [pá]	d	d	d	*d*ou [dó]
ɐ	6	a	c*a*ma [cɐ́mɐ]	k	k	c,k	*c*asa [kázɐ]
ɨ	@	e	d*e* [dɨ]	g	g	g	gato [gátu]
ɔ	O	ó,o	p*ó* [pɔ́]				
o	o	ô,o	av*ô* [ɐvó]	f	f	f	*f*é [fɛ́]
u	u	ú,u	t*u*do [túdu]	v	v	v	*v*ê [vé]
j	j	i,e	pa*i* [páj]	s	s	s,ç,c	*s*ol [sɔ́ɬ]
w	w	u,o	pa*u* [páw]	z	z	z,s,x	ca*s*a [kázɐ]
				ʃ	S	ch,s,z,x	*ch*ave [ʃávɨ]
ĩ	i~	i,í	s*im* [sĩ́]	ʒ	Z	j,g,s,z,x	*j*á [ʒá]
ẽ	e~	e,ê	p*en*te [pẽ́tɨ]				
ɐ̃	6~	ã,a,e	br*an*co [brɐ̃́ku]	l	l	l	*l*á [lá]
õ	o~	õ,o,ô	p*on*te [pṍtɨ]	ɬ	l~	l	ma*l* [máɬ]
ũ	u~	u,ú	at*um* [ɐtṹ]	ʎ	L	lh	va*lh*a [váʎɐ]
j̃	j~	i,e	põ*e* [pṍj̃]				
w̃	w~	o	mã*o* [mɐ̃́w̃]	m	m	m	*m*ão [mɐ̃́w̃]
				n	n	n	*n*ão [nɐ̃́w̃]
				ɲ	J	nh	se*nh*a [sɐ́ɲɐ]
				ɾ	r	r	ca*r*o [káɾu]
				ʀ	R	r	ca*rr*o [káʀu]

Na tabela 3.3 estão representados, na primeira coluna todos os grafemas presentes no alfabeto português, na segunda todos os fonemas do Alfabeto Fonético Internacional (IPA ou AFI) que esses grafemas podem simbolizar e na última coluna palavras portuguesas como exemplo.

Tabela 3.3 – Exemplos de correspondências entre símbolos gráficos e sons na ortografia do português europeu padrão. Na primeira coluna estão representados todos os grafemas simples, na segunda coluna as suas correspondências fonéticas de acordo com o alfabeto Fonético Internacional e na terceira coluna são indicados alguns exemplos de palavras portuguesas para cada um dos fonemas. Em cada palavra o fonema que se pretende exemplificar está a escrito a negrito.

Grafemas	Sons	Palavras
< a, A >	[a] [ɐ]	Arte, sapato Ametista, farelo
< b, B >	[b]	Batata, cabide
< c, C >	[k] [s]	Caruma, faca Cigarra, hélice
< d, D >	[d]	Dama, amador
< e, E >	[ɛ] [e] [i] [j] [ɐ] [ɾ̃] [i]	Canela Caneta Pedal Meada Espelho Mãe E
< f, F >	[f]	Família, afinador
< g, G >	[g] [ʒ]	Gavião, apagador Girafa, mágico
< i, I >	[i] [j]	Igreja, sinal Iate, gaivota
< j, J >	[ʒ]	Janela, poejos
< l, L >	[l] [ɫ]	Laranja, mala, Altura
< m, M >	[m]	Morango, camélia
< n, N >	[n]	Notário, canário
< o, O >	[ɔ] [o] [u] [w] [ŵ]	Porta Sopa Porteira Água Limão
< p, P >	[p]	Panela, mapa
< r, R >	[ɾ] [ʁ]	Maroto, artista Rodada, honra
< s, S >	[s] [z] [ʃ] [ʒ]	Sapato, balsa Riso, casa Lápis, astro Asno, esmero
< t, T >	[t]	Teatro, caneta
< u, U >	[u] [w]	Unha, mula Pauta, vau
< v, V >	[v]	Viola, avião
< x, X >	[ʃ] [z] [ks]	Xaile, vexame Exame Tóxico
< z, Z >	[z]	Zebra, azeite

Na tabela 3.4 estão representados, na primeira coluna todos as sequências de grafemas, grafemas compostos e dígrafos do português europeu que têm uma única correspondência fonética no Alfabeto Fonético Internacional (IPA ou AFI), na segunda essas mesmas correspondências e na última coluna palavras portuguesas como exemplo dessa correspondência.

Tabela 3.4 – Exemplos de correspondências entre símbolos gráficos e sons na ortografia do português europeu padrão. Na primeira coluna estão representados as sequências de grafemas e grafemas compostos, na segunda coluna as suas correspondências fonéticas de acordo com o alfabeto Fonético Internacional e na terceira coluna são indicados alguns exemplos de palavras portuguesas para cada um dos fonemas. Em cada palavra o fonema que se pretende exemplificar está a escrito a negrito.

Sequências de grafemas e alguns grafemas compostos		
< ã, an, am, Ã, AN, AM >	[ɐ̃]	**Fã**, **banco**, **campo**
< en, em, EN, EM >	[ɛ̃]	**Vento**, **emprego**
< õ, on, om, Õ, ON, OM >	[ɔ̃]	**Anões**, **lontra**, **compras**
< in, im, IN, IM >	[ĩ]	**Pintor**, **imposto**
< un, um, UN, UM >	[ũ]	**Mundo**, **umbigo**
Grafemas compostos		
< á, à, Á, À >	[a]	Armário, **à**
< é, É >	[ɛ]	**Café**
< í, Í >	[i]	**Saída**
< ó, Ó >	[ɔ]	**Órbita**
< ú, Ú >	[u]	**Saúde**
< â, Â >	[ɐ]	**Âmago**
< ê, Ê >	[e]	**Inglês**
< ô, Ô >	[o]	**Bisavô**
< ç, Ç >	[s]	**Palhaço**
Dígrafos		
< ch, Ch, CH >	[ʃ]	**Chapéu**
< lh, LH >	[ʎ]	**Gargalhada**
< nh, NH >	[ɲ]	**Galinha**
< qu, Qu, QU >	[k]	**Qu**intal
< rr, RR >	[ʁ]	**Garrafa**
< ss, SS >	[s]	**Assador**

Consultando as tabelas 3.1, 3.2, 3.3 e 3.4 é possível retirar algumas conclusões sobre as diferenças entre os fonemas existentes no inglês americano e no português europeu, bem como os grafemas utilizados em ambas as línguas.

Comparando os grafemas é de salientar que na língua inglesa não existe nem acentuação nem til e por esse motivo não há grafemas compostos. Dos dígrafos indicados na tabela 3.4 apenas o /ch/ e o /qu/ existem também em inglês e não com a mesma correspondência fonética e as letras “k”, “w” e “y” não existem no alfabeto português.

Em termos de fonemas as principais diferenças entre as duas línguas são as vogais e o facto de no português europeu não haver consoantes aspiradas, /h/, nem africadas, /C/ e /J/. Na figura 3.2 pode ser observado o mapeamento das vogais orais de acordo com o Alfabeto Fonético Internacional. As vogais estão distribuídas horizontalmente quanto ao ponto de articulação, ou seja, quanto à localização de maior obstrução do tracto vocal provocada pela língua, que pode ser à frente, no centro ou a trás e verticalmente quanto à elevação da língua, “aberta” se a língua estiver para baixo e gradualmente mais elevada até estar muito próxima do palato duro, que corresponde à posição “fechado”. As vogais com um rectângulo ponteado à sua volta existem no português europeu, mas não no inglês americano, as com um rectângulo tracejado correspondem à situação inversa e as com um rectângulo com traço contínuo existem em ambas as línguas.

Existem outras diferenças entre as duas línguas, mas não afectam uma classe fonética inteira, apenas alguns fonemas isolados, como é o caso dos fonemas /T/, /D/ e /G/ que apenas existem em inglês e dos fonemas /ɬ/, /ʎ/, /ɲ/, /ɾ/ e /ʀ/ que apenas existem no português europeu.

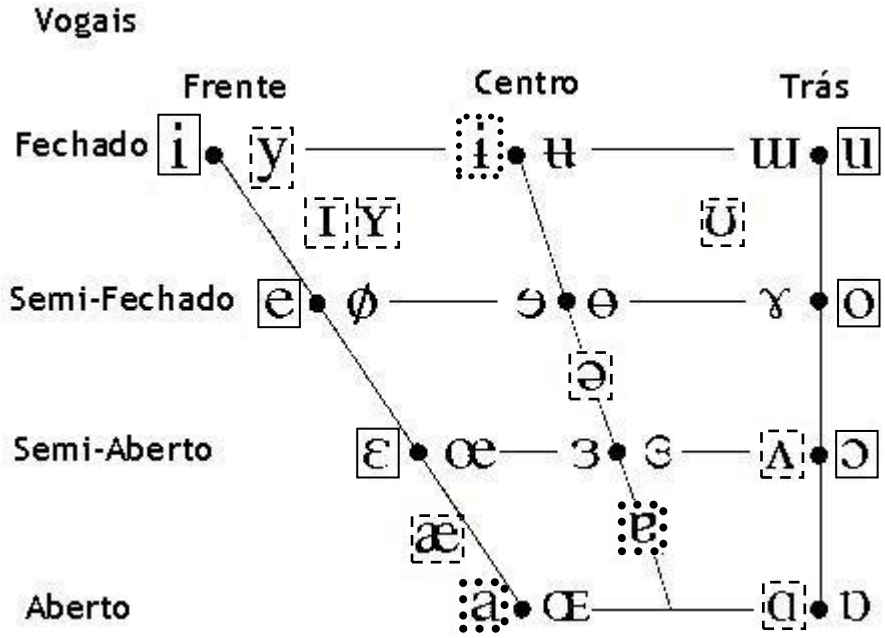


Figura 3.2 – Mapeamento das vogais orais em função do ponto de articulação e do grau de obstrução. Quando os símbolos aparecem em pares, o símbolo da direita representa a vogal arredondada. Adaptado de [8].

3.3 - Formantes típicas dos fonemas

Esta secção tem por base o estudo dos perfis do tracto vocal para fonemas do inglês americano, realizado por Lindbom and Sandeberg em 1971 e mencionado em [7], bem como as formas de onda típicas e os espectros dos mesmos. Apesar de se basear num estudo sobre as formantes do inglês americano grande parte dos perfis dos tractos vocais são válidos também para o português europeu, pois a maioria das consoantes mencionadas existem em ambas as línguas. Toda a informação respeitante a uma classe de fonemas e não a um fonema específico também é válida para as duas línguas.

Existem 13 vogais no inglês americano, sendo que uma delas é uma "vogal degenerada", Na figura 3.1 e na figura 3.3 essa "vogal degenerada", também conhecida como vogal *schwa*, está representada como /x/. A vogal *schwa* surge quando o orador pronuncia, por vezes, uma das restantes 12 vogais demasiado depressa e os articuladores não têm tempo para alcançarem a sua posição de destino, ficando numa posição intermédia em que o tracto vocal se assemelha a um tubo uniforme. A vogal *schwa* tem como características ser mais curta em duração e ter uma amplitude de sinal inferior à das outras vogais.

3.3.1 – Formantes das vogais

As vogais distinguem-se dos restantes fonemas por normalmente terem uma duração mais longa, entre 40 e 400 milissegundos, por serem todas vozeadas e por possuírem uma maior amplitude de sinal que os outros fonemas. Essa maior amplitude tem como origem o facto da obstrução do tracto vocal ser menos acentuada nas vogais. Essa obstrução, que é sempre determinada pela posição da língua, é a principal característica diferenciadora entre as vogais. Pode ser à frente, no meio e atrás e pode ser mais ou menos pronunciada. A distribuição das 13 vogais presentes no inglês americano segundo esses dois critérios pode ser visualizada na figura 3.3. A variação da área de corte transversal determina os formantes de uma vogal.

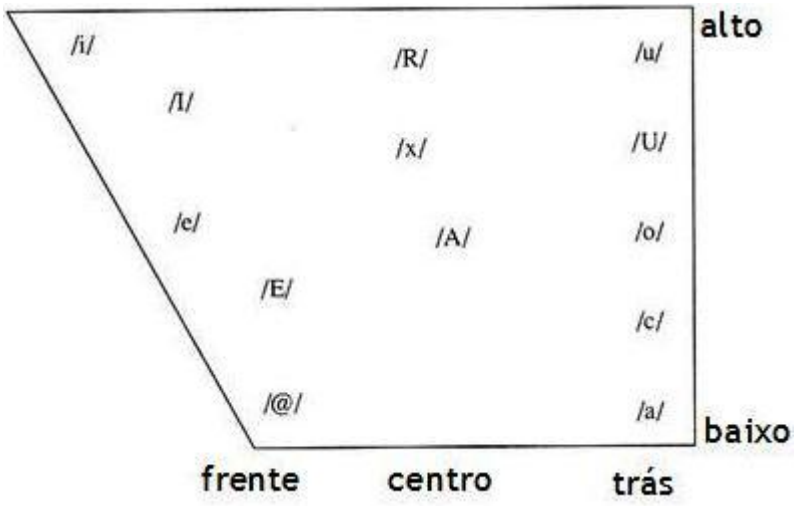


Figura 3.3 – Diagrama que mostra a localização e o grau de obstrução provocado pela língua para as diferentes vogais do inglês americano. Adaptado de [7].

A configuração do tracto vocal, a resposta no domínio dos tempos e a resposta no domínio das frequências durante a produção das 12 principais vogais do inglês americano podem ser comparadas, respectivamente, nas colunas (a), (b) e (c) da figura 3.4. O funcionamento cíclico das pregas vocais, que excita o tracto vocal durante o vozeamento, está presente no gráfico do domínio dos tempos de todas as 12 vogais, pois todas elas representam ondas quase periódicas. Observando os gráficos da coluna (b) também é possível verificar que a estrutura ressonante do tracto vocal varia com a localização e o grau de estreitamento do tracto vocal. Essa variação também confirma-se com os gráficos da coluna (c), que demonstram que a localização das frequências formantes e a sua largura de banda se alteram conforme a configuração da estrutura ressonante.

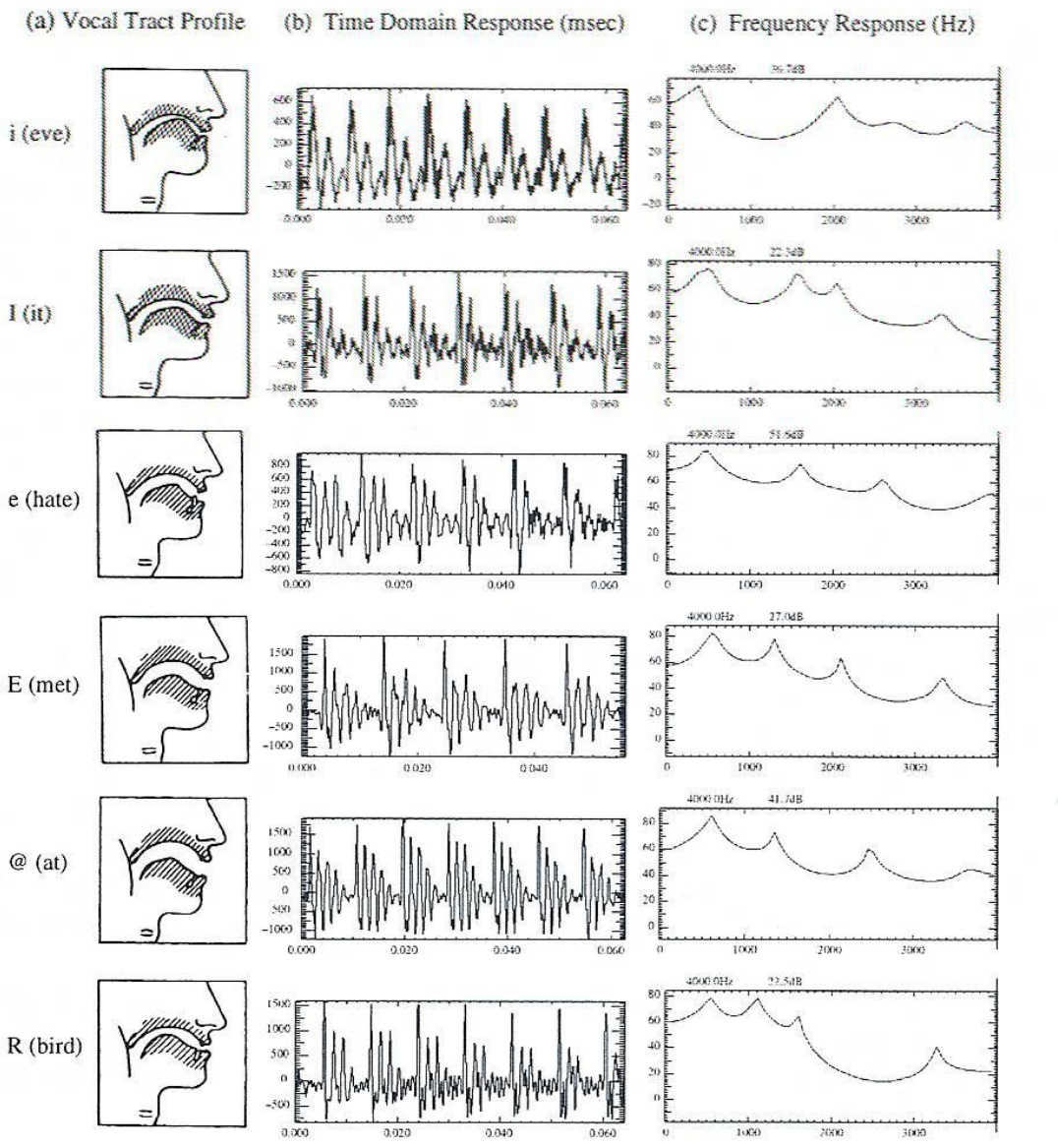


Figura 3.4 – Representa para as 12 principais vogais do inglês americano, um esquema da localização dos articuladores, na coluna (a), um gráfico com a resposta no domínio dos tempos, na coluna (b), e um gráfico com a resposta no domínio das frequências, na coluna (c) [7].

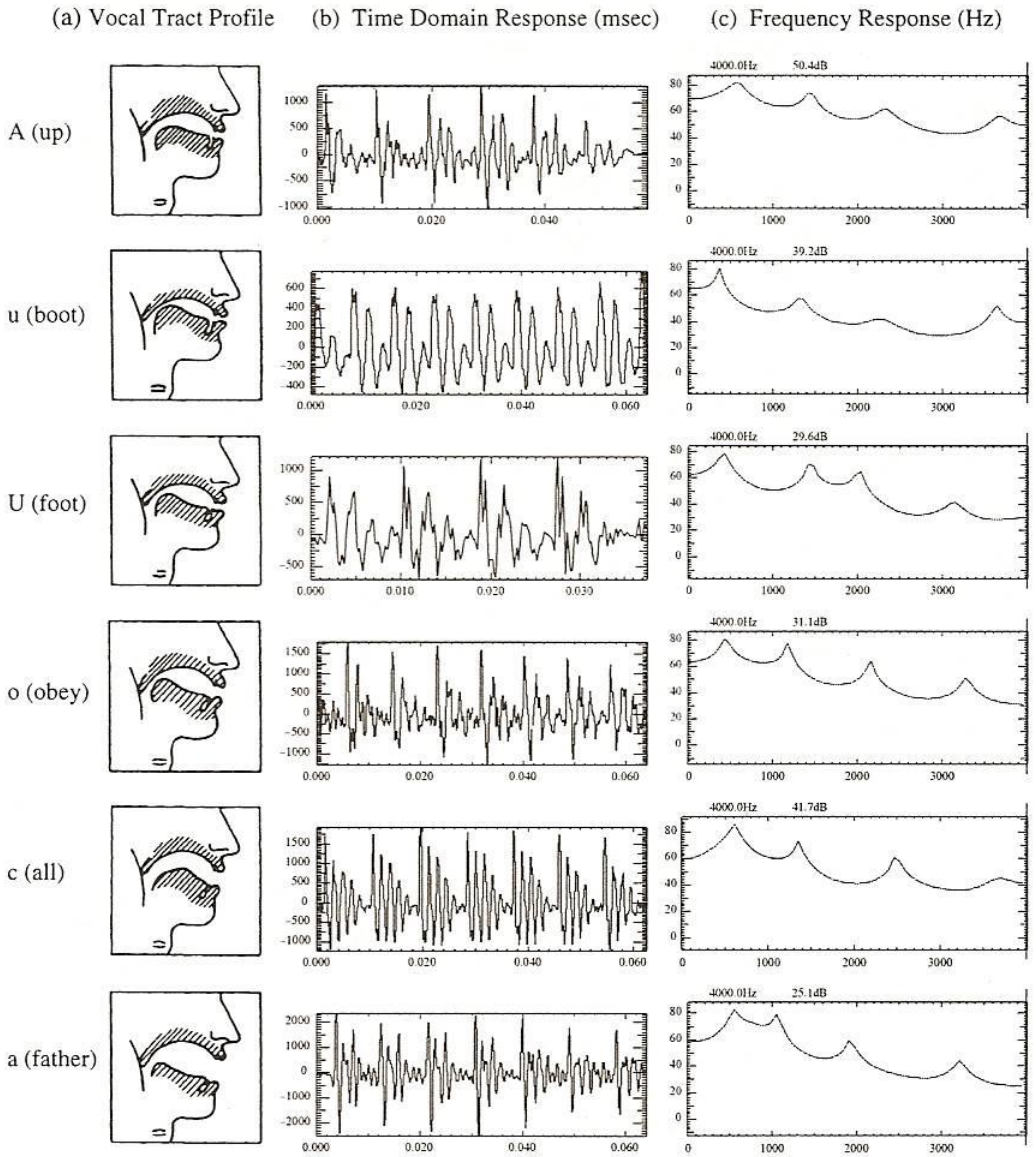


Figura 3.4 – (continuação)

Os três principais factores para a alteração das frequências formantes das vogais são o comprimento total do tracto vocal e a localização e grau da obstrução do mesmo. A localização dessas frequências formantes, especialmente das três primeiras, é normalmente suficiente para identificar as vogais. Quanto maior for o comprimento total do tracto vocal menor é a média das frequências das formantes da vogal em questão. Por essa razão, as crianças têm, em média, as frequências das formantes mais altas do que oradores adultos e os adultos do sexo masculino têm essas frequências mais baixas do que de oradores adultos do sexo feminino. Esta correlação entre o comprimento do tracto vocal e a localização das frequências das formantes e do espaçamento entre elas é menos notório para as primeiras duas formantes, pois estas têm uma dependência mais forte com o local e o grau da obstrução ao longo do tracto vocal.

A frequência da primeira formante é mais baixa se o estreitamento for na metade da frente da cavidade oral e é tanto menor quanto maior for essa obstrução. Se a obstrução for

na cavidade faríngea a primeira formante tem uma frequência mais alta e quanto maior for essa obstrução mais elevada é a frequência. Por seu lado, a frequência da segunda formante tem a tendência a baixar se o estreitamento for provocado pela parte de trás da língua e a aumentar se for a parte da frente da língua e em ambos os casos essa tendência é intensificada pelo aumento do grau de obstrução. Outra conclusão que se retira é que quanto mais arredondados forem os lábios na produção da vogal, mais baixas são as frequências de todas as formantes.

Observando com atenção a coluna (a) da figura 3.4 repara-se que a sequência das vogais /i/, /I/, /e/, /E/ e /@/ é alcançada com a língua a obstruir a parte frontal da cavidade oral e que essa obstrução é progressivamente menos acentuada quando se passa o primeiro fonema da sequência para o último. Nota-se também que as duas vogais finais têm muito mais estreitamento na cavidade faríngea do que as primeiras. Não é portanto de estranhar que à medida que se avança na sequência, o espaçamento entre a primeira e a segunda formante, que no início é grande, seja cada vez mais pequena, com uma contínua diminuição da frequência da segunda formante e um aumento também contínuo da primeira. A vogal posterior /u/ é gerada com os lábios muito mais arredondados do que o fonema /o/ e tem a língua mais próxima do palato que a vogal /U/. Comparando as vogais posteriores /o/, /c/ e /a/ repara-se que o estreitamento faríngeo é progressivamente mais pronunciado e que as duas últimas têm os lábios muito menos arredondados que o fonema /o/. Quando existe mais do que uma alteração simultânea da posição dos articuladores, nem sempre é fácil identificar que consequências é que terá no gráfico da resposta no domínio das frequências, pois uma das alterações pode ter um efeito superior do que a outra no resultado final. Estas últimas observações confirmam isso mesmo, por esse motivo para confirmar estas últimas afirmações o melhor é extrair as conclusões da figura 3.5, que representa um resumo de um estudo realizado por Peterson e Barney em 1952 que estudou a localização das frequências das três primeiras formantes, bem como a amplitude relativa das mesmas para 10 das vogais do inglês americano num universo de 33 oradores masculinos. Apesar deste estudo não conter dados sobre o fonema /o/, as informações extraídas sobre as outras vogais vêm de encontro ao que foi afirmado anteriormente, ou seja, a redução contínua da obstrução no palato para pronunciar a sequência de vogais /i/, /I/, /E/ e /@/ faz com que o espaçamento entre as frequências da primeira e segunda formantes seja cada vez mais pequena, com um aumento progressivo da primeira formante e uma redução igualmente progressiva da segunda. Comparando as vogais posteriores /u/ e /U/ verifica-se que devido ao maior estreitamento de /u/ a frequência da sua segunda formante é mais baixa. O fonema /a/ tem uma maior obstrução da cavidade faríngea do que a vogal /c/ e, por isso, a frequência da sua primeira formante é mais elevada. Observando a amplitude relativa entre as diferentes vogais da figura 3.5 conclui-se que para todas as vogais a amplitude é sempre mais elevada na primeira formante e que a da segunda também é sempre superior à da terceira.

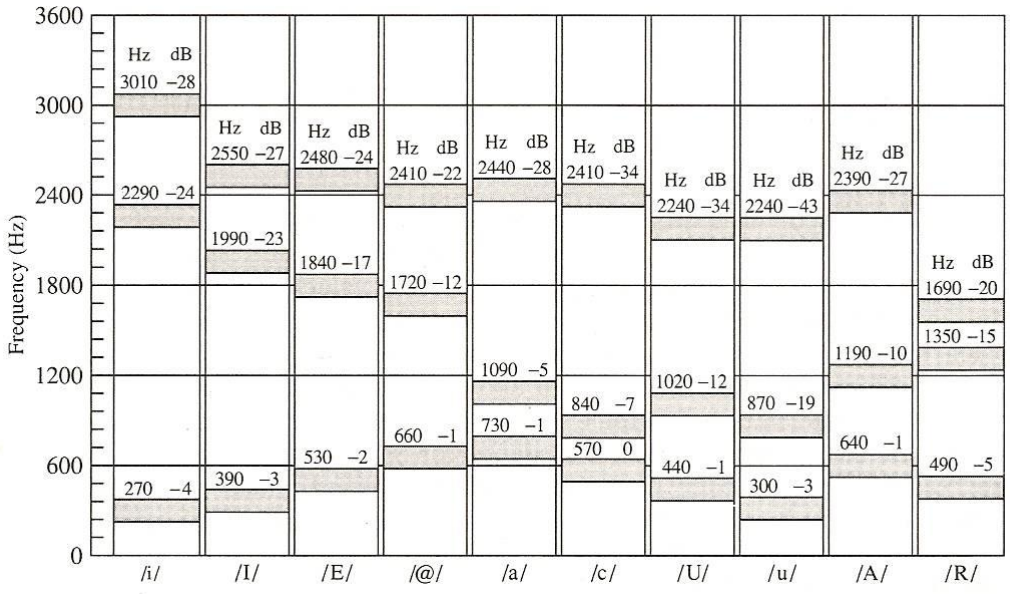
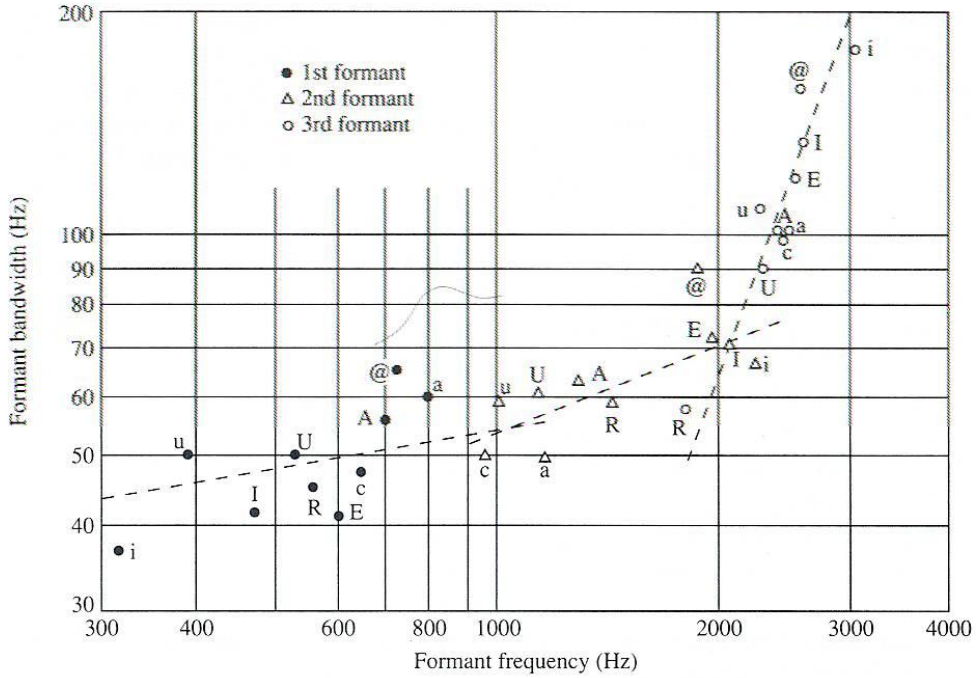


Figura 3.5 – Representa as frequências médias e a amplitude média relativa das três primeiras formantes de 10 das principais vogais do inglês americano [7].

A figura 3.6 relaciona as diferentes larguras de banda das primeiras três formantes das mesmas 10 vogais mencionadas na figura 3.5. Estas diferenças podem ser usadas para distinguir entre as diferentes vogais. Os resultados expostos na figura 3.6 foram retirados do estudo realizado por Dunn em 1961 e teve a participação de 20 oradores masculinos que repetiram cada vogal por duas vezes. A localização das frequências das formantes em relação à largura de banda é mostrada na parte superior da figura 3.6. Observando os resultados verifica-se a tendência da largura de banda aumentar com o aumento da frequência central da formante e que esse comportamento é bastante mais acentuado para a terceira formante. Conclui-se portanto que a comparação entre larguras de banda de diferentes vogais ajuda na diferenciação entre elas, não tendo no entanto a mesma precisão que a observação da localização das três primeiras formantes.



Vowels	F_1 Avg.	F_1 Extremes		F_2 Avg.	F_2 Extremes		F_3 Avg.	F_3 Extremes	
i	38	30	80	66	30	120	171	60	300
I	42	30	100	71	40	120	142	60	300
E	42	30	120	72	30	140	126	50	300
@	65	30	140	90	40	200	156	50	300
a	60	30	160	50	30	80	102	40	300
c	47	30	120	50	30	200	98	40	240
u	50	30	120	58	30	200	107	50	200
U	51	30	100	61	30	140	90	40	200
A	56	30	140	63	30	140	102	50	300
R	46	30	80	59	30	120	58	40	120
Avg.	49.7			64.0			115.2		

Figura 3.6 – Largura de banda em relação às frequências médias das três primeiras formantes de 10 das principais vogais do inglês americano [7].

3.3.2 – Formantes das fricativas não-vozeadas

Na figura 3.7 estão representadas características das quatro fricativas não-vozeadas e da consoante aspirada /h/. A consoante aspirada /h/ por vezes é considerada uma fricativa glotal não-vozeada. Tanto a fricativa não-vozeadas /T/ como a consoante aspirada /h/, representadas na figura 3.7, não existem na língua portuguesa. Como é possível identificar, a resposta no domínio dos tempos para qualquer uma delas é semelhante ao gráfico de um sinal de ruído e a sua amplitude é relativamente baixa comparando com a amplitude das vogais. Comparando as respostas no domínio das frequências é de notar que também estas apresentam características aproximadas de ruído branco, ou seja, respostas em frequência semelhantes em praticamente todo o espectro analisado. As fricativas não-vozeadas, ao

contrário das vogais, não apresentam uma maior predominância de conteúdo às baixas frequências, nem evidenciam um comportamento quase periódico no domínio dos tempos.

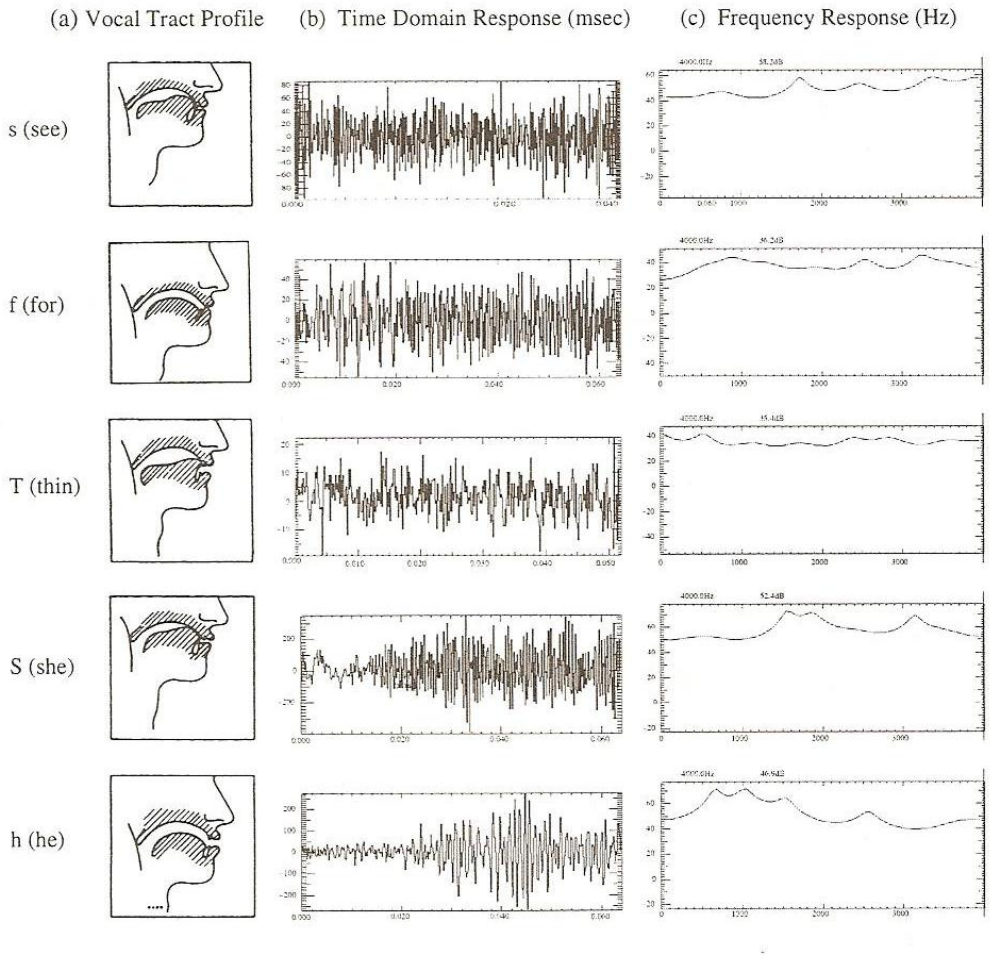


Figura 3.7 – Representação das quatro fricativas não-vozeadas e da consoante aspirada /h/ do inglês americano, um esquema da localização dos articuladores durante a sua produção, na coluna (a), um gráfico com a resposta no domínio dos tempos, na coluna (b), e um gráfico com a resposta no domínio das frequências, na coluna (c) [7].

3.3.3 – Formantes das fricativas vozeadas

As fricativas vozeadas /v/, /D/, /z/ e /Z/ diferem apenas no vozeamento das consoantes fricativas não-vozeadas /f/, /T/, /s/ e /S/, respectivamente. Analisando os gráficos da resposta temporal das fricativas vozeadas da figura 3.8 verifica-se que as fricativas vozeadas são fonemas com excitação mista, pois possuem características de vozeamento, como a quase periodicidade do sinal, e características de não-vozeamento, que as torna semelhantes a ruído. Estas características de fonemas de excitação mista afectam estas fricativas vozeadas de forma desigual, sendo que a fricativa vozeada labiodental /v/ tem características mais e de fonema vozeado, com a resposta temporal claramente quase periódica e uma maior componente espectral às baixas frequências, enquanto que as fricativas interdental /D/, alveolar /z/ e palatal /Z/ apresentam maiores semelhanças com fonemas não-vozeados, como uma maior parecença com ruído e um conteúdo distribuído mais uniformemente ao longo do espectro. A fricativa vozeada /D/ não tem correspondência no português europeu.

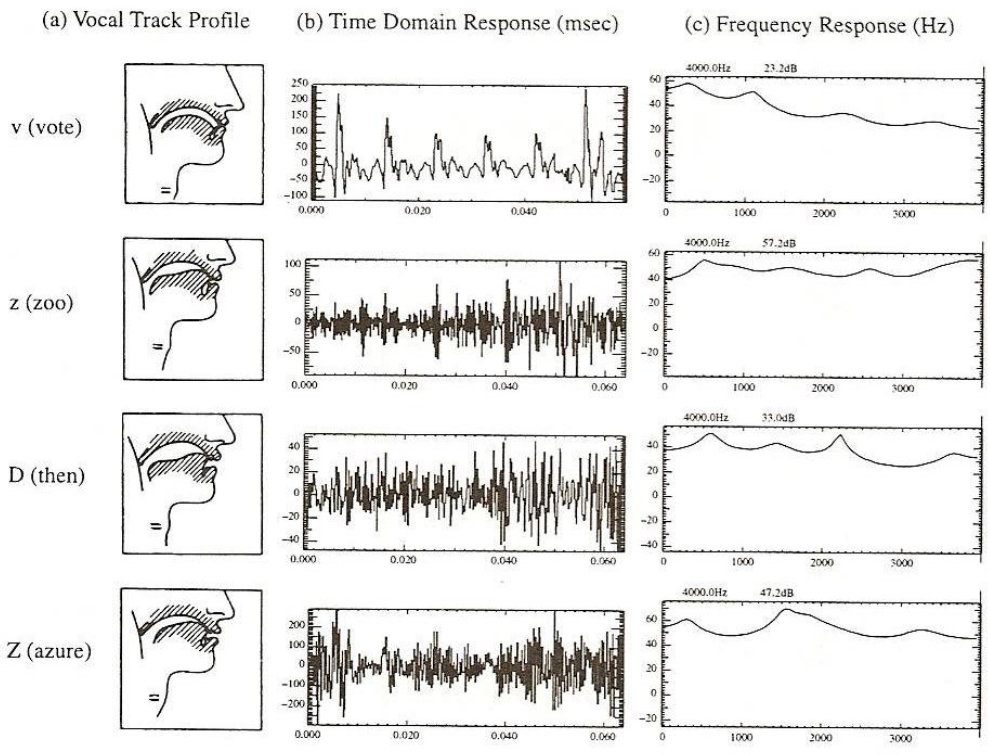


Figura 3.8 – Representação das quatro fricativas vozeadas do inglês americano, um esquema da localização dos articuladores durante a sua produção, na coluna (a), um gráfico com a resposta no domínio dos tempos, na coluna (b), e um gráfico com a resposta no domínio das frequências, na coluna (c) [7].

3.3.4 – Formantes das plosivas

As consoantes plosivas, tal como as fricativas, podem ser classificadas como não-vozeadas ou vozeadas e, tal como nas fricativas, a diferença entre as não-vozeadas, /p/, /t/ e /k/ e as vozeadas /b/, /d/ e /g/ é apenas o vozeamento, pois a posição dos articuladores durante a fase de obstrução total do tracto vocal e a movimentação dos mesmos articuladores durante a fase de libertação da pressão de ar é a mesma. Tanto as plosivas não-vozeadas, como as plosivas vozeadas existem também na língua portuguesa. A localização da obstrução total do tracto vocal nos fonemas plosivos pode ser bilabial (/p/ e /b/), alveolar (/t/ e /d/) ou velar (/k/ e /g/). A figura 3.9 demonstra algumas características das diferentes plosivas. Apesar de na maioria dos casos não ser perceptível um comportamento quase periódico das plosivas vozeadas devido aos movimentos cíclicos das pregas vocais, é fácil distinguir entre as plosivas não-vozeadas e as vozeadas. De facto, estas últimas, mesmo durante a fase de aumento da pressão de ar no ponto da obstrução, têm as suas pregas vocais a vibrar ciclicamente e apesar de não haver radiação de ar pela boca nem pelas narinas, existe uma radiação de energia reduzida pelas paredes da garganta. Esta radiação, apesar de pequena, é perceptível durante a fase de acumulação de pressão e identificável tanto no gráfico temporal através de uma perturbação não residual da onda, que no exemplo do fonema /b/ até se distingue a componente quase periódica da onda, como no gráfico do domínio das frequências que detecta algumas componente espectrais de baixa amplitude.

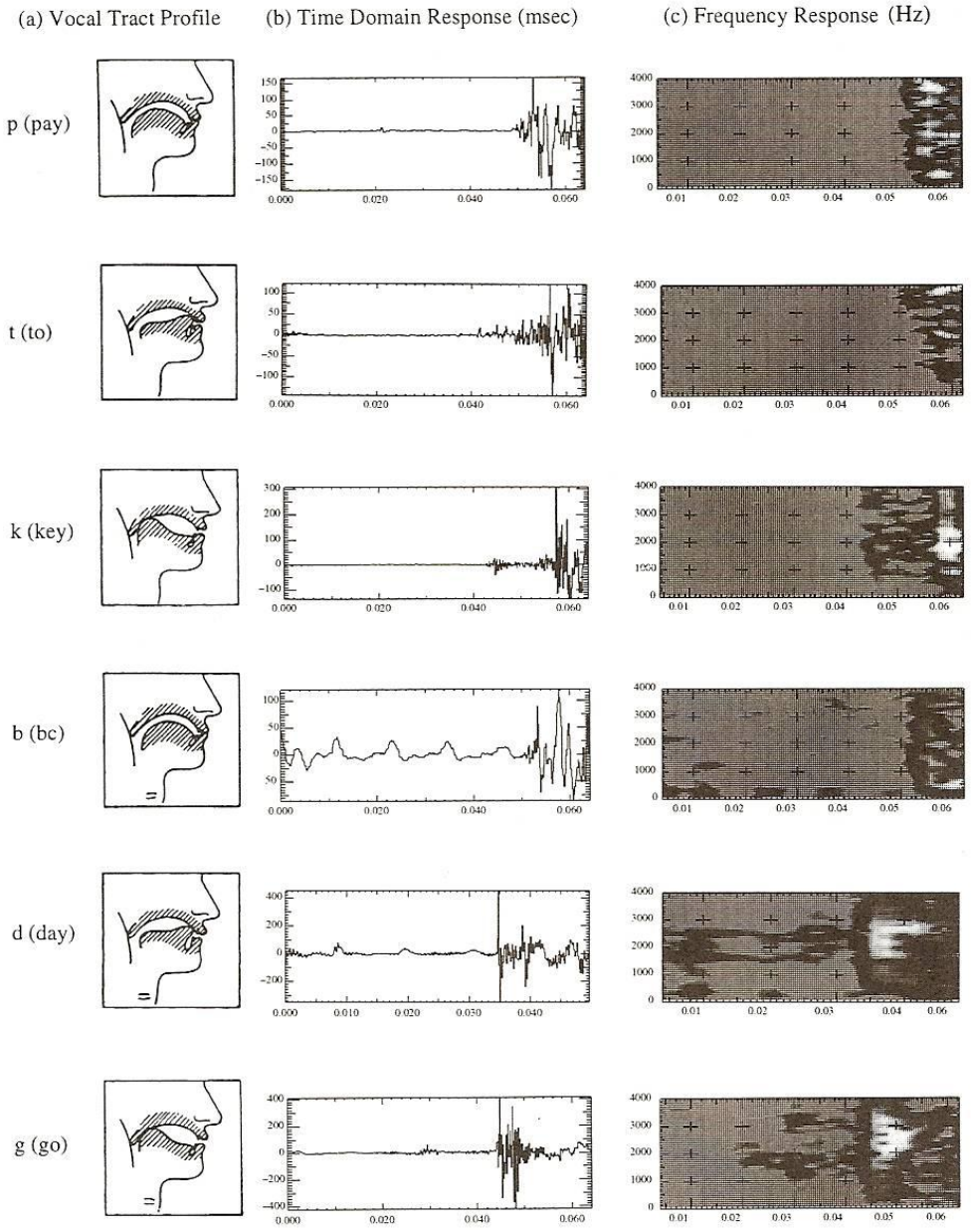


Figura 3.9 – Representação das três plosivas não-vozeadas e das três plosivas vozeadas do inglês americano, um esquema da localização dos articuladores durante a sua produção, na coluna (a), um gráfico com a resposta no domínio dos tempos, na coluna (b), e um gráfico com a resposta no domínio das frequências, na coluna (c) [7].

Apesar de por vezes até ser possível distinguir entre uma plosiva não-vozeada e uma plosiva vozeada, analisando tanto o gráfico temporal como o espectral é praticamente impossível fazer a distinção entre plosivas não-vozeadas ou entre plosivas vozeadas, devido à característica explosiva da fase de libertação do ar. Este tipo de excitação (plosiva) é semelhante a uma consoante fricativa, pois também neste caso o fluxo de ar ao passar pelo estreitamento do tracto vocal fica com características turbulentas, ou seja, semelhantes a ruído e isto é identificável tanto no gráfico temporal como no espectral através da

distribuição relativamente uniforme a todas as frequências. A identificação também é dificultada por causa da segunda fase de uma consoante plosiva ser de duração bastante reduzida. Para além das razões já mencionadas existem outras que acentuam ainda mais a dificuldade em identificar as consoantes plosivas, como o facto de as suas características variarem bastante consoante a sua posição na palavra ou frase. A maioria das plosivas não chega sequer a ser devidamente produzida quando ocorre no final de uma sílaba, isto acontece porque no final de uma sílaba a pressão pulmonar é inferior e reduz a pressão na obstrução do tracto vocal que é necessária para a produção correcta de uma plosiva. Por vezes, as plosivas também são alteradas quando ocorrem entre duas vogais. Por todos estes motivos, as plosivas são os fonemas mais difíceis de identificar.

Capítulo 4

Métodos de extracção de características do sinal de voz

4.1 – Introdução à extracção de características do sinal de voz

Os sinais de fala são sinais não-estacionários, ou seja, ao longo do tempo os seus principais atributos estatísticos e, em particular, a sua forma de onda, estão permanentemente a ser alterados. Estas modificações das propriedades dos sinais de fala são realizadas pelos diferentes articuladores envolvidos no processo fonatório. As ferramentas matemáticas utilizadas no processamento de sinais tipicamente requerem que estes permaneçam invariantes no tempo para que as suas características possam ser convenientemente analisadas. Na produção da fala estão envolvidos diferentes órgãos, ossos e músculos e, devido à inércia destes articuladores, não é possível alterar as suas posições de forma abrupta nem instantaneamente. Modificar o posicionamento dos diversos articuladores e consequentemente alterar a forma do tracto vocal é, portanto, um processo contínuo e com alterações relativamente suaves. Por esse motivo, se um sinal de fala for dividido em segmentos de duração suficientemente curta (aproximadamente 20ms), estes "novos" sinais de duração curta podem ser considerados quase estacionários, pois durante a sua duração os articuladores movem-se suficientemente pouco e devagar para que as características acústicas do "novo" sinal de voz possam ser consideradas praticamente invariantes no tempo.

Neste trabalho serão apresentados diversos métodos de extracção de características de sinais de voz e para todos eles é necessário efectuar primeiro a segmentação dos sinais de fala em segmentos de duração suficientemente curta. Para que os sinais de fala possam ser processados é necessário proceder à amostragem e à quantização do mesmo, neste trabalho todos os sinais de voz foram amostrados a 32.000 amostras por segundo e a quantização dos mesmos foi de 16 bit por amostra. A segmentação do sinal de fala é conseguida aplicando uma janela deslizante ao sinal de voz completo. Para todos os métodos e experiências realizadas neste trabalho foram utilizadas na segmentação janelas Hanning com sobreposição de 50% entre segmentos. A cada um destes segmentos, que se obtém multiplicando a sequência de voz com a janela de Hanning chama-se *frame*. A segmentação do sinal de voz é feita multiplicando a janela de Hanning com a sequência da fala, ou seja, se a janela de Hanning tiver N pontos a primeira *frame*, $fs(1;N)$, é constituída multiplicando um a um os

primeiros *N* pontos da sequência de voz com os pontos da janela. A *frame* criada tem, portanto, o mesmo número de pontos da janela utilizada na segmentação. A segunda *frame* é gerada deslizando a janela de Hanning sobre a sequência da fala e realizando o mesmo processo, ou seja, multiplicando a janela com a sequência de pontos do sinal de fala que se inicia na amostra $N/2+1$ até à amostra $N+N/2$, $fs(N/2+1; N+N/2)$, resultando numa sobreposição entre *frames* consecutivas de 50%. As *frames* seguintes são construídas da mesma forma até se atingir o fim da sequência de fala total. A definição de uma *frame* de um sinal de fala é:

$$f_S(n;m) \stackrel{\text{def}}{=} s(n)w(m-n) \tag{4.1}$$

sendo $s(n)$ o sinal de fala total e $w(n-m)$ a janela deslizante.

Há vários métodos que permitem extrair diversas características de um sinal de voz. Neste trabalho serão abordados os métodos: Linear Predictive Coding (LPC), Mel-Frequency *Cepstral* Coefficients (MFCC), Perceptual Linear Prediction (PLP) e o "método das sinusóides".

Para cada um dos métodos o objectivo é configurar um procedimento de análise e síntese de sinal e actuar nos parâmetros de análise/síntese de cada um deles para concluir sobre a sua influência no sinal de voz em termos de qualidade e inteligibilidade.

4.2 – Método LPC

4.2.1 – Introdução teórica ao método LPC

Um esquema de um modelo linear de tempo discreto razoavelmente geral utilizado para representar a produção da fala está apresentado na figura 4.1. Este modelo tem o nome de "modelo terminal analógico" e representa o processo de produção de um sinal de fala tendo como base as suas características de saída. Neste "modelo terminal analógico" o modelo do tracto vocal H(z) e o modelo de radiação R(z) são excitados por um sinal glotal de tempo discreto u(n) = u_{glotis}(n). Para produzir um sinal de fala vozeado é utilizada uma estimativa do *pitch* que serve como parâmetro de entrada a um gerador de trens de impulsos. Estes impulsos são modelados posteriormente por um modelo de pulso glotal antes de passarem para o modelo do tracto vocal. Na produção da fala não-vozeada a fonte de excitação é tipicamente um gerador de sinal aleatório. Este modelo é limitado na sua representação da produção da fala, pois não permite mais do que uma fonte de excitação. Os fonemas de excitação mista, como por exemplo as fricativas vozeadas, são deste modo mal caracterizados.

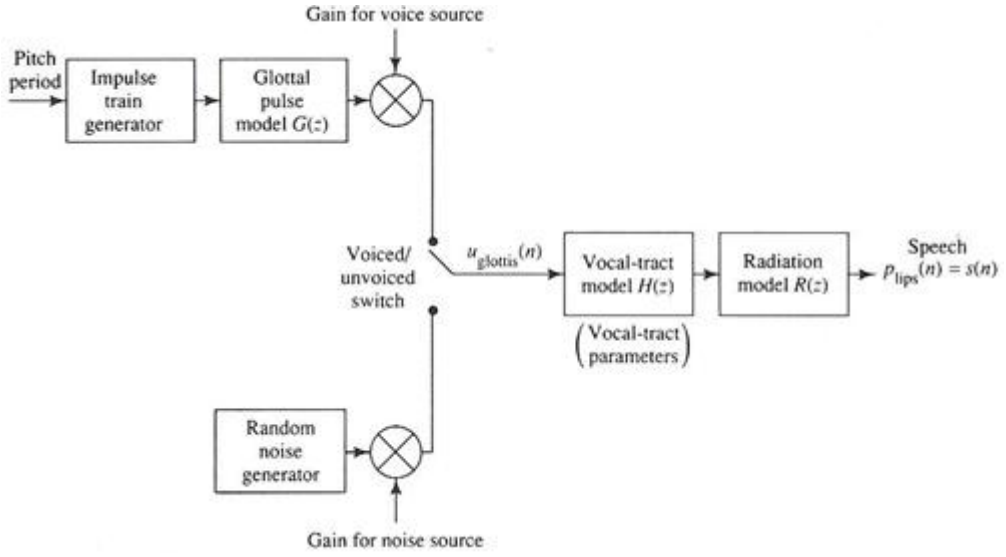


Figura 4.1– Modelo genérico de tempo discreto da produção de fala. Segundo Rabiner and Schafer (1978) [7].

A função de transferência do modelo do tracto vocal usada no modelo é:

$$H(z) = \frac{H_o}{\prod_{k=1}^{N}(1 - p_k z^{-1})} \tag{4.2}$$

H_o é um ganho geral e p_k é a localização complexa dos pólos no plano-z. Como já foi referido, este modelo tem limitações para alguns tipos de fonemas, mas mesmo assim é frequentemente utilizado para modelar todo o tipo de fonemas, pois há vários métodos analíticos poderosos que dependem da sua utilização. Cada par de pólos complexos conjugados localizados no plano Z corresponde aproximadamente a uma formante no espectro de H(z) e como $H(z)$ é um sistema estável, todos os pólos estão localizados no interior do círculo unitário do plano Z [9].

No caso da fala vozeada para algumas aplicações é utilizado para o filtro G(z), que pretende simular o comportamento da laringe, um modelo só com pólos semelhante ao usado como modelo do tracto vocal. Por vezes é usado o modelo com dois pólos, [7]

$$g(n) = [\alpha^n - \beta^n]u(n), \qquad \beta < \alpha < 1, \qquad \alpha \approx 1 \tag{4.3}$$

sendo $u(n)$ o degrau unitário. Este modelo consegue produzir um impulso com uma magnitude espectral parecida com os resultados obtidos empiricamente, mas tendo fase mínima não consegue produzir um impulso com uma fase de abertura mais prolongada do que a fase de fechamento [7], como já foi anteriormente ilustrado na figura 2.4. Um modelo só com pólos não consegue portanto reproduzir de forma fiel um ciclo vibratório das pregas vocais. Por esse motivo um modelo mais usual [7] é

$$g(n) = \begin{cases} \frac{1}{2}\left[1 - cos\left(\frac{\pi n}{P}\right)\right], & 0 \le n \le P \\ cos\left[\frac{\pi(n-P)}{2(K-P)}\right], & P \le n \le K, \\ 0\,, & outros\ valores \end{cases} \tag{4.4}$$

no qual P corresponde ao tempo de pico do impulso e K ao tempo de fechamento completo.

O modelo de radiação R(z) pode ser modelado por

$$R(z) = 1 - z_0 z^{-1}, \qquad z_0 \approx 1, \qquad z_0 < 1, \tag{4.5}$$

mas a equação pode ser reescrita na forma

$$R(z) = 1 - z_0 z^{-1} \approx \frac{1}{\sum_{k=0}^{K} z_0^k z^{-k}}, \tag{4.6}$$

com K teoricamente infinito, mas na prática finito, pois $z_0 < 1$.

A produção da fala pode ser modelada recorrendo apenas a filtros só com pólos. Na produção da fala não-vozeada o sinal de saída é constituído pela filtragem do sinal de excitação recorrendo a dois filtros lineares e independentes um do outro,

$$S(z) = E(z)H(z)R(z), \tag{4.7}$$

enquanto que a produção da fala vozeada recorre a três filtros lineares e independentes entre si,

$$S(z) = E(z)G(z)H(z)R(z). \tag{4.8}$$

Apesar de algumas limitações, os filtros só com pólos são frequentemente usados na reprodução da fala, pois estes permitem a utilização de uma técnica simples e bastante útil, a análise de predição linear.

Como foi referido anteriormente, o método mais correcto para modelar a produção de fala requer a utilização de um modelo de pólos e zeros, mas se for usado um modelo só com pólos os resultados obtidos são também razoavelmente bons. Para a percepção do conteúdo da fala, isto é, na perspectiva da inteligibilidade, as relações de fase entre as componentes da mesma não têm praticamente nenhuma importância [7]. Se não for necessário preservar as relações de fase entre componentes da fala durante a análise, é possível obter praticamente os mesmos resultados com o modelo só com pólos que com o modelo com pólos e zeros recorrendo à análise de predição linear (análise LP).

Qualquer sistema causal e racional do tipo

$$\Theta(z) = \Theta'_0 \frac{1+\sum_{i=1}^{L} b(i) z^{-i}}{1-\sum_{i=1}^{R} a(i) z^{-i}} \tag{4.9}$$

pode ser alterado para a forma

$$\Theta(z) = \Theta_0 \Theta_{min}(z) \Theta_{ap}(z) \tag{4.10}$$

onde $\Theta_{min}(z)$ tem fase mínima e $\Theta_{ap}(z)$ é um passa-tudo, ou seja, $\left|\Theta_{ap}(e^{j\omega})\right| = 1 \; \forall \, \omega$ e Θ_0 é uma constante relacionada com Θ'_0 e com as singularidades de $\Theta(z)$ [7].

A componente de fase mínima pode ser expressa como um sistema só com pólos,

$$\Theta_{min}(z) = \frac{1}{1-\sum_{i=1}^{I} a(i) z^{-i}} \tag{4.11}$$

com *I*, apesar de teoricamente infinito, na prática é um inteiro relativamente pequeno (e.g., 14).

O sistema com pólos e zeros inicial (4.9) pode ser então reescrito como

$$\Theta(z) = \Theta_0 \frac{1}{1-\sum_{i=1}^{I} a(i) z^{-i}} \Theta_{ap}(z) \tag{4.12}$$

e como $\left|\Theta_{ap}(\omega)\right| = 1 \; \forall \; \omega$, então $|\Theta(z)| = \Theta_0 |\Theta_{min}(\omega)|$. O que se perde ao passar de (4.9) para (4.12) é a informação sobre a fase, o que não é muito relevante para a percepção do conteúdo da fala.

Uma sequência de fala pode ser vista como

$$S(z) = \Theta(z) E(Z) = \Theta_0 \Theta_{min}(z) E'(z) \tag{4.13}$$

onde S(z) é a transformada z da sequência de fala de saída e E'(z) é a sequência de excitação de entrada e é definida como:

$$E'(z) \stackrel{\text{def}}{=} E(z) \Theta_{ap}(z) \tag{4.14}$$

No domínio temporal fica:

$$s(n) = \sum_{i=1}^{I} a(i) s(n-1) + \Theta_0 e'(n) \tag{4.15}$$

com excepção do termo $\Theta_0 e'(n)$, que é o sinal excitador com a fase modificada, a sequência de fala de saída pode ser predita através da combinação linear dos seus I valores passados, por este motivo este modelo também é conhecido como modelo autoregressivo (modelo AR). Na análise de predição linear resolvem-se as equações para determinar os parâmetros $a(i)$, também conhecidos como coeficientes LPC (Linear Predictive Coding), pois os coeficientes LPC passam a determinar a sequência de saída e como o número de coeficientes é reduzido (cerca de 14 por *frame*), relativamente ao número de amostras por *frame*, existe uma codificação eficiente da sequência, no caso de se usar pouca informação para representar $\Theta_0 e'(n)$.

4.2.2 – Introdução prática ao método LPC utilizado nas experiências

A análise LPC é efectuada pelo programa Matlab ***proclpc.m*** [10] e inicia-se com a pré-ênfase do sinal de fala. A pré-ênfase do sinal de fala consiste em aumentar a energia relativa do sinal de fala às altas frequências. Existem duas razões para se realizar a filtragem de pré-ênfase. Em primeiro lugar, porque a filtragem de pré-ênfase introduz um zero perto de z=1 que, em conjunto com o zero introduzido pelo modelo de radiação labial igualmente perto de z=1, cancelam os dois pólos na proximidade de z=1 da componente de fase mínima do modelo glotal [7]. Em segundo lugar, porque previne a instabilidade numérica, que pode ocorrer com o método da autocorrelação e também com o método da covariância [7]. Após a filtragem de pré-ênfase segmenta-se o sinal completo da fala em *frames* de 25 ms recorrendo a um janelamento rectangular e um incremento entre *frames* de 12,5 ms, ou seja, uma sobreposição de 50%. Para cada *frame* utiliza-se o método de Levinson (autocorrelação) para determinar os coeficientes LPC. Os coeficientes LPC correspondem aos parâmetros $a(i)$ que minimizam o erro, ou seja, que minimizam o erro quadrático médio da diferença entre a sequência de entrada do filtro (*frame* do sinal) e a sequência de saída do filtro (sequência

predita). Depois calcula-se o ganho de cada *frame* e, utilizando o método da autocorrelação, tenta-se calcular o valor do *pitch* da *frame*. Caso não seja detectado um valor para o *pitch* atribui-se o valor zero para identificar que a *frame* analisada corresponde a um segmento de fala não vozeado. Por fim, calcula-se o resíduo de cada *frame* do sinal, que é o vector do erro do filtro, ou seja, a diferença entre a *frame* de entrada do filtro e a sequência predita.

O diagrama de blocos da análise LPC explicada anteriormente está representado na figura 4.2.

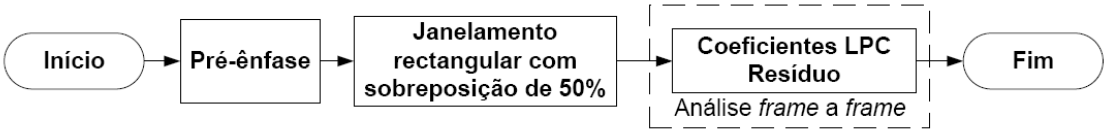


Figura 4.2– Diagrama de blocos da análise LPC.

O programa Matlab ***synlpc.m***, que realiza a síntese LPC, começa por sintetizar as *frames* a partir da matriz de coeficientes LPC e da matriz do resíduo. A criação do sinal de fala completo é conseguida, aplicando uma janela triangular a essas *frames* e sobrepondo as *frames* consecutivas em 50% de modo a eliminar a sobreposição utilizada durante a fase da análise LPC. Por fim procede-se à de-ênfase da sequência sintetizada final. A de-ênfase realiza a filtragem inversa da pré-ênfase, de modo a eliminar os efeitos da mesma e assim retomar a relação de energia entre as diferentes frequências aos valores do sinal de fala original.

Como a síntese do sinal de fala original é obtida também com o recurso ao resíduo resultante da análise LPC, ou seja, com o recurso à matriz dos erros de predição, o sinal resultante da síntese LPC é uma réplica exacta do sinal de fala original, na ausência de quantização ou modificação do resíduo ou dos coeficientes LPC.

O diagrama de blocos da síntese LPC explicada anteriormente está representado na figura 4.3.

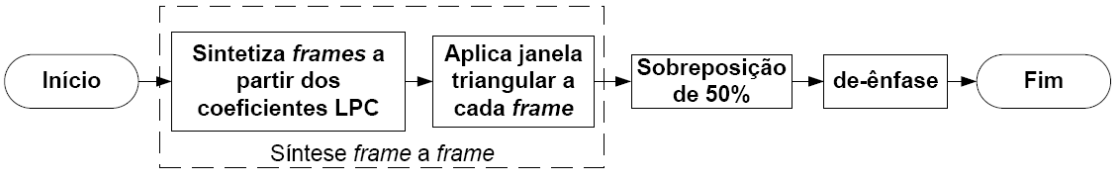


Figura 4.3– Diagrama de blocos da síntese LPC.

4.2.3 – Experiências baseadas no método LPC

4.2.3.1 - Quantização do resíduo

O método LPC descrito na secção 4.2.2 consegue obter, após a síntese, uma réplica exacta do sinal de fala original, mas para tal precisa para cada *frame*, para além dos valores dos coeficientes LPC calculados, da matriz dos erros de perdição (resíduo).

O objectivo desta primeira experiência é classificar a importância da quantização em termos de inteligibilidade e qualidade do sinal ressintetizado comparativamente com o sinal original, utilizando na ressíntese um resíduo com diferentes graus de quantização (2, 3 e 4 bits). A quantização é realizada *frame* a *frame* durante o processo de análise LPC e a ressíntese do sinal é feita como indicado na figura 4.3.

O código Matlab do método LPC de quantização do resíduo pode ser consultado no ANEXO A.

4.2.3.2 – Excitação por ruído branco

Na segunda experiência baseada no método LPC cada *frame* ressintetizada é substituída por uma outra de igual média e desvio padrão, mas de um sinal aleatório. Tal como na experiência do método LPC de quantização do resíduo, o objectivo é verificar se as diferentes ressínteses são inteligíveis e, caso sejam, classificar a qualidade das mesmas relativamente ao sinal original.

O código Matlab do método LPC de sinal aleatório pode ser consultado no ANEXO B e o processo de síntese é realizado como demonstrado na figura 4.3.

4.2.3.3 – Importância do *pitch* na percepção

A quarta e quinta experiência servem para tentar identificar se é a fala vozeada ou a fala não-vozeada que tem mais impacto em termos de percepção no sinal de voz. Tal como nas experiências anteriormente descritas este estudo foca-se na inteligibilidade e na qualidade do sinal ressintetizado. A quarta e quinta experiência têm o mesmo código Matlab base, em que, utilizando a função da autocorrelação, se tenta identificar a presença de *pitch* no sinal de fala e se faz a separação entre *frames* de fala vozeada e de fala não-vozeada. A quarta experiência mantém as *frames* do resíduo em que se identificou o *pitch* inalteradas, substituindo as restantes por *frames* de igual média e desvio padrão, mas de um sinal aleatório, como no caso da segunda experiência. A quinta experiência por seu lado mantém as *frames* em que não se identificou a presença de *pitch* no sinal inalteradas, aplicando às *frames* vozeadas uma máscara, que mantém por cada ciclo vibratório das pregas vocais os 25% de maior amplitude e elimina os restantes 75%.

O código Matlab para a quarta experiência do método LPC pode ser visualizado no ANEXO C e o código da quinta no ANEXO D.

4.3 – Método MFCC

4.3.1 – Introdução teórica ao método MFCC

Um sinal de fala corresponde a uma sequência de excitação convolvida com a resposta impulsional do sistema vocal. Por vezes é conveniente separar as duas componentes, para que seja possível manipular apenas uma das partes, mas este processo não é trivial. A análise *cepstral* foi desenvolvida para resolver este problema.

A análise *cepstral* representa (idealmente) uma transformação do sinal de fala com duas propriedades importantes:

-as representações das componentes do sinal estarão separadas no *cepstrum*,

-as representações das componentes de sinal no *cepstrum* vão corresponder a uma combinação linear [7].

Depois do sinal de fala estar representado no *cepstrum* é possível seleccionar determinadas componentes do *cepstrum*, aplicando um filtro linear para remover as partes indesejadas. Às componentes que não foram eliminadas aplica-se uma transformação inversa à da produção do *cepstrum*. Todo este processo respeita o princípio da sobreposição, que no caso da convolução é:

$$H[x(n)] = H[x_1(n) * x_2(n)] = H[x_1(n)] * H[x_2(n)], \tag{4.16}$$

em que "$H[\,]$" representa o sistema homomórfico e o símbolo "$*$" é o sinal de convolução.

Aos sistemas que obedecem ao princípio da sobreposição para a convolução chamam-se de sistemas homomórficos.

Qualquer sistema homomórfico pode ser representado por três sistemas homomórficos, como ilustrado na figura 4.4.

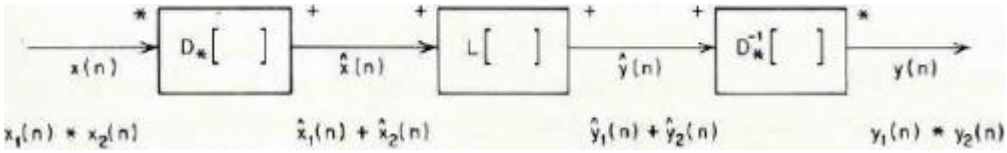


Figura 4.4– Forma canónica para um sistema para desconvolução homomórfica [9].

Nesta figura, o primeiro sistema recebe na entrada sinais combinados por convolução e transforma-os numa soma de termos. O segundo sistema é um sistema linear, que obedece ao princípio da sobreposição e o terceiro sistema realiza o inverso do primeiro sistema. Ao primeiro sistema chama-se sistema característico da desconvolução e na forma canónica é fixo, tal como o terceiro sistema, ou seja, só o sistema linear é que difere entre sistemas homomórficos [9]. O sistema característico da desconvolução homomórfica transforma a convolução na entrada numa soma na saída. Se o sinal de entrada de um sistema homomórfico for a convolução

$$x(n) = x_1(n) * x_2(n) \tag{4.17}$$

então aplicando a transformada-z, a entrada passa a ser a multiplicação das respectivas transformadas-z

$$X(Z) = X_1(Z) \cdot X_2(Z). \tag{4.18}$$

A passagem para a soma final é conseguida com o recurso a uma função logarítmica, pois o logaritmo de um produto é igual à soma dos respectivos logaritmos. A soma pode então ser manipulada por um sistema linear. O inverso do sistema característico da desconvolução homomórfica utiliza a função exponencial e depois a transformada-z inversa para passar a soma para um produto e por fim para uma convolução.

A um *cepstrum* construído com base num processo homomórfico dá-se o nome de *cepstrum complexo*, que difere do *cepstrum real* pois ao utilizar um logaritmo complexo não descarta as relações de fase. A maioria das análises *cepstrais* no entanto utiliza um *cepstrum real*, pois é bastante mais simples e a perda da informação sobre a fase não é relevante para muitas aplicações.

O *MEL-Cepstrum* é uma variação do *cepstrum* normal, que tira proveito da percepção auditiva humana. A verdadeira frequência de um som e a percepção que um humano tem dessa frequência não têm uma correspondência linear. A frequência "percebida" pelos humanos, também conhecida como *pitch*, tem como unidade de medição o *MEL*. Na figura 4.5 está representada a escala mel, criada por Stevens e Volkman em 1940 e que relaciona o *pitch* com a frequência real. A relação entre as duas é praticamente linear até aos 1000 Hz, ficando depois logarítmica para frequências superiores a esse valor. Também se descobriu que a percepção que se tem de uma determinada frequência é influenciada pela energia de uma banda crítica de frequências em torno dessa mesma frequência (Schoroeder, 1977; Allen, 1985; O'Shaughnessy, 1987) e que a largura de banda dessas bandas críticas varia com a frequência. Com base nestas descobertas foram desenvolvidos novos métodos de análise de sinais de fala.

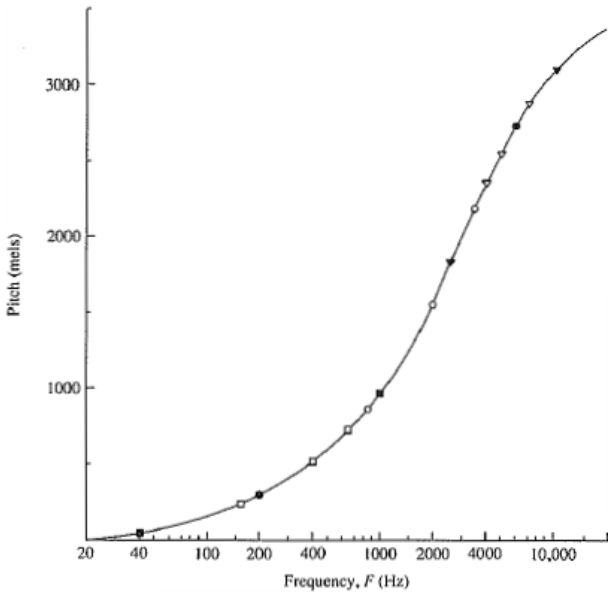


Figura 4.5– A escala mel. Segundo Stevens e Volkman (1940) [7].

4.3.2 – Introdução prática ao método MFCC utilizado nas experiências

A análise MFCC é realizada pelo programa Matlab ***Melfcc.m*** [11] e tem como único parâmetro de entrada obrigatório os dados do ficheiro de voz **(samples)**. A análise MFCC começa com a pré-ênfase do sinal de fala **(preemph ⟹ 0.97)**. Após esta filtragem segmenta-se o sinal completo da fala em *frames* de 25 ms **(wintime ⟹ 0,025),** recorrendo a um janelamento Hamming e um incremento entre *frames* de 12,5 ms **(hoptime ⟹ 0,0125)**, ou seja, uma sobreposição de 50%. Calcula-se depois o espectro de potência com FFT*s* de 512 pontos. Posteriormente é feita uma análise espectral de uma frequência mínima de 0Hz **(minfreq ⟹ 0)** até à frequência máxima de metade da frequência de amostragem **(maxfreq ⟹ sr/2)** e são usados 80 filtros Mel **(fbtype ⟹ 'mel')** para a sua realização **(nbands ⟹ 80)**.

As larguras de banda desses filtros relativamente aos valores de referência permanecem iguais **(bwidth ⟹ 1.0)**. Como a *flag* sumpower tem o valor 1 (**sumpower ⟹1)**, o mapeamento das potências do espectro para a escala Mel é feito multiplicando as potências do espectro com uma matriz de pesos, caso contrário (se a *flag* sumpower tiver o valor 0) o mapeamento das potências do espectro para a escala Mel seria feito elevando ao quadrado a multiplicação da raiz quadrada das potências do espectro com uma matriz de pesos. Por fim são extraídos os coeficientes MFCC **(numcep ⟹ 13)**, aplicando a função logaritmo às diferentes *frames* do espectro e depois retirando a DCT (transformada discreta de cosseno), que neste caso é ortogonal e de norma unitária **(dcttype ⟹ 2)**.

O diagrama de blocos da análise MFCC explicada anteriormente está representado na figura 4.6.

O único parâmetro de saída obrigatório é a matriz **cepstra**, em que as colunas da matriz representam as *frames* analisadas e as linhas da matriz representam os coeficientes calculados para cada *frame*. Para além do parâmetro de saída obrigatório é possível obter a matriz **aspectrum** e a matriz **pspectrum**, que representam resultados intermédios do programa ***Melfcc.m***, a primeira é o resultado após a conversão para a escala MEL e a segunda o resultado após o espectro de potência FFT.

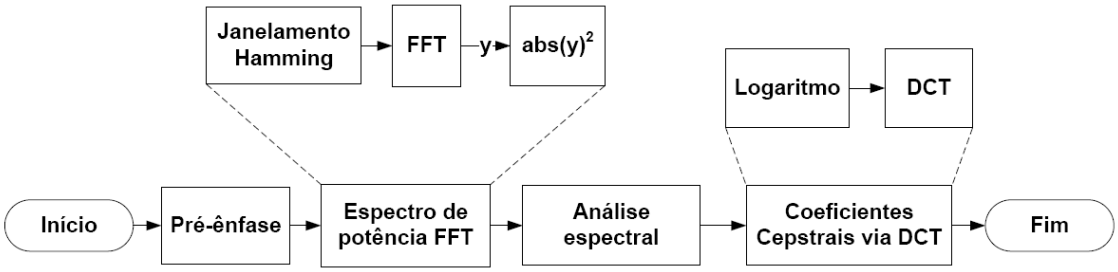


Figura 4.6– Diagrama de blocos da análise MFCC.

A ressíntese do sinal áudio a partir da matriz **cepstra** após a análise MFCC é feita pelo programa Matlab ***invmelfcc.m*** [11]. Os argumentos da função ***invmelfcc*** são a matriz do *cepstrum* e os argumentos utilizados durante a análise MFCC com os mesmos valores, para que o resultado final da ressíntese seja o mais próximo possível do ficheiro áudio original.

Inicialmente converte-se o *cepstrum* em espectro, multiplicando o *cepstrum* com a IDCT (transformada discreta de cosseno inversa) e usando o resultado como argumento da função exponencial. A fase seguinte consiste em tentar eliminar os efeitos da realização da análise espectral, para tal multiplica-se o resultado com o inverso da matriz de pesos, invertendo assim o mapeamento para a escala Mel. Os valores desta matriz são utilizados para modelar um espectrograma de ruído branco. Cada *frame* do espectrograma é convertida numa porção da onda do sinal áudio usando uma IFFT e depois é acrescentada ao vector final sobrepondo os segmentos. Por fim realiza-se a de-ênfase do sinal ressintetizado.

O sinal ressintetizado é um sinal sintético, pois, como foi obtido utilizando excitação de ruído branco, comparando-o com o sinal original perde-se a informação de fase e da frequência fundamental durante o processo de ressíntese.

O diagrama de blocos da síntese MFCC explicada anteriormente está representado na figura 4.7.

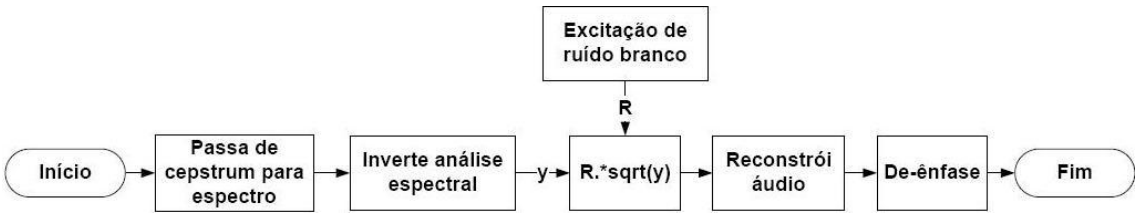


Figura 4.7– Diagrama de blocos da síntese MFCC.

4.3.3 – Experiências baseadas no método MFCC

Os sinais ressintetizados utilizando o método MFCC, devido à perda da informação de fase e da frequência fundamental, são consideravelmente diferentes dos sinais originais. O objectivo principal desta experiência é, portanto, tentar perceber em que medida é que esta perda de informação influencia a inteligibilidade e a qualidade dos sinais e também fazer uma avaliação às alterações ocorridas nos diferentes tipos de fonemas.

O código Matlab para o método MFCC está reproduzido no ANEXO E.

4.3.3.1 – Influência do número de coeficientes na percepção

Para todos os sinais de fala analisados foram produzidas duas ressínteses baseadas no método MFCC em que o único parâmetro alterado foi o número de coeficientes utilizados (6 e 13 coeficientes).

A intenção desta experiência é averiguar as diferenças na percepção dos dois tipos de ressínteses tanto ao nível da inteligibilidade como da qualidade dos sinais de voz.

4.4 – Método PLP

4.4.1 – Introdução teórica ao método PLP

A análise de predição linear perceptual (PLP) foi desenvolvida por Hermansky em 1989. Esta técnica utiliza três conceitos da psicoacústica, ou seja, do estudo subjectivo das características sonoras, para criar um espectro auditivo. A primeira é a resolução espectral das bandas críticas, a segunda é uma curva de igualização de sonoridade e a terceira é a lei da potência intensidade-sonoridade. O espectro auditivo é depois aproximado por um modelo autorregressivo só com pólos [12].

A análise PLP proposta por Hermansky consiste em segmentar o sinal de fala usando um janelamento de Hamming, com janelas de aproximadamente 20ms. A transição para o domínio das frequências é feita usando a transformada discreta de Fourier, normalmente a FFT de 256 pontos para uma frequência de amostragem de 10kHz. O espectro de potência de termo curto é obtido somando o quadrado das componentes real e imaginária do espectro de sinal de termo curto.

$$P(\omega) = Re[S(\omega)]^2 + Im[S(\omega)]^2 \tag{4.19}$$

O espectro P(w) é distorcido ao longo do seu eixo de frequência *w* para o eixo de frequências Bark *Ω*, de acordo com a relação

$$\Omega(\omega) = 6ln\left\{\frac{\omega}{1200\pi} + \left[\left(\frac{\omega}{1200\pi}\right)^2 + 1\right]^{0.5}\right\} \tag{4.20}$$

onde ω é a frequência angular em rad/s. Convolui-se depois o espectro resultante com um espectro de potência de uma curva de mascaramento de banda crítica *Ψ(Ω)* simulado. A curva de mascaramento tem a forma

$$\Psi(\Omega) = \begin{cases} 0 & \Omega < -1.3, \\ 10^{2.5(\Omega+0.5)} & -1.3 \le \Omega \le -0.5, \\ 1 & -0.5 < \Omega < 0.5, \\ 10^{-1.0(\Omega-0.5)} & 0.5 \le \Omega \le 2.5, \\ 0 & \Omega > 2.5. \end{cases} \tag{4.21}$$

A convolução dos dois espectros permite uma reamostragem a intervalos de aproximadamente um Bark. O sinal reamostrado é depois pré-enfatizado pela curva de igualização de sonoridade, que simula a sensibilidade auditiva humana para valores de aproximadamente 40 dB,

$$E(\omega) = [\omega^2 + 56{,}8 \cdot 10^6)\omega^4]/[(\omega^2 + 6{,}3 \cdot 10^6)^2 \cdot (\omega^2 + 0{,}38 \cdot 10^9)] \tag{4.22}$$

resultando no sinal

$$\Xi[\Omega(\omega)] = E(\omega)\Theta[\Omega(\omega)]. \tag{4.23}$$

Por fim utiliza-se uma compressão de raiz cúbica para simular a relação não-linear entre a intensidade de um determinado som e a percepção da *"loudness"* e do mesmo pelo ouvido humano.

$$\Phi(\Omega) = \sqrt[3]{\Xi(\Omega)} \tag{4.24}$$

O sinal Φ(Ω) é aproximado pelo espectro de um modelo autorregressivo só com pólos usando o método da autocorrelação e no final é possível extrair determinadas características, como é o caso dos coeficientes do modelo autorregressivo ou coeficientes *cepstrais*.

4.4.2 - Introdução prática ao método PLP utilizado nas experiências

A análise PLP, tal como a MFCC, é realizada pelo programa Matlab ***Melfcc.m*** [11] e tem como único parâmetro de entrada obrigatório os dados do ficheiro de voz **(samples)**. O processo de análise desenrola-se como no método MFCC até à conclusão da análise espectral, com a diferença que no método PLP são usados filtros Bark **(fbtype ⟹ 'bark')**.

Após a conclusão da análise de banda crítica a análise PLP diverge da análise MFCC. Como a *flag* usecmp tem valor 1 (**usecmp ⟹1)**, é efectuada uma igualização de sonoridade e uma compressão de raiz cúbica. Posteriormente, são extraídos coeficientes LPC por cada *frame* do espectro utilizando um modelo autorregressivo de ordem 12 **(modelorder ⟹ 12)** e, por fim, com uma técnica específica, esses coeficientes LPC são convertidos em coeficientes *cepstrais* **(numcep ⟹ 13)** [7]. Todos os restantes parâmetros utilizados na análise PLP têm os mesmos valores que os usados na análise MFCC.

O único parâmetro de saída obrigatório é a matriz **cepstra**, em que as colunas representam as *frames* analisadas e as linhas representam os coeficientes calculados. Para além do parâmetro de saída obrigatório é possível obter a matriz **aspectrum** e a matriz **pspectrum**, que representam resultados intermédios do programa ***Melfcc.m***, a primeira é o resultado após a realização da igualização de sonoridade e a compressão de raiz cúbica e a segunda o resultado após o espectro de potência *FFT*.

O diagrama de blocos da análise PLP explicada anteriormente está representado na figura 4.8.

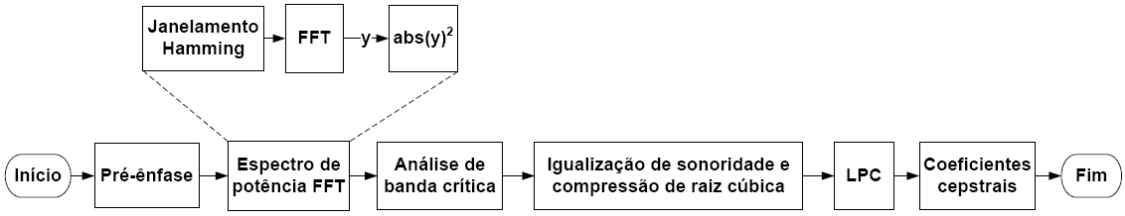


Figura 4.8– Diagrama de blocos da análise PLP.

A ressíntese do sinal áudio a partir da matriz **cepstra** após a análise PLP é feita pelo programa Matlab ***invmelfcc.m*** [11]. Os argumentos da função ***invmelfcc*** são a matriz do *cepstrum* e os argumentos utilizados durante a análise PLP com os mesmos valores, para que o resultado final da ressíntese seja o mais próximo possível do sinal de fala original.

Inicialmente converte-se o *cepstrum* em espectro, multiplicando o *cepstrum* com a IDCT (transformada discreta de cosseno inversa) e usando o resultado como argumento da função exponencial. Depois tenta-se inverter os efeitos da igualização de sonoridade e da compressão de raiz cúbica e também da realização da análise de banda crítica. Os valores desta matriz são utilizados para modelar um espectrograma de ruído branco. Cada *frame* do espectrograma é convertida numa porção da onda do sinal áudio usando uma IFFT e depois é acrescentada ao vector final sobrepondo os segmentos. Por fim realiza-se a de-ênfase do sinal ressintetizado.

Tal como no método MFCC o sinal ressintetizado após a síntese PLP é um sinal sintético, pois foi obtido utilizando excitação de ruído branco. Em relação ao sinal de fala original a ressíntese perde a informação da fase e da frequência fundamental.

O diagrama de blocos da análise PLP explicada anteriormente está representado na figura 4.9.

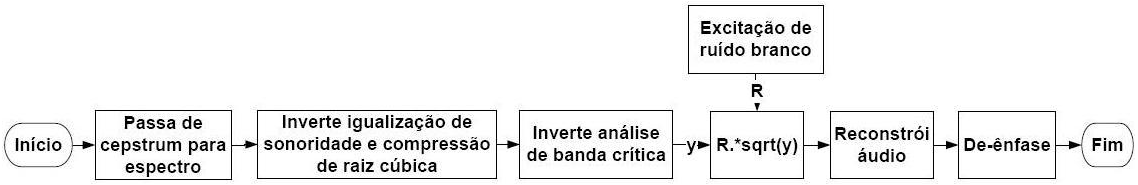


Figura 4.9– Diagrama de blocos da síntese PLP.

4.4.3 – Experiências baseadas no método PLP

As experiências baseadas no método PLP têm o mesmo objectivo que as baseadas no método MFCC, ou seja, tentar identificar qual as alterações que ocorrem na qualidade dos sinais, se a inteligibilidade dos mesmos é afectada e se há modificações específicas na percepção de determinados tipos de fonemas.

O código Matlab para o método PLP é o mesmo que para o método MFCC e pode ser visualizado no ANEXO E.

4.4.3.1 – Influência do número de coeficientes na percepção

Foram produzidas duas ressínteses distintas para cada um dos sinais de fala analisados baseadas no método PLP em que o único parâmetro alterado foi o número de coeficientes usados (6 e 13 coeficientes).

4.5 – "Método das sinusóides"

4.5.1 – Introdução teórica ao "método das sinusóides"

Este método pretende calcular a frequência, a fase e a magnitude de sinusóides quase estacionárias, com o objectivo de analisar, modificar e codificar sinais de áudio. Dada uma sinusóide discreta na forma:

$$x(n) = A \sin\left[\frac{2\pi}{N}(\ell + \Delta\ell)\, n + \phi\right] \tag{4.25}$$

sendo A a magnitude, ℓ a parte inteira e $\Delta\ell$ a parte fraccionária dos bins de frequência da DFT e ϕ a fase inicial do sinal depois de janelado por uma função real $h(n)$ de tamanho N e transformado para o domínio de frequências complexas usando um banco de filtros uniforme de N canais.

Este método realiza uma análise *frame* a *frame* e só utiliza para o cálculo dos parâmetros a informação resultante da transformação complexa da *frame* que está a ser analisada, ou seja, não utiliza informações das *frames* adjacentes.

Inicialmente o sinal é segmentado utilizando uma "janela de seno", que é a raiz quadrada da janela de Hanning. Com esta janela, o banco de filtros uniforme de N canais consegue uma reconstrução perfeita do sinal.

Este método utiliza uma transformada, ou banco de filtros, que consiste na ODFT e cuja expressão geral é a seguinte:

$$X_0(k) = \sum_{n=0}^{N-1} x(n)h(n)e^{-j\frac{2\pi}{N}\left(k+\frac{1}{2}\right)n} \tag{4.26}$$

A janela $h(n)$utilizada para segmentar o sinal de entrada é "janela de seno" (4.27), que é a raiz quadrada da janela de Hanning. Com esta janela, o banco de filtros ODFT consegue uma reconstrução perfeita do sinal.

$$h(n) = \sin\ \frac{\pi}{N}\left(n+\frac{1}{2}\right), \qquad 0 \le n \le N-1 \tag{4.27}$$

A resposta em frequência $H(\omega)$ é dada por,

$$H(\omega) = \frac{\cos\ N\frac{\omega}{2}}{2}\left[\frac{1}{\sin\ \frac{1}{2}\left(\frac{\pi}{N}-\omega\right)}+\frac{1}{\sin\ \frac{1}{2}\left(\frac{\pi}{N}+\omega\right)}\right] \tag{4.28}$$

com zeros em $\omega = \frac{\pi}{N} + k\frac{2\pi}{N}$, com k inteiro e dois pólos em $\omega = \frac{\pi}{N}$ e $\omega = -\frac{\pi}{N}$. Os dois pólos são cancelados pelos dois zeros às mesmas frequências.

A magnitude normalizada da resposta em frequência $H(\omega)$ da "janela de seno" é

$$\left|\widehat{H(\omega)}\right| = \frac{|H(\omega)|}{|H(0)|} = \frac{\left|\cos\ N\frac{\omega}{2}\right|}{2}\sin\left(\frac{\pi}{2N}\right)\left|\frac{1}{\sin\ \frac{1}{2}\left(\frac{\pi}{N}-\omega\right)}+\frac{1}{\sin\ \frac{1}{2}\left(\frac{\pi}{N}+\omega\right)}\right| \tag{4.29}$$

e está representada na figura 4.10.

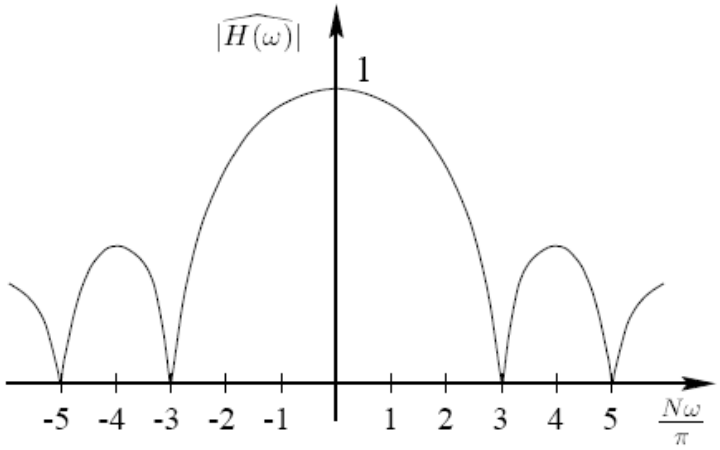


Figura 4.10 – Resposta em frequência normalizada da "janela de seno".

A figura 4.10 mostra que $\widehat{H(\omega)}$ tem características de filtro passa-baixo com uma largura de banda do lobo principal de $6\pi/N$, que a envolvente da banda de rejeição é monotonamente decrescente e que tem zeros às frequências: $\omega = \pm\left(\frac{\pi}{N} + k\frac{2\pi}{N}\right), k = 1, 2, 3 \ldots$

Cada canal do banco de filtros ODFT é obtido modulando $H(\omega)$ às frequências centrais discretas $\omega = \left(k+\frac{1}{2}\right)\frac{2\pi}{N}$, com $k = 1,2,3,\ldots,N-1$ [13]. Na figura 4.11 está ilustrada a modulação e também uma sinusóide com frequência $\omega = \frac{4\pi}{N}$.

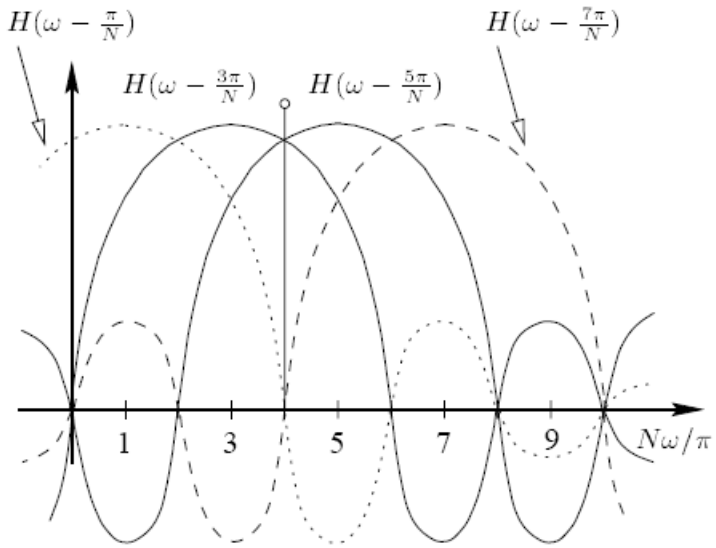


Figura 4.11 – Resposta em frequência dos primeiros quatro canais do banco de filtros ODFT.

A separação entre canais ODFT vizinhos é $2\pi/N$ e os zeros de todas as funções moduladas são múltiplos inteiros de $2\pi/N$.

Uma sinusóide com frequência $\omega = \frac{2\pi}{N}(\ell+\Delta\ell)$, com $1 \leq \ell \leq \frac{N}{2} - 1$ e $0.0 \leq \Delta\ell < 1.0$,, será representada pelo menos por duas sub-bandas abaixo da frequência de Nyquist. Na figura 4.12 estão ilustradas as quatro possibilidades possíveis para as relações entre magnitudes das sub-bandas dos canais com índices $\ell - 1$, ℓ e $\ell + 1$. Observando a figura 4.12 concluímos que, para todos os valores $\Delta\ell$, com excepção de $\Delta\ell = 0.0$, a magnitude da sub-banda ℓ é um máximo local, o que permite que o seu valor possa ser extraído facilmente do espectro da ODFT. O valor de $\Delta\ell$ pode ser estimado tendo em conta as magnitudes relativas das sub-bandas $\ell - 1$ e $\ell + 1$.

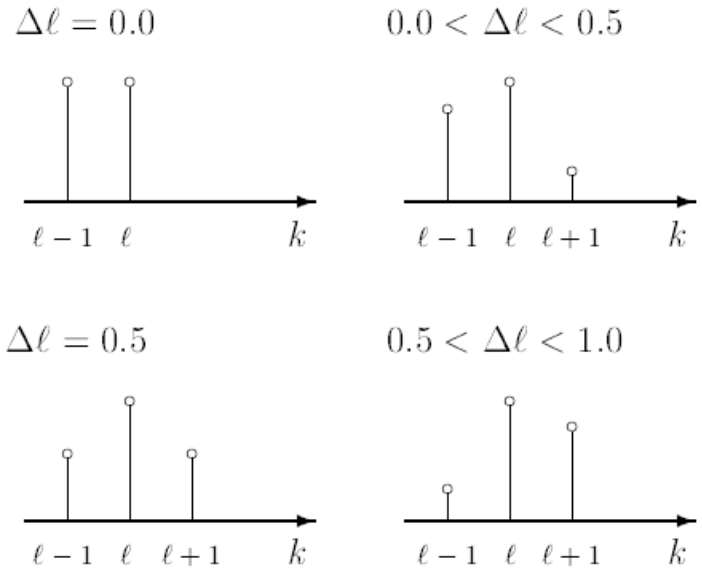


Figura 4.12 – Relação entre as magnitudes dos canais ODFT $\ell - 1$, ℓ e $\ell + 1$ quando o sinal de entrada é uma sinusóide com frequência dada por $\frac{2\pi}{N}(\ell + \Delta\ell)$.

4.5.1.1 - Estimação da frequência

A sinusóide estacionária é projectada nas diferentes sub-bandas como uma função de dois parâmetros:

- a diferença exacta entre a frequência da sinusóide e a frequência central de cada sub-banda ODFT: $\frac{2\pi}{N}(\ell + \Delta\ell) - \frac{2\pi}{N}(k+\frac{1}{2})$ e,
- a forma da resposta em frequência da janela de análise temporal $|H(\omega)|$

Dado que a magnitude do sinal da sub-banda ℓ é um máximo local, $\Delta\ell$ é estimado calculando o rácio entre as magnitudes do sinal da sub-banda $\ell - 1$ e da sub-banda $\ell + 1$,

$$\frac{|X_0(\ell-1)|}{|X_0(\ell+1)|} = \frac{\left|H\left(\frac{2\pi}{N}\left(\ell+\frac{1}{2}\right)\right)\right|}{\left|H\left(\frac{2\pi}{N}\left(\ell-\frac{3}{2}\right)\right)\right|} \tag{4.30}$$

e extraindo o valor de $\Delta\ell$. De modo a reduzir a complexidade do cálculo de $H(\omega)$, utiliza-se a simplificação da forma do lobo principal de $\widehat{H(\omega)}$:

$$\left|\widehat{H(\omega)}\right| \simeq \left[\cos\frac{N}{6}\omega\right]^G, |\omega| < \frac{3\pi}{N} \tag{4.31}$$

como uma aproximação para o lobo principal de $\widehat{H(\omega)}$, sendo *G* uma constante real.
Usando esta aproximação a equação (4.18) pode ser simplificada para:

$$\sqrt[G]{\frac{|X_0(\ell-1)|}{|X_0(\ell+1)|}} + \frac{1}{2} \simeq \frac{\sqrt{3}}{2}\cot\frac{\pi\Delta\ell}{3} \tag{4.32}$$

e $\Delta\ell$ fica então:

$$\Delta\ell \simeq \frac{3}{\pi} arctan \frac{\sqrt{3}}{1+2\left[\frac{|X_0(\ell-1)|}{|X_0(\ell+1)|}\right]^{1/G}} \tag{4.33}$$

O valor utilizado para a constante real *G* foi *27.4/20.0* de modo a minimizar o máximo erro absoluto. Foi conseguido experimentalmente um máximo erro absoluto menor a 1% da largura do *bin* e praticamente independente dos valores de *N*, da frequência *bin*, da amplitude *A* e da fase ϕ. [13]

4.5.1.2 - Estimação da fase

A expressão analítica de $X_0(k)$ considerando apenas o espectro ODFT abaixo da frequência de Nyquist é:

$$X_0(k) = \frac{A}{4}\sin(\pi\Delta\ell)e^{\phi+\tau(\Delta\ell)}\chi\left\{\frac{e^{-j\left(\frac{\pi}{2N}+\tau\left(\frac{\ell+\Delta\ell-k-1}{N}\right)\right)}}{\sin\frac{\pi}{N}(\ell+\Delta\ell-k-1)} + \frac{e^{j\left(\frac{\pi}{2N}-\pi-\tau\left(\frac{\ell+\Delta\ell-k}{N}\right)\right)}}{\sin\frac{\pi}{N}(\ell+\Delta\ell-k)}\right\}, \tag{4.34}$$

com

$$\tau(\alpha) = arctan\frac{-\sin(2\pi\alpha)}{1-\cos(2\pi\alpha)}. \tag{4.35}$$

Sabendo que $\tau(\alpha) = \pi\alpha - \pi/2$, então

$$\angle X_0(\ell - 1) = \phi - \frac{\pi}{2N} + \pi\Delta\ell\left(1 - \frac{1}{N}\right) \tag{4.36}$$

$$\angle X_0(\ell) = \phi - \pi\left(1 - \frac{1}{2N}\right) + \pi\Delta\ell\left(1 - \frac{1}{N}\right) \tag{4.37}$$

Quando $\Delta\ell \neq 0.0$:

- existe mais um termo do que com $\Delta\ell = 0.0$ na expressão da fase que é preciso ter em conta,
- a diferença de fase $\angle X_0(\ell) - \angle X_0(\ell - 1)$ é exactamente $\pi(1/N - 1)$, independentemente do valor de ℓ e ϕ.

O erro de estimação da fase ϕ está, assim, apenas dependente do erro na estimação de $\Delta\ell$.

4.5.1.3 - Estimação da magnitude

A aproximação (4.34) de $\left|\widehat{H(\omega)}\right|$ utilizada para a estimação de $\Delta\ell$ é também usada para a estimação da magnitude com a diferença que no caso da magnitude é usado toda a largura do lobo principal de $\left|\widehat{H(\omega)}\right|$, ou seja, 6π/N.

- Se $\Delta\ell = 0.0$ a equação (4.22) pode ser simplificada para,

$$X_0(k) = \frac{NA}{4} X\left\{e^{-j\left(\frac{\pi}{2N} + \tau\left(\frac{\ell+\Delta\ell-k-1}{N}\right)\right)} + e^{j\left(\frac{\pi}{2N} - \pi - \tau\left(\frac{\ell+\Delta\ell-k}{N}\right)\right)}\right\}, \tag{4.38}$$

  e a magnitude fica então: $|X_0(\ell)| = |NA|/4$, pelo que $A = \frac{4|X_0(\ell)|}{N}$

- Se $0.0 \leq \Delta\ell \leq 1.0$, como sugerido pela figura 4.11 e usando o modelo anterior, então

$$|X_0(\ell)| \simeq \frac{|NA|}{4}\left|\frac{2}{\sqrt{3}}\cos\frac{\pi}{6}(2\Delta\ell - 1)\right|^F, \tag{4.39}$$

sendo *F* uma constante real.
A magnitude A fica então:

$$A \simeq \frac{4|X_0(\ell)|}{N}\left|\frac{\sqrt{3}}{2\cos\frac{\pi}{6}(2\Delta\ell - 1)}\right|^F \tag{4.40}$$

O valor óptimo para a constante real $F = 33.0/20.0$ não é igual ao valor óptimo para a constante real G [13].

4.5.2 - Introdução prática ao "método das sinusóides"

No "método das sinusóides" a fase da análise inicia-se com a segmentação do sinal de entrada em *frames* aplicando-lhes uma "janela de seno". Depois calcula-se a ODFT (Odd Discrete Fourier Transform) da *frame*, a sua envolvente espectral e tenta-se determinar a frequência fundamental e a estrutura harmónica mais proeminente no espectro. Estima-se de seguida a envolvente espectral usando o *cepstrum*, para tal extrai-se o logaritmo ao módulo da ODFT, aplica-se o inverso da ODFT (Inverse Odd Discrete Fourier Transform - IODFT), faz-se uma "lifteragem" de banda estreita e calcula-se novamente a ODFT. Procuram-se todos os máximos relevantes da envolvente espectral inicialmente calculada e para todos eles estima-

se o $\Delta\ell$. Depois, tendo em conta o valor da frequência fundamental estimada, verifica-se quais dos máximos fazem de facto parte da estrutura harmónica e, para os que fizerem, determinam-se os seus verdadeiros valores de ℓ, $\Delta\ell$, ϕ e A. Se se desejar pode-se agora modificar os valores de A, ω ou ϕ e sintetizam-se as sinusóides da nova estrutura harmónica. Por fim, aplica-se a IODFT à *frame* e coloca-se o resultado no vector de saída. Procede-se depois para a análise da *frame* seguinte e sobrepõe-se à *frame* anterior com 50% de sobreposição para haver uma reconstrução correcta do sinal. Este ciclo repete-se até ser processada a última *frame*.

O diagrama de blocos do processo de análise e síntese do "método das sinusóides" descrito anteriormente está representado na figura 4.13.

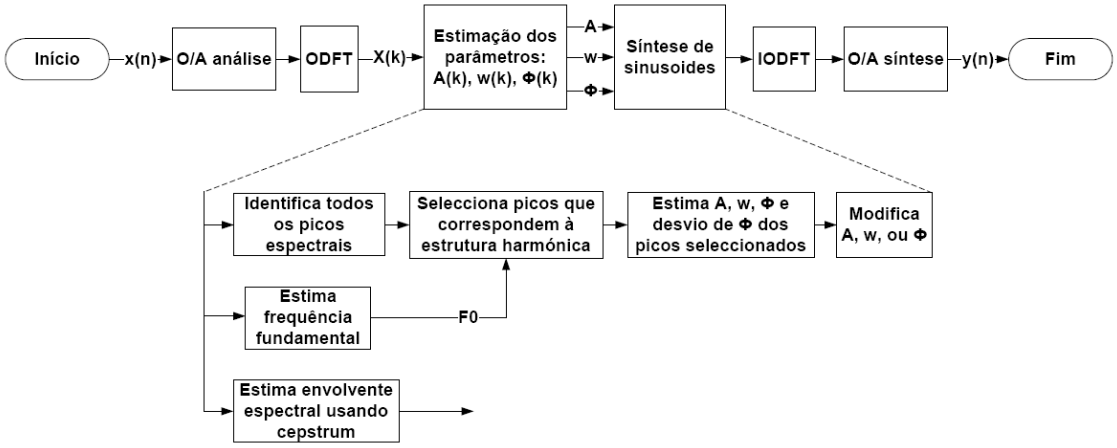


Figura 4.13 – Diagrama de blocos do "método das sinusóides".

4.5.3 – Experiências baseadas no "método das sinusóides"

O "método das sinusóides" difere dos métodos descritos anteriormente, pois ao contrário desses não vai ser experimentado em consoantes, mas nas vogais portuguesas /a/, /ɛ/, /i/, /ɔ/ e /u/. As experiências baseadas no "método das sinusóides" têm como objectivo determinar quais as alterações que ocorrem na percepção, quando as sinusóides que constituem as estruturas harmónicas das vogais portuguesas são geradas do processo de síntese.

O código Matlab para o "método das sinusóides" pode ser consultado no ANEXO F.

4.5.3.1 – Alteração da magnitude

A primeira experiência consiste em reduzir a amplitude de determinadas sinusóides da estrutura harmónica das vogais. Pretende-se com esta experiência identificar quais as sinusóides típicas para cada uma das vogais e que alterações ocorrem em termos de percepção quando a magnitude dessas sinusóides é modificada.

Os testes de alteração da magnitude consistem em reduzir a amplitude dos parciais (1 a 5, 6 a 10, 1 a 10, 11 a 15 e 16 a 30) da estrutura harmónica para 80%, 60%, 40% e 20% dos seus valores originais.

4.5.3.2 – Alteração da fase

Com a segunda experiência do "método das sinusóides" pretende-se identificar quais as modificações de percepção que ocorrem quando as fases de determinadas sinusóides que constituem as vogais portuguesas são alteradas.

Os testes de alteração da fase consistem em iniciar cada sinusóide sintetizada dos parciais (1 a 5, 6 a 10, 1 a 10, 11 a 15 e 16 a 30) da estrutura harmónica com uma fase de 0º, $\frac{\pi}{2}$ e por último um teste em que a fase tem o valor de $\frac{\pi}{4}$ multiplicado pelo número do parcial, ou seja, o valor do primeiro parcial corresponde a $\frac{\pi}{4}$ e o do quarto parcial π.

Capítulo 5

Discussão de resultados

5.1 – Introdução à metodologia usada durante a fase de testes

Este trabalho aborda diferentes métodos de extracção de características de sinais de voz do ponto de vista da percepção auditiva do sinal. A percepção auditiva é algo subjectivo e, como tal, o mesmo sinal de voz é naturalmente interpretado de maneira distinta por ouvintes diferentes. Para reduzir o grau de subjectividade dos resultados obtidos com este trabalho todos os sinais de fala são classificados quanto à sua inteligibilidade e qualidade por três ouvintes diferentes, que não têm acesso às avaliações dos restantes. Para aumentar o rigor da avaliação todas as avaliações são feitas em ambiente silencioso e é permitido aos avaliadores ouvirem os sinais de voz novamente sempre que desejarem.

Os métodos caracterizados do capítulo 4 foram desenvolvidos para serem testados em ambientes diferentes. Os métodos LPC, MFCC e PLP foram testados em classes específicas de fonemas pertencentes às consoantes, enquanto que o "método das sinusóides" foi aplicado a cinco vogais do Português.

5.1.1 – Metodologia usada na avaliação das consoantes

O teste às consoantes foca-se em três classes diferentes de consoantes: fricativas, plosivas e aspiradas. Os sinais de voz utilizados no teste às consoantes representam palavras e não os próprios fonemas isolados. As sessenta e oito palavras inglesas utilizadas nos testes são retiradas da base de dados TIMIT e foram produzidas por quatro oradores, dois do sexo masculino e dois do sexo feminino, de regiões diferentes dos Estados Unidos da América. O facto das palavras utilizadas serem inglesas não tem influência nas conclusões retiradas deste estudo, pois os três avaliadores participantes têm um bom conhecimento da língua inglesa.

As experiências usando os métodos LPC, MFCC e PLP estão divididas em três fases distintas. A primeira tem como objectivo avaliar a inteligibilidade dos sinais ressintetizados. Este é o primeiro teste a ser realizado. A palavra ressintetizada é reproduzida sem que a pessoa que está a fazer a avaliação tenha conhecimento que palavra é, apenas sabe que se trata de uma palavra inglesa e é-lhe pedida que identifique a palavra. A palavra é considerada inteligível se for correctamente identificada. O teste da inteligibilidade parte do pressuposto que o avaliador não conhece à priori a palavra que terá que identificar e, por

esse motivo, não faz sentido efectuar o teste para todas as palavras com todas as ressínteses de todos os métodos e cada uma das suas variantes. Com o objectivo de se retirar a maior quantidade possível de informação em termos de inteligibilidade o teste da inteligibilidade foi efectuado com as ressínteses do método que inicialmente, durante a fase de treino e antes da fase de testes, parecia produzir as ressínteses de pior qualidade. Caso a pessoa que está a fazer o teste identifique a palavra que está a ser reproduzida, essa palavra e todas as suas ressínteses são consideradas inteligíveis, caso contrário a ressíntese em questão é classificada como incompreensível e o teste é repetido com a ressíntese do segundo método que inicialmente parecia indicar os piores resultados. Este processo é repetido até que o avaliador identifique a palavra de uma das ressínteses ou até todas elas serem classificadas como não compreensíveis. Se o avaliador conseguir identificar a palavra de uma das ressínteses, a ressíntese em avaliação e todas as que ainda não tenham sido avaliadas são consideradas como inteligíveis.

O método PLP com seis coeficientes usados na análise/síntese era o que aparentava inicialmente reproduzir os piores resultados, em segundo lugar era o método MFCC com seis coeficientes e depois dos mesmos dois métodos, mas com treze coeficientes, e foi essa a ordem que foi seguida na realização das avaliações.

A segunda fase prende-se com a classificação da qualidade dos sinais ressintetizados comparativamente com os respectivos sinais originais. Nesta segunda fase a metodologia utilizada para aumentar a objectividade das avaliações segue a recomendação da BS.1116 da ITU-R [5], com a diferença que neste trabalho, para tornar o procedimento de teste menos moroso, os avaliadores utilizam uma escala discreta para a degradação subjectiva de áudio. Para cada avaliação da qualidade de uma ressíntese são dados a ouvir aos avaliadores, primeiro o sinal de fala original e, de seguida, o sinal de voz sintetizado e o sinal de voz original, sendo que a ordem pela qual são dados a ouvir o sinal sintetizado e o original é aleatória e desconhecida para o avaliador. É depois pedido que, dos dois sinais de voz reproduzidos por ordem aleatória, identifiquem o sinal de voz original e classifiquem a qualidade da ressíntese comparativamente ao sinal original de acordo com a escala de percepção subjectiva representada na tabela 5.1. Para que cada avaliador tenha a mesma noção do que representa cada um dos patamares da escala de percepção subjectiva foram seleccionados e dados a ouvir antes da avaliação dois exemplos para cada um dos patamares de ressínteses de sinais de voz extraídos da base de dados TIMIT, mas de um quinto orador que somente foi usado na fase de treino. Sempre que desejarem é permitido aos avaliadores requisitarem uma nova audição dos exemplos dos patamares da escala de percepção subjectiva.

Tabela 5.1 - Escala absoluta para a degradação subjectiva de áudio codificado. Adaptado de [5].

Classificação	Descrição
5	É impossível distinguir o sinal de voz original da ressíntese.
4	É possível distinguir o sinal de voz original da ressíntese, mas a diferença é mínima e não perturba a qualidade do sinal de fala.
3	As diferenças entre o sinal original e a ressíntese são perceptíveis, mas apenas algo perturbadoras.
2	As diferenças entre o sinal original e a ressíntese são perturbadoras.
1	As diferenças entre o sinal original e a ressíntese são muito perturbadoras, podendo mesmo ser imperceptível toda ou parte da palavra ressintetizada.

Como é possível verificar a tabela 5.1 é composta por cinco patamares, sendo que a classificação 5 apenas é atribuída quando o avaliador não consegue distinguir o sinal original da ressíntese ou quando identifica incorrectamente o sinal original.

A terceira, e última, fase da avaliação das consoantes pretende que os avaliadores indiquem as alterações que conseguem identificar em fonemas específicos para todos os métodos utilizados. Os dezanove sinais de fala utilizados nesta fase são um subconjunto dos usados na segunda fase e cada fonema que se pretende avaliar está presente em pelo menos duas palavras distintas.

5.1.2 – Metodologia usada na avaliação das vogais

A avaliação das vogais /a/, /ɛ/, /i/, /ɔ/ e /u/ da língua portuguesa segue a metodologia descrita na secção 5.1. As vogais foram obtidas de dois oradores, um do sexo masculino e outro do sexo feminino. Aos três avaliadores é pedido que identifiquem o sinal original e que indiquem as alterações perceptíveis na ressíntese, tendo especial atenção a possíveis alterações da própria inteligibilidade do fonema em questão. A avaliação é efectuada para todas as ressínteses que têm a sua magnitude alterada e também para as que têm a sua fase modificada.

5.2 – Resultados obtidos

Os resultados serão apresentados como um conjunto das três avaliações individuais e não como três avaliações em separado, pois o objectivo de haver três avaliadores é apenas reduzir a subjectividade do estudo.

5.2.1 – Avaliação das consoantes

5.2.1.1 – Inteligibilidade

A tabela 5.2 representa um resumo dos resultados da primeira fase da avaliação das consoantes. Na primeira coluna estão indicadas as palavras avaliadas e os respectivos oradores e nas colunas seguintes os diferentes métodos utilizados com as respectivas variantes. Cada par Palavra-Método está classificado com um "SIM" ou um "Não", sendo que um "SIM" indica que a ressíntese correspondente ao par Palavra-Método foi classificada por dois ou pelos três avaliadores como inteligível e um "NÃO" indica a ressíntese foi classificada por dois ou pelos três avaliadores como não-inteligível.

Tabela 5.2 - Resumo dos resultados da primeira fase da avaliação das consoantes.

Palavras	Métodos				
	LPC				
	Quantização do resíduo			Excitação por ruído branco	*Frames* com pitch inalteradas
	2 bit	3 bit	4 bit		
orador 1					
amounts	SIM	SIM	SIM	SIM	SIM
church	SIM	SIM	SIM	SIM	SIM
direction	SIM	SIM	SIM	SIM	SIM
exchanged	SIM	SIM	SIM	SIM	SIM
greasy	SIM	SIM	SIM	SIM	SIM
initiative	SIM	SIM	SIM	SIM	SIM
involved	SIM	SIM	SIM	SIM	SIM
lawful	SIM	SIM	SIM	SIM	SIM
manufacturer	SIM	SIM	SIM	SIM	SIM
pointing	SIM	SIM	SIM	SIM	SIM
simpler	SIM	SIM	SIM	SIM	SIM
taken	SIM	SIM	SIM	SIM	SIM
terms	SIM	SIM	SIM	NÃO	SIM
worship	SIM	SIM	SIM	SIM	SIM
orador 2					
archeological	SIM	SIM	SIM	SIM	SIM
blistered	SIM	SIM	SIM	NÃO	NÃO
common	SIM	SIM	SIM	SIM	SIM
divorced	SIM	SIM	SIM	SIM	SIM
fuming	SIM	SIM	SIM	SIM	SIM
greasy	SIM	SIM	SIM	SIM	SIM
helpless	SIM	SIM	SIM	SIM	SIM
museum	SIM	SIM	SIM	SIM	SIM
nevada	SIM	SIM	SIM	NÃO	NÃO
relaxed	SIM	SIM	SIM	SIM	SIM
underbrush	SIM	SIM	SIM	SIM	SIM
wealth	SIM	SIM	SIM	SIM	SIM
orador 3					
advisement	SIM	SIM	SIM	SIM	SIM
arbitrate	SIM	SIM	SIM	SIM	SIM
brother	SIM	SIM	SIM	NÃO	SIM
choreographer	SIM	SIM	SIM	SIM	SIM
circle	SIM	SIM	SIM	SIM	SIM
confirm	SIM	SIM	SIM	NÃO	SIM
government	SIM	SIM	SIM	SIM	SIM
greasy	SIM	SIM	SIM	SIM	SIM
imagination	SIM	SIM	SIM	SIM	SIM
masquerade	SIM	SIM	SIM	NÃO	SIM

other	SIM	SIM	SIM	SIM	SIM
outdoors	SIM	SIM	SIM	SIM	SIM
parties	SIM	SIM	SIM	SIM	SIM
policy	SIM	SIM	SIM	SIM	SIM
regarding	SIM	SIM	SIM	SIM	SIM
removal	SIM	SIM	SIM	SIM	SIM
repainted	SIM	SIM	SIM	SIM	SIM
shimmers	SIM	SIM	SIM	SIM	SIM
straight	SIM	SIM	SIM	SIM	SIM
sunshine	SIM	SIM	SIM	SIM	SIM
symbolize	SIM	SIM	SIM	SIM	SIM
uniqueness	SIM	SIM	SIM	SIM	SIM
universality	SIM	SIM	SIM	SIM	SIM
orador 4					
almost	SIM	SIM	SIM	SIM	SIM
aprons	SIM	SIM	SIM	SIM	SIM
assume	SIM	SIM	SIM	NÃO	SIM
available	SIM	SIM	SIM	SIM	SIM
becoming	SIM	SIM	SIM	SIM	SIM
change	SIM	SIM	SIM	SIM	SIM
coeducational	SIM	SIM	SIM	SIM	SIM
colleges	SIM	SIM	SIM	SIM	SIM
drawing	SIM	SIM	SIM	NÃO	SIM
famous	SIM	SIM	SIM	SIM	SIM
greasy	SIM	SIM	SIM	SIM	SIM
holiday	SIM	SIM	SIM	SIM	SIM
ignored	SIM	SIM	SIM	SIM	SIM
resolute	SIM	SIM	SIM	SIM	SIM
resolved	SIM	SIM	SIM	SIM	SIM
slavery	SIM	SIM	SIM	NÃO	SIM
thinker	SIM	SIM	SIM	SIM	SIM
unappreciated	SIM	SIM	SIM	SIM	SIM
wishful	SIM	SIM	SIM	SIM	SIM

Palavras	**Métodos**				
	LPC	**MFCC**		**PLP**	
	***Frames* sem pitch inalteradas**	**6 coeficientes**	**13 coeficientes**	**6 coeficientes**	**13 coeficientes**
orador 1					
amounts	SIM	SIM	SIM	SIM	SIM
church	SIM	SIM	SIM	SIM	SIM
direction	SIM	SIM	SIM	SIM	SIM
exchanged	SIM	SIM	SIM	SIM	SIM
greasy	SIM	SIM	SIM	SIM	SIM
initiative	SIM	SIM	SIM	SIM	SIM

involved	SIM	SIM	SIM	SIM	SIM
lawful	SIM	NÃO	SIM	NÃO	NÃO
manufacturer	SIM	SIM	SIM	SIM	SIM
pointing	SIM	SIM	SIM	SIM	SIM
simpler	SIM	SIM	SIM	SIM	SIM
taken	SIM	SIM	SIM	SIM	SIM
terms	SIM	NÃO	NÃO	NÃO	NÃO
worship	SIM	SIM	SIM	SIM	SIM
orador 2					
archeological	SIM	SIM	SIM	SIM	SIM
blistered	SIM	NÃO	NÃO	NÃO	NÃO
common	SIM	NÃO	SIM	NÃO	SIM
divorced	SIM	SIM	SIM	SIM	SIM
fuming	SIM	SIM	SIM	SIM	SIM
greasy	SIM	SIM	SIM	SIM	SIM
helpless	SIM	SIM	SIM	SIM	SIM
museum	SIM	SIM	SIM	SIM	SIM
nevada	SIM	NÃO	NÃO	NÃO	NÃO
relaxed	SIM	SIM	SIM	SIM	SIM
underbrush	SIM	SIM	SIM	SIM	SIM
wealth	SIM	SIM	SIM	SIM	SIM
orador 3					
advisement	SIM	SIM	SIM	SIM	SIM
arbitrate	SIM	SIM	SIM	SIM	SIM
brother	NÃO	NÃO	NÃO	NÃO	NÃO
choreographer	SIM	NÃO	NÃO	NÃO	NÃO
circle	SIM	SIM	SIM	SIM	SIM
confirm	SIM	NÃO	NÃO	NÃO	NÃO
government	SIM	NÃO	SIM	NÃO	SIM
greasy	SIM	SIM	SIM	SIM	SIM
imagination	SIM	SIM	SIM	SIM	SIM
masquerade	SIM	NÃO	NÃO	NÃO	NÃO
other	SIM	SIM	SIM	SIM	SIM
outdoors	SIM	SIM	SIM	SIM	SIM
parties	SIM	SIM	SIM	SIM	SIM
policy	SIM	NÃO	NÃO	NÃO	NÃO
regarding	SIM	SIM	SIM	SIM	SIM
removal	SIM	SIM	SIM	SIM	SIM
repainted	SIM	SIM	SIM	SIM	SIM
shimmers	SIM	SIM	SIM	SIM	SIM
straight	SIM	SIM	SIM	SIM	SIM
sunshine	SIM	SIM	SIM	SIM	SIM
symbolize	SIM	SIM	SIM	SIM	SIM
uniqueness	SIM	SIM	SIM	SIM	SIM
universality	SIM	SIM	SIM	SIM	SIM
orador 4					

almost	SIM	SIM	SIM	SIM	SIM
aprons	SIM	SIM	SIM	SIM	SIM
assume	SIM	NÃO	SIM	NÃO	NÃO
available	SIM	SIM	SIM	SIM	SIM
becoming	SIM	SIM	SIM	SIM	SIM
change	SIM	SIM	SIM	SIM	SIM
coeducational	SIM	SIM	SIM	SIM	SIM
colleges	SIM	SIM	SIM	SIM	SIM
drawing	SIM	NÃO	SIM	NÃO	NÃO
famous	SIM	SIM	SIM	SIM	SIM
greasy	SIM	SIM	SIM	SIM	SIM
holiday	SIM	NÃO	NÃO	NÃO	NÃO
ignored	SIM	SIM	SIM	SIM	SIM
resolute	SIM	NÃO	SIM	NÃO	NÃO
resolved	SIM	NÃO	SIM	NÃO	NÃO
slavery	SIM	NÃO	NÃO	NÃO	NÃO
thinker	SIM	NÃO	SIM	NÃO	NÃO
unappreciated	SIM	SIM	SIM	SIM	SIM
wishful	SIM	SIM	SIM	SIM	SIM

A tabela 5.3 consiste num resumo estatístico da tabela 5.2, indicando qual a percentagem de palavras inteligíveis.

Tabela 5.3 - Resumo estatístico do estudo da inteligibilidade.

	Métodos				
	LPC				
	Quantização do resíduo			**Excitação por ruído aleatório**	***Frames* com pitch inalteradas**
	2 bit	**3 bit**	**4 bit**		
orador 1	100,0%	100,0%	100,0%	92,9%	100,0%
orador 2	100,0%	100,0%	100,0%	83,3%	83,3%
orador 3	100,0%	100,0%	100,0%	87,0%	100,0%
orador 4	100,0%	100,0%	100,0%	84,2%	100,0%
total	100,0%	100,0%	100,0%	86,8%	97,1%

	Métodos				
	LPC	MFCC		PLP	
	Frames sem pitch inalteradas	6 coeficientes	13 coeficientes	6 coeficientes	13 coeficientes
orador 1	100,0%	85,7%	92,9%	85,7%	85,7%
orador 2	100,0%	75,0%	83,3%	75,0%	83,3%
orador 3	95,7%	73,9%	78,3%	73,9%	78,3%
orador 4	100,0%	63,2%	89,5%	63,2%	63,2%
total	98,5%	73,5%	85,3%	73,5%	76,5%

5.2.1.2 – Qualidade das ressínteses comparativamente aos sinais originais

A tabela 5.4 representa as classificações da segunda fase da avaliação das consoantes. Na primeira coluna estão indicadas as palavras avaliadas e os respectivos oradores e nas colunas seguintes os diferentes métodos utilizados com as respectivas variantes. Cada par Palavra-Método está classificado com um número compreendido entre 1 e 5, correspondendo o número à média das classificações dadas pelos três avaliadores.

Tabela 5.4 – Classificações da segunda fase da avaliação das consoantes.

Palavras	Métodos				
	LPC				
	Quantização do resíduo			Excitação por ruído branco	*Frames* com pitch inalteradas
	2 bit	3 bit	4 bit		
orador 1					
amounts	3,3	4,0	4,0	1,7	5,0
church	3,0	3,7	4,0	1,0	4,0
direction	3,7	4,0	4,3	1,3	3,7
exchanged	3,3	4,0	4,3	1,3	3,7
greasy	3,7	4,0	4,0	1,7	4,0
initiative	3,7	4,0	4,0	1,3	4,0
involved	4,0	4,0	4,0	1,0	3,3
lawful	3,3	4,0	4,0	1,3	4,0
manufacturer	3,0	3,7	4,0	1,3	4,0
pointing	3,7	4,0	4,0	1,7	4,0
simpler	3,3	4,0	4,0	1,3	4,0
taken	3,3	4,0	4,3	1,7	3,7
terms	3,0	4,0	4,3	1,7	4,0
worship	3,3	4,0	4,0	2,0	4,0
orador 2					
archeological	3,3	4,0	4,0	2,0	3,3

blistered	3,3	4,0	4,3	1,7	4,0
common	3,7	4,0	4,0	1,0	4,0
divorced	3,7	4,0	4,0	1,7	3,7
fuming	4,0	4,0	4,3	2,0	4,0
greasy	3,3	4,0	4,7	1,7	3,3
helpless	3,7	4,0	4,0	1,7	4,0
museum	4,0	4,0	4,7	2,0	3,3
nevada	3,3	4,0	4,0	1,3	3,0
relaxed	3,7	4,0	4,3	1,0	3,0
underbrush	4,0	4,0	4,3	1,7	4,0
wealth	4,0	4,0	4,3	1,7	4,0
orador 3					
advisement	3,7	4,0	4,0	1,7	4,0
arbitrate	3,7	4,0	4,0	1,0	4,0
brother	3,0	4,0	4,0	1,0	4,3
choreographer	3,0	3,3	4,0	1,3	4,0
circle	3,7	4,0	4,0	1,0	4,0
confirm	3,7	4,0	4,0	1,3	4,0
government	3,3	4,0	4,0	1,0	4,0
greasy	4,0	4,0	4,3	1,7	4,0
imagination	3,7	4,0	4,0	1,7	4,0
masquerade	3,3	4,0	4,0	1,0	4,0
other	2,3	4,0	4,0	2,0	4,0
outdoors	3,3	4,0	4,0	1,0	4,0
parties	3,3	4,0	4,0	1,7	4,0
policy	3,3	4,0	4,0	1,7	4,0
regarding	3,0	3,7	4,0	1,3	4,0
removal	3,7	4,0	4,0	1,7	4,0
repainted	3,0	4,0	4,0	1,7	4,0
shimmers	3,7	4,0	4,0	1,3	4,0
straight	3,0	4,0	4,0	1,3	3,7
sunshine	3,7	4,0	4,0	1,3	4,0
symbolize	3,3	4,0	4,0	1,3	3,7
uniqueness	3,7	4,0	4,0	1,3	4,0
universality	3,3	4,0	4,0	1,0	3,7
orador 4					
almost	3,3	3,7	4,0	1,3	3,3
aprons	3,7	3,7	4,0	1,0	4,0
assume	4,0	4,0	4,0	1,0	3,7
available	3,3	4,0	4,0	1,0	3,3
becoming	4,0	4,0	4,0	1,3	4,0
change	4,0	4,0	4,0	1,7	4,0
coeducational	3,0	4,0	4,0	1,3	3,3
colleges	3,7	4,0	4,0	1,0	4,0
drawing	3,3	4,0	4,0	1,3	3,3
famous	3,7	4,0	4,0	1,7	4,0

greasy	4,0	4,0	4,0	1,3	3,7
holiday	3,0	3,7	4,0	1,0	3,3
ignored	3,0	4,0	4,0	1,3	4,0
resolute	3,3	4,0	4,0	1,0	3,7
resolved	3,0	3,7	4,0	1,0	4,0
slavery	3,7	4,0	4,0	1,3	4,0
thinker	3,0	4,0	4,7	1,0	2,3
unappreciated	3,7	4,0	4,3	1,3	3,3
wishful	3,7	4,0	4,3	1,7	4,0

Palavras	**Métodos**				
	LPC	**MFCC**		**PLP**	
	***Frames* sem pitch inalteradas**	**6 coeficientes**	**13 coeficientes**	**6 coeficientes**	**13 coeficientes**
orador 1					
amounts	3,0	1,7	2,0	1,0	1,7
church	3,3	1,3	1,7	1,0	1,0
direction	3,7	1,7	2,0	1,0	1,0
exchanged	3,7	1,7	2,0	1,0	1,0
greasy	4,0	1,7	2,0	1,0	1,0
initiative	4,0	1,0	1,3	1,0	1,0
involved	3,3	1,0	1,7	1,0	1,0
lawful	3,0	1,3	2,0	1,0	1,0
manufacturer	2,7	2,0	2,0	1,0	1,0
pointing	3,3	1,3	2,0	1,0	1,3
simpler	3,7	1,3	2,0	1,0	1,0
taken	4,0	1,7	1,7	1,0	1,3
terms	3,3	1,0	1,7	1,0	1,0
worship	3,7	2,0	2,0	1,0	1,0
orador 2					
archeological	3,7	1,7	2,0	1,3	1,0
blistered	3,3	1,3	2,0	1,3	1,0
common	3,7	1,0	1,3	1,0	1,3
divorced	3,7	2,0	2,0	1,3	1,3
fuming	4,0	1,7	1,7	1,3	1,0
greasy	3,7	1,7	2,0	1,3	1,3
helpless	4,0	1,3	2,0	1,0	1,0
museum	4,0	1,7	1,7	1,0	1,0
nevada	4,0	1,0	1,7	1,0	1,0
relaxed	4,0	1,3	2,0	1,0	1,0
underbrush	3,7	1,7	2,0	1,0	1,7
wealth	3,3	1,3	2,0	1,0	1,3
orador 3					

advisement	4,0	1,3	2,0	1,0	1,3
arbitrate	4,0	1,7	2,0	1,0	1,0
brother	4,0	1,0	2,0	1,0	1,3
choreographer	3,7	1,3	2,0	1,0	1,0
circle	4,0	1,7	2,0	1,0	1,0
confirm	3,7	1,3	1,7	1,0	1,0
government	3,7	1,0	1,7	1,0	1,0
greasy	4,0	1,7	2,0	1,0	1,3
imagination	3,3	1,7	2,0	1,0	1,3
masquerade	3,0	1,3	1,7	1,0	1,0
other	4,0	1,0	2,0	0,7	1,3
outdoors	4,0	1,7	1,7	1,0	1,7
parties	4,0	1,3	2,0	1,0	1,0
policy	4,0	1,7	2,0	1,0	1,3
regarding	3,3	1,3	2,0	1,0	1,0
removal	3,3	2,0	2,0	1,3	1,3
repainted	3,0	2,0	2,0	1,3	1,0
shimmers	4,0	1,7	1,7	1,3	1,0
straight	3,7	1,7	1,7	1,3	1,3
sunshine	3,7	2,0	2,0	1,3	1,0
symbolize	3,3	2,0	2,0	1,0	1,3
uniqueness	3,7	1,3	1,7	1,0	1,0
universality	3,7	2,0	2,0	1,0	1,3
orador 4					
almost	3,3	1,0	1,7	1,0	1,0
aprons	3,0	1,0	1,3	1,0	1,0
assume	3,3	1,0	1,3	1,0	1,0
available	3,3	1,0	1,7	1,0	1,0
becoming	3,7	1,3	2,0	1,3	1,0
change	4,0	1,3	2,0	1,3	1,3
coeducational	4,0	1,3	1,7	1,0	1,0
colleges	3,3	1,0	1,7	1,0	1,0
drawing	3,3	1,0	1,7	1,0	1,0
famous	4,0	1,3	2,0	1,3	1,0
greasy	4,0	1,3	1,7	1,0	1,3
holiday	3,7	1,0	1,3	1,0	1,0
ignored	3,3	1,0	1,7	1,0	1,0
resolute	4,0	1,0	1,3	1,0	1,0
resolved	3,0	1,0	1,7	1,0	1,0
slavery	3,7	1,0	1,3	1,0	1,0
thinker	4,0	1,0	1,3	1,0	1,0
unappreciated	4,0	1,0	1,7	1,0	1,0
wishful	3,7	1,0	2,0	1,0	1,0

A tabela 5.5 consiste num resumo estatístico da tabela 5.4, indicando qual a média e variância da classificação atribuída a cada método. Os valores são dados por cada orador individualmente e os valores totais também são indicados.

Tabela 5.5 – Valores da média e variância da classificação atribuída aos diferentes métodos..

		Métodos				
		LPC				
		Quantização do resíduo			**Excitação por ruído branco**	***Frames* com pitch inalteradas**
		2 bit	**3 bit**	**4 bit**		
orador 1	**média**	3,40	3,96	4,09	1,45	3,96
	variância	0,10	0,01	0,02	0,09	0,13
orador 2	**média**	3,67	4,00	4,24	1,63	3,63
	variância	0,09	0,00	0,07	0,12	0,18
orador 3	**média**	3,38	3,96	4,01	1,36	3,97
	variância	0,15	0,02	0,00	0,10	0,02
orador 4	**média**	3,50	3,94	4,07	1,24	3,64
	variância	0,15	0,02	0,03	0,06	0,20
total	**média**	3,47	3,96	4,08	1,39	3,82
	variância	0,13	0,02	0,03	0,10	0,14

		Métodos				
		LPC	**MFCC**		**PLP**	
		***Frames* sem pitch inalteradas**	**6 coeficientes**	**13 coeficientes**	**6 coeficientes**	**13 coeficientes**
orador 1	**média**	3,48	1,48	1,86	1,00	1,09
	variância	0,17	0,12	0,05	0,00	0,04
orador 2	**média**	3,76	1,48	1,87	1,13	1,16
	variância	0,07	0,10	0,05	0,02	0,05
orador 3	**média**	3,70	1,55	1,91	1,05	1,16
	variância	0,12	0,12	0,02	0,02	0,04
orador 4	**média**	3,61	1,08	1,64	1,05	1,03
	variância	0,13	0,02	0,07	0,01	0,01
total	**média**	3,64	1,39	1,82	1,05	1,11
	variância	0,13	0,12	0,06	0,02	0,03

5.2.1.3 – Anotações às alterações de fonemas específicos

A tabela 5.6 representa o subconjunto das palavras utilizadas na avaliação das alterações fonéticas nas diferentes ressínteses dos sinais de fala. Cada linha representa um fonema e na

primeira coluna estão indicados o tipo de fonema e a classe fonética a que pertence. Nas restantes colunas estão representadas as palavras que foram utilizadas na avaliação de cada um dos fonemas, sendo que a coluna em que essas palavras estão inseridas indicam qual o orador que as pronunciou.

Tabela 5.6 – Subconjunto das palavras utilizadas na avaliação das alterações fonéticas.

Fonemas	Palavras			
	orador 1	orador 2	orador 3	orador 4
Fricativas não-vozeadas				
/f/	manufacturer	fuming	confirm	
/T/		wealth		thinker
/s/	simpler	divorced, helpless	symbolyze	
/S/	direction, worship			
Fricativas vozeadas				
/v/	involved	divorced	government	
/D/			brother, other	
/z/		museum	greasy, symbolize	
/Z/	exchanged			
Plosivas não-vozeadas				
/p/	simpler, worship	helpless		
/t/	Terms	blistered		
/k/	direction		confirm	thinker
Plosivas vozeadas				
/b/		blistered	brother, symbolyze	
/d/	direction	divorced		holiday
/g/			government, greasy	
Aspiradas				
/h/		helpless		holiday

Após a análise pormenorizada dos três avaliadores relativamente a todas as palavras seleccionadas é possível concluir que alguns métodos influenciam todos os fonemas de igual modo. É o caso dos métodos LPC com 3 e 4 bit utilizados na quantização do resíduo e o método LPC em que as *frames* em que se identifica *pitch* permanecem inalteradas. Estes métodos praticamente não alteram os fonemas, incluindo no máximo um ruído muito ligeiro. O método LPC em que as *frames* sem *pitch* ficam inalteradas e as com *pitch* ficam com o seu ciclo vibratório reduzido a 25% não influencia os fonemas não-vozeados e atenua a intensidade dos fonemas vozeados, esta alteração era esperada, pois reduzindo numa *frame* cada ciclo vibratório para 25% das suas amostras reduz-se significativamente a energia dessa *frame*. As ressínteses em que é utilizada excitação por ruído branco influenciam os fonemas da mesma forma, independentemente do método ou do número de coeficientes usados.

Nos restantes métodos a avaliação individual aos diferentes fonemas permitiu concluir que existem bastantes semelhanças no tipo de alterações que são produzidas numa mesma classe fonética. A classe das fricativas não-vozeadas não sofre qualquer tipo de alteração, independentemente do fonema específico ou do método utilizado.

As fricativas vozeadas, por outro lado, sofrem a influência de ruído. O fonema /D/ nos métodos de excitação por ruído branco fica praticamente irreconhecível. Nesses mesmos métodos os fonemas /z/ e /Z/ ficam com semelhanças ou ficam mesmo indiferenciáveis dos fonemas /s/ e /S/, respectivamente.

Nas plosivas não-vozeadas todos os fonemas sofrem a influência de ruído, especialmente nos métodos com excitação por ruído branco. Nesses métodos todos os fonemas perdem parte da "explosão" típica das plosivas e apesar de continuar evidente que se trata de uma consoante plosiva, por vezes é difícil identificar qual. O fonema /p/ soa várias vezes como uma mistura entre o fonema /p/ e o fonema /b/ e o fonema /k/ quando ocorre no início da palavra parece uma mistura entre /k/ e /h/.

As plosivas vozeadas sofrem o mesmo tipo de alterações que as plosivas não-vozeadas, mas o efeito do ruído é mais evidente nas vozeadas, especialmente nos fonemas /b/ e /g/, que nos métodos de sinal aleatório ficam várias vezes irreconhecíveis. Nesses mesmos métodos o fonema /d/ parece por vezes um /t/.

A consoante aspirada /h/ fica em qualquer destes métodos com bastante ruído e nos de excitação por ruído branco o fonema /h/ fica também mais intenso.

5.2.2 – Avaliação das vogais

5.2.2.1 – Alteração da magnitude

As experiências com a alteração de magnitude têm diversas vertentes de análise: os parciais modificados, o grau da atenuação da amplitude dos mesmos e as alterações provocadas na percepção das diferentes vogais.

A tabela 5.7 contém as modificações registadas durante as observações às experiências com sinusóides sintéticas com magnitude alterada. A tabela compara os parciais alterados com o efeito que as alterações na magnitude das sinusóides provocaram nas vogais analisadas.

O grau da redução da amplitude das sinusóides não foi registado na tabela 5.7, pois a percentagem da atenuação apenas altera a intensidade das observações. Sendo que quanto maior é a redução da amplitude, mais notório é o efeito descrito na tabela 5.7 nos sinais sintetizados.

Existem algumas diferenças entre o orador masculino e o feminino no grau de modificação que ocorre nas vogais. A vogal /ɛ/ com os parciais 16 a 30 alterados não sofre uma modificação tão acentuada como no caso do orador masculino, por outro lado com os parciais 11 a 15 alterados o /ɛ/ do orador masculino fica pouco modificado. A vogal /ɔ/ com os parciais 6 a 10 alterados é quase irreconhecível para o orador masculino. Em todos os outros casos as diferenças entre os oradores são apenas ligeiras.

Tabela 5.7 – Observações das experiências com magnitude alterada para o orador feminino, tendo em consideração os parciais alterados e as modificações ocorridas nas vogais.

<table>
<tr><td></td><td colspan="5">Vogais</td></tr>
<tr><td>Parciais</td><td>/a/</td><td>/ɛ/</td><td>/i/</td><td>/ɔ/</td><td>/u/</td></tr>
<tr><td>1 a 5</td><td colspan="5">Som é mais baixo
Não se ouve a vibração das pregas vocais
Voz tem um som artificial, nasalado
Apesar de alterados ainda se identificam os fonemas</td></tr>
<tr><td rowspan="2">1 a 10</td><td>Irreconhecível</td><td>Quase irreconhecível</td><td>Bastante alterado</td><td>Irreconhecível</td><td>Irreconhecível</td></tr>
<tr><td colspan="5">Som é mais baixo
Não se ouve a vibração das pregas vocais
Voz tem um som muito artificial</td></tr>
<tr><td>6 a 10</td><td>Irreconhecível</td><td colspan="4">Algo nasalado, mas menos do que nos casos dos parciais 1 a 5
Ainda se identificam os fonemas</td></tr>
<tr><td>11 a 15</td><td>Pouco alterado</td><td>Alterado</td><td>Alterado</td><td>Pouco alterado</td><td>Pouco alterado</td></tr>
<tr><td>16 a 30</td><td>Quase igual</td><td>Muito alterado</td><td>Quase irreconhecível</td><td>Quase igual</td><td>Quase igual</td></tr>
</table>

5.2.2.2 – Alteração da fase

Os diferentes testes efectuados alterando a fase inicial das sinusóides das vogais, sugerem que a alteração não parece introduzir diferenças assinaláveis em termos de percepção dos sinais de voz. Existe uma pequena diferença entre os sinais ressintetizados e o sinal de fala original, mas essa alteração resulta na limitação a trinta parciais na ressíntese. Essa limitação produz nas ressínteses uma redução da estrutura harmónica. Como o intuito desta experiência era identificar as modificações na percepção auditiva provocadas por uma alteração da fase, as ressínteses foram comparadas entre si e entre a ressíntese do sinal original, mas sem que houvesse alterações de qualquer dos seus parâmetros.

5.3 – Discussão dos resultados obtidos

5.3.1 – Consoantes

Analisando os resultados obtidos no teste da inteligibilidade, é possível verificar que em todas as ressínteses em que foram usados métodos LPC de quantização de resíduo a inteligibilidade da palavra foi mantida e que só em três ressínteses no total dos dois métodos PLC de alteração das *frames* não foi possível identificar a palavra, o que corresponde a cerca de 97,8% de ressínteses inteligíveis. É possível concluir que a perda de inteligibilidade praticamente só acontece quando é usada excitação por ruído branco e mesmo nesses casos em 79,1% das situações os avaliadores identificaram correctamente a palavra. O número de palavras ininteligíveis do método MFCC com seis coeficientes e do método PLP com seis

coeficientes é exactamente igual e as palavras que as originaram são as mesmas. Estes dois métodos tiveram os piores resultados, como é possível verificar na tabela 5.3, com uma percentagem de palavras identificadas de 73,5%. Com os mesmos métodos, mas com treze coeficientes verifica-se que o método MFCC obtém melhores resultados que o PLP e ambos conseguem níveis de inteligibilidade superiores aos métodos que usam apenas seis coeficientes. Também é preciso ter em consideração a relevância do tipo de voz na percepção da inteligibilidade. Como o teste da inteligibilidade não é um teste comparativo, se o sinal original já for difícil de identificar o ressintetizado não poderá ter muita qualidade. Como se demonstra na tabela 5.3, o orador 4 tem percentagens mais baixas de palavras inteligíveis e, em parte, isso deve-se ao facto desse orador ter uma voz que torna as palavras menos claras.

A tabela 5.8 contém as médias das classificações atribuídas a cada um dos métodos avaliados.

Tabela 5.8 – Média das classificações dos métodos avaliados.

	Métodos									
	LPC						**MFCC**		**PLP**	
	Quantização			**Excitação por ruído branco**	***Frames* com pitch inalteradas**	***Frames* sem pitch inalteradas**				
	2 bit	**3 bit**	**4 bit**				**6 coef.**	**13 coef.**	**6 coef.**	**13 coef.**
Total	3,47	3,96	4,09	1,39	3,82	3,64	1,39	1,81	1,06	1,12

Analisando as médias das classificações, é notória a grande diferença existente entre os métodos LPC que não usam sinal aleatório e os restantes métodos que utilizam esse tipo de excitação. Essa diferença é ainda mais pronunciada no orador 4, que devido às características da sua voz é especialmente sensível à excitação por ruído branco, como indicam os seus valores nesses métodos, que nunca estão acima da média geral. Conforme esperado, no caso da quantização do resíduo, as classificações melhoram com o aumento dos bits utilizados na quantização. Também se verifica que o método MFCC obtém melhores resultados que o método PLP e que em ambos as ressínteses têm melhor qualidade quanto maior o número de coeficientes usados na análise. No entanto ao contrário do que se verifica no teste da inteligibilidade o método PLP com treze coeficientes utilizados na análise obtém pior média do que o método MFCC com apenas seis coeficientes. Esta diferença pode ter origem no facto do teste da qualidade ser um teste comparativo, ou seja, um teste em que se avalia a qualidade da ressíntese comparativamente ao sinal original e o teste da inteligibilidade é um teste em que se avalia apenas a ressíntese, apesar da inteligibilidade da mesma depender obviamente da inteligibilidade do sinal original. Apesar da percepção auditiva ser, por natureza, subjectiva os resultados da variância indicados na tabela 5.5 demonstram que as opiniões dos avaliadores foram maioritariamente coincidentes.

As avaliações de fonemas específicos permitiram extrair informações bastante interessantes. Tal como no teste da inteligibilidade e no da qualidade das ressínteses, os melhores resultados obtêm-se com os mesmos métodos, mas existe também uma clara relação entre o tipo de modificação da ressíntese e a classe à qual o fonema ressintetizado pertence.

As fricativas não-vozeadas são claramente as consoantes que obtêm melhores resultados. Mesmo utilizando excitação por ruído branco na maior parte dos casos não são detectadas diferenças em relação aos fonemas originais. A natureza turbulenta das consoantes não-vozeadas permite que os métodos que usam excitação por ruído branco consigam ressintetizar muito bem este tipo de fonemas.

As fricativas vozeadas devido à natureza periódica do seu sinal são mais difíceis de ressintetizar. Ruídos ligeiros são mais perceptíveis do que no caso das fricativas não-vozeadas. A excitação de ruído branco provoca a perda da informação periódica do sinal, que provoca alterações muito significativas em termos de percepção. Essas modificações incluem um nível elevado de ruído, que por vezes torna o fonema irreconhecível, mas também, nalguns casos, da alteração da natureza do fonema, passando de um fonema vozeado para um não-vozeado. Nos casos avaliados essa alteração é notória no fonema /z/ e /Z/, que transitam para os fonemas /s/ e /S/, respectivamente. As diferenças na produção dos fonemas /z/ e /Z/ e dos fonemas /s/ e /S/ prende-se, na verdade, com o tipo de excitação de cada um deles, pois o posicionamento dos articuladores envolvidos na sua produção é o mesmo.

No caso das consoantes plosivas a diferença entre sinais vozeados e não-vozeados não é tão evidente. A natureza "explosiva" das consoantes plosivas dificulta a análise e sínteses destes fonemas, mesmo para os métodos LPC que com os outros fonemas obtêm resultados excelentes. A produção de uma consoante plosiva inclui duas fases distintas, na primeira, existe uma obstrução total num ponto do tracto vocal e a ligação para o tracto nasal estão fechadas e há uma acumulação de pressão na boca, na segunda fase, a fase da "explosão", a obstrução desaparece e a pressão acumulada é libertada, podendo essa libertação ser acompanhada pela vibração ou não das pregais vocais, dependendo se se trata de uma plosiva vozeada ou não. Apesar de existirem plosivas vozeadas a duração do vozeamento destas consoantes é consideravelmente mais reduzido do que no caso das fricativas e ainda menos notório, por coincidir com a fase da "explosão". Tal como nas fricativas, também no caso das plosivas existem pares de fonemas que apresentam o mesmo posicionamento dos articuladores durante a sua produção, variando apenas no tipo de excitação. O fonema /p/ faz par com o /b/, o mesmo sucedendo com o /t/ e /d/ e o /k/ e /g/. Devido a estas dificuldades na análise/síntese das plosivas verifica-se que por vezes é difícil identificar se o fonema ressintetizado é vozeado ou não, assemelhando-se a uma mistura dos entre os dois fonemas. Como já foi referido anteriormente, o fonema /k/ nalgumas ressínteses parece uma mistura entre um /k/ e um /h/. Esta aproximação entre os dois fonemas reside no facto do fonema /h/ ter também uma fase de silêncio, já que este fonema só ocorre no início das palavras, seguido por uma fase em que o ar é libertado repentinamente e do fonema /k/ corresponder a uma plosiva velar, que, de todas as plosivas, é a que tem a obstrução mais atrás no tracto vocal.

Os métodos que utilizam excitação por ruído branco têm uma dificuldade acrescida na ressíntese deste tipo de fonemas. A fase da "explosão" nas plosivas tem uma duração inferior à duração das *frames* utilizadas pelos métodos e como a média e variância do ruído é calculada por *frame*, depois da ressíntese a energia concentrada na porção do sinal correspondente à fase da "explosão" é muito inferior à do sinal original, resultando em termos de percepção numa perda da intensidade da "explosão" típica das consoantes plosivas.

5.3.2 – Vogais

A redução da amplitude das sinusóides dos parciais mais baixos do sinal de voz provoca claras modificações na percepção dos sinais. Esses parciais contêm uma percentagem elevada da energia do sinal de fala e a redução da amplitude dessas sinusóides implica uma redução da energia total do sinal e uma diminuição da sua intensidade. No sinal sintetizado deixa de ser perceptível o som da vibração das pregas vocais tipicamente presentes nas vogais e a voz parece nasalada e artificial. Apesar de todas estas alterações quando se reduzem os primeiros cinco parciais ainda é possível distinguir entre as diferentes vogais, pois apesar de importantes em todas as vogais não são estes os parciais que distinguem umas vogais das outras, como é possível observar na figura 5.1. Observando o espectrograma da figura 5.1 do orador masculino é notório que a amplitude das vogais /ɛ/, /i/ e /u/ aproximadamente a partir do quinto parcial atenuam bastante e é por esse motivo que reduzindo para 20% a amplitude dos parciais 6 a 10 não provoca grandes alterações de percepção nestas vogais. As vogais /a/ e /ɔ/ por outro lado já ficam praticamente irreconhecíveis. As vogais /ɛ/ e /i/ a redução dos parciais 16 a 30 provoca alterações evidentes no sinal ressintetizado e visualizando o espectro que na região correspondente aos 1800 - 2100Hz os parciais voltam a ter amplitudes mais elevadas. O orador masculino tem uma frequência fundamental de aproximadamente 115 Hz, por isso, a redução dos parciais 16 a 30 corresponde aproximadamente à região 1850 - 3450Hz, incluindo portanto uma zona de frequências bastante relevante na identificação das vogais /ɛ/ e /i/.

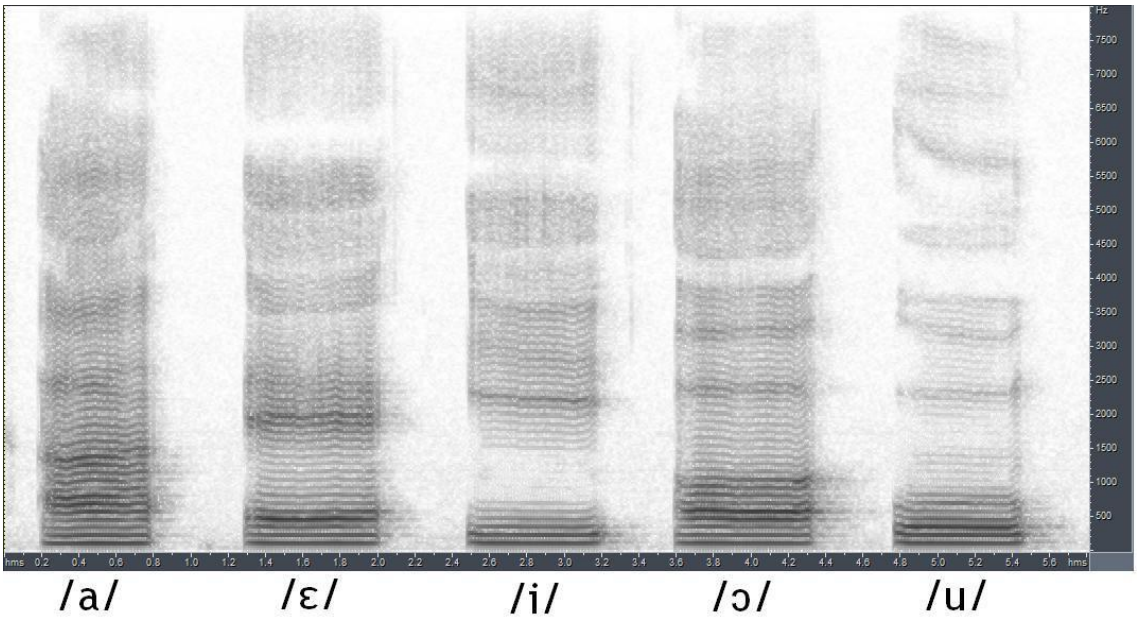


Figura 5.1 – Espectrograma das vogais /a/, /ɛ/, /i/, /ɔ/ e /u/ do sinal original do orador masculino.

As diferenças verificadas na percepção dos fonemas /ɛ/ e /ɔ/ nos oradores masculino e feminino parecem indicar que a importância na identificação das vogais não são os números dos parciais alterados, mas a região de frequências a que as sinusóides modificadas pertencem. No orador masculino os parciais 11 a 15 correspondem aproximadamente à região de frequências 1250 – 1700Hz e os parciais 16 a 30, como já foi referido, à região 1850 – 3450Hz, pois a sua frequência fundamental é 115Hz, enquanto que no orador feminino

correspondem às regiões 1550 – 2100Hz e 2250 – 4200hz, respectivamente, pois a sua frequência fundamental é de 140Hz. Esta observação explica as alterações verificadas entre os dois oradores, pois a região de frequências 1800 - 2100Hz importante para o fonema /ε/ para o orador masculino corresponde aos parciais 16 a 30 e no feminino aos parciais 11 a 15.

Não são perceptíveis diferenças entre as diferentes ressínteses com a fase alterada, independentemente do valor da fase inicial, dos parciais alterados ou do sexo do orador, o que vem confirmar que a fase dos parciais não tem grande importância em termos da percepção dos sinais de voz [7]. Apesar das avaliações perceptuais realizadas aos sinais sintéticos não terem revelado diferenças entre eles, observando as formas de onda das diferentes ressínteses são notórias diferenças entre elas.

Capítulo 6

Conclusões e trabalho futuro

Neste capítulo que encerra esta dissertação apresentam-se as conclusões que foi possível retirar dos resultados obtidos aos testes efectuados e tenta-se dar resposta aos objectivos que esta dissertação pretendia atingir.

6.1 – Conclusões

Os resultados obtidos com as diferentes experiências realizadas permitem concluir que os métodos que utilizam excitação por ruído branco não produzem resultados satisfatórios, tanto ao nível da qualidade do sinal sintetizado, como por vezes da própria inteligibilidade, pois a componente periódica dos sinais vozeados tem uma enorme importância em termos de percepção auditiva. Estes métodos, por outro lado, conseguem caracterizar os sinais não-vozeados sem grandes alterações na percepção e, dada a sua enorme vantagem em termos de eficiência na utilização dos recursos, pode fazer sentido utilizar um sistema híbrido em que as porções não-vozeadas dos sinais de fala sejam sintetizadas com excitação de ruído branco. Este princípio geral é actualmente usado em diversos algoritmos de codificação de voz e áudio [5]. Para que isso seja possível sem perdas na qualidade da ressíntese, há no entanto que ter um grande cuidado na correcta identificação das *frames* vozeadas e não-vozeadas e há também que encontrar melhores soluções para a análise e síntese de fonemas com características plosivas, ou seja, de fonemas que sofrem variações acentuadas e de duração muito curta.

As experiências de quantização do resíduo do método LPC revelam resultados interessantes, pois utilizando 4 ou mesmo 3 bit por amostra no resíduo obtêm-se ressínteses com alterações pouco significativas na percepção e uma poupança grande comparando com os 16 bit por amostra dos sinais de voz originais. Esta evidência confirma o grande sucesso da abordagem LPC na codificação de sinais de fala [5].

Os testes que envolveram a análise de fonemas específicos foram também de uma grande relevância, pois permitiram solidificar as conclusões extraídas dos testes da inteligibilidade e da qualidade das ressínteses e estabeleceram uma relação próxima entre a natureza fonética e o tipo de alterações ocorridas nos sinais sintetizados.

O estudo das vogais teve menos profundidade do que estava inicialmente previsto, pois não foi possível identificar exaustivamente as frequências típicas de todas as vogais analisadas e também não foi possível determinar se existem formantes que, se presentes, tornam a vogal imediatamente identificável. Os testes permitiram no entanto concluir que são as frequências e não os parciais da estrutura harmónica que tornam as vogais diferentes entre si. Foi possível também saber as regiões de frequências que têm mais impacto nas características auditivas das vogais em geral e algumas das regiões que tornam determinadas vogais únicas.

6.2 – Trabalho futuro

No seguimento do estudo que foi realizado nesta dissertação e dos resultados obtidos o trabalho futuro deve incidir no melhoramento da caracterização das consoantes que tenham uma fonação mais "explosiva" (plosivas vozeadas, plosivas não-vozeadas e fonema aspirado /h/). O estudo das vogais deve continuar a ser aprofundado, com o objectivo de identificar as características que diferenciam as vogais umas das outras. Este estudo deve também ser alargado às restantes vogais da língua portuguesa que não foram estudadas nesta dissertação.

ANEXO A

```
function [aCoeff,resid,pitch,G,parcor,stream,m_n,m_x] =
proclpc(data,sr,L,fr,fs,preemp,q)
%
% This code was graciously provided by:
%   Delores Etter (University of Colorado, Boulder) and
%   Professor Geoffrey Orsak (Southern Methodist University)
% It was first published in
%   Orsak, G.C. et al. "Collaborative SP education using the Internet
and
%   MATLAB" IEEE SIGNAL PROCESSING MAGAZINE  Nov. 1995. vol.12, no.6,
pp.
%   23-32.
% Modified and debugging plots added by Kate Nguyen and Malcolm Slaney

% (c) 1998 Interval Research Corporation

% Modified by Bártolo Maia - 2010

if (nargin<3), L = 13; end
if (nargin<4), fr = 20; end
if (nargin<5), fs = 30; end
if (nargin<6), preemp = .9378; end

[row col] = size(data);
if col==1 data=data'; end

nframe = 0;
msfr = round(sr/1000*fr);           % Convert ms to samples
msfs = round(sr/1000*fs);           % Convert ms to samples
duration = length(data);
speech = filter([1 -preemp], 1, data)';     % Preemphasize speech
msoverlap = msfs - msfr;
ramp = [0:1/(msoverlap-1):1]';          % Compute part of window

 for frameIndex=1:msfr:duration-msfs+1
  frameData = speech(frameIndex:(frameIndex+msfs-1
  nframe = nframe+1;
  autoCor = xcorr(frameData);  % Compute the cross correlation
  autoCorVec = autoCor(msfs+[0:L]);

                        % Levinson's method
  err(1) = autoCorVec(1);
  k(1) = 0;
  A = [];
  for index=1:L
    numerator = [1 A.']*autoCorVec(index+1:-1:2);
    denominator = -1*err(index);
    k(index) = numerator/denominator;       % PARCOR coeffs
    A = [A+k(index)*flipud(A); k(index)];
    err(index+1) = (1-k(index)^2)*err(index);
  end

  aCoeff(:,nframe) = [1; A];
  parcor(:,nframe) = k';
```

```
  errSig = filter([1 A'],1,frameData);      % find excitation noise

  G(nframe) = sqrt(err(L+1));           % gain
  autoCorErr = xcorr(errSig); % calculate pitch & voicing information

  resid(:,nframe) = errSig/G(nframe);

  [m_x(nframe),imax] = max(resid(:,nframe));
  [m_n(nframe),imin] = min(resid(:,nframe));
  resid(:,nframe) = round((resid(:,nframe)-m_n(nframe))/(m_x(nframe)-
m_n(nframe))*(2^q-1));

  if(frameIndex==1)             % add residual frames using a
    stream = resid(1:msfr,nframe);      % trapezoidal window
  else
    stream = [stream;
              overlap+resid(1:msoverlap,nframe).*ramp;
              resid(msoverlap+1:msfr,nframe)];
  end
  if(frameIndex+msfr+msfs-1 > duration)
    stream = [stream; resid(msfr+1:msfs,nframe)];
  else
    overlap = resid(msfr+1:msfs,nframe).*flipud(ramp);
  end
end
stream = filter(1, [1 -preemp], stream)';
```

ANEXO B

```
function [aCoeff,resid,pitch,G,parcor,stream] =
proclpc(data,sr,L,fr,fs,preemp)
%
% This code was graciously provided by:
%   Delores Etter (University of Colorado, Boulder) and
%   Professor Geoffrey Orsak (Southern Methodist University)
% It was first published in
%   Orsak, G.C. et al. "Collaborative SP education using the Internet
and
%   MATLAB" IEEE SIGNAL PROCESSING MAGAZINE  Nov. 1995. vol.12, no.6,
pp.
%   23-32.
% Modified and debugging plots added by Kate Nguyen and Malcolm Slaney

% A more complete set of routines for LPC analysis can be found at
%   http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html

% (c) 1998 Interval Research Corporation


% Modified by Bártolo Maia - 2010


if (nargin<3), L = 13; end
if (nargin<4), fr = 20; end
if (nargin<5), fs = 30; end
if (nargin<6), preemp = .9378; end

[row col] = size(data);
if col==1 data=data'; end

nframe = 0;
msfr = round(sr/1000*fr);           % Convert ms to samples
msfs = round(sr/1000*fs);           % Convert ms to samples
duration = length(data);
speech = filter([1 -preemp], 1, data)';     % Preemphasize speech
msoverlap = msfs - msfr;
ramp = [0:1/(msoverlap-1):1]';          % Compute part of window


for frameIndex=1:msfr:duration-msfs+1
  frameData = speech(frameIndex:(frameIndex+msfs-1));
  nframe = nframe+1;
  autoCor = xcorr(frameData); % Compute the cross correlation
  autoCorVec = autoCor(msfs+[0:L]);

                        % Levinson's method
  err(1) = autoCorVec(1);
  k(1) = 0;
  A = [];
  for index=1:L
    numerator = [1 A.']*autoCorVec(index+1:-1:2);
    denominator = -1*err(index);
    k(index) = numerator/denominator;       % PARCOR coeffs
    A = [A+k(index)*flipud(A); k(index)];
```

```
    err(index+1) = (1-k(index)^2)*err(index);
  end

  aCoeff(:,nframe) = [1; A];
  parcor(:,nframe) = k';

  errSig = filter([1 A'],1,frameData);      % find excitation noise

  G(nframe) = sqrt(err(L+1));           % gain
  autoCorErr = xcorr(errSig);% calculate pitch & voicing information

  resid(:,nframe) = errSig/G(nframe);

  tam = size(resid(:,nframe));
  sigma = sqrt(var(resid(:,nframe)));
  media = mean(resid(:,nframe));
  resid(:,nframe) = normrnd(media,sigma,tam);

  if(frameIndex==1)             % add residual frames using a
    stream = resid(1:msfr,nframe);      % trapezoidal window
  else
    stream = [stream;
              overlap+resid(1:msoverlap,nframe).*ramp;
              resid(msoverlap+1:msfr,nframe)];
  end
  if(frameIndex+msfr+msfs-1 > duration)
    stream = [stream; resid(msfr+1:msfs,nframe)];
  else
    overlap = resid(msfr+1:msfs,nframe).*flipud(ramp);
  end
end
stream = filter(1, [1 -preemp], stream)';
```

ANEXO C

```
function [aCoeff,resid,pitch,G,parcor,stream] =
proclpc(data,sr,L,fr,fs,preemp)

%
% This code was graciously provided by:
%   Delores Etter (University of Colorado, Boulder) and
%   Professor Geoffrey Orsak (Southern Methodist University)
% It was first published in
%   Orsak, G.C. et al. "Collaborative SP education using the Internet
and
%   MATLAB" IEEE SIGNAL PROCESSING MAGAZINE  Nov. 1995. vol.12, no.6,
pp.
%   23-32.
% Modified and debugging plots added by Kate Nguyen and Malcolm Slaney

% A more complete set of routines for LPC analysis can be found at
%   http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html

% (c) 1998 Interval Research Corporation

% Modified by Bártolo Maia - 2010

if (nargin<3), L = 13; end
if (nargin<4), fr = 20; end
if (nargin<5), fs = 30; end
if (nargin<6), preemp = .9378; end

[row col] = size(data);
if col==1 data=data'; end

nframe = 0;
msfr = round(sr/1000*fr);           % Convert ms to samples
msfs = round(sr/1000*fs);           % Convert ms to samples
duration = length(data);
speech = filter([1 -preemp], 1, data)';     % Preemphasize speech
msoverlap = msfs - msfr;
ramp = [0:1/(msoverlap-1):1]';          % Compute part of window

for frameIndex=1:msfr:duration-msfs+1
  frameData = speech(frameIndex:(frameIndex+msfs-1));
  nframe = nframe+1;
  autoCor = xcorr(frameData);  % Compute the cross correlation
  autoCorVec = autoCor(msfs+[0:L]);

                          % Levinson's method
  err(1) = autoCorVec(1);
  k(1) = 0;
  A = [];
  for index=1:L
    numerator = [1 A.']*autoCorVec(index+1:-1:2);
    denominator = -1*err(index);
    k(index) = numerator/denominator;       % PARCOR coeffs
    A = [A+k(index)*flipud(A); k(index)];
    err(index+1) = (1-k(index)^2)*err(index);
  end
```

```
  aCoeff(:,nframe) = [1; A];
  parcor(:,nframe) = k';


  errSig = filter([1 A'],1,frameData);      % find excitation noise

  G(nframe) = sqrt(err(L+1));           % gain
  autoCorErr = xcorr(errSig); % calculate pitch & voicing information

%  figure(1);
%  plot(autoCorErr);
  [B,I] = findpeaks(autoCorErr(1:msfs),'minpeakdistance',50);
  if isempty(I) == 0 && length(I) >= 2
    if abs(msfs - I(end))*2-2 <= abs(msfs - I(end-1)) && abs(msfs -
I(end))*2+2 >= abs(msfs - I(end-1))
        pitch(nframe) = abs(msfs - I(end));
    elseif length(B) >= 3 && abs(msfs - I(end-1))*2-2 <= abs(msfs -
I(end-2)) && abs(msfs - I(end-1))*2+2 >= abs(msfs - I(end-2))
        pitch(nframe) = abs(msfs - I(end-1));
    else
        pitch(nframe) = 0;
    end
  else
    pitch(nframe) = 0;
  end
    resid(:,nframe) = errSig/G(nframe);
  %  figure(2);
  if pitch(nframe) == 0
     ppitch(nframe) = pitch(nframe);
  else
     ppitch(nframe) = msfs*1000/(pitch(nframe)*fs);
  end
%  plot(ppitch);
%  figure(3);
%  plot(resid(:,nframe));
  if(pitch(nframe) > 0)
     resid(:,nframe) = resid(:,nframe);
  else
       tam = size(resid(:,nframe));
       sigma = sqrt(var(resid(:,nframe)));
       media = mean(resid(:,nframe));
       resid(:,nframe) = normrnd(media,sigma,tam);
  end
    if(frameIndex==1)             % add residual frames using a
    stream = resid(1:msfr,nframe);      % trapezoidal window
  else
    stream = [stream;
              overlap+resid(1:msoverlap,nframe).*ramp;
              resid(msoverlap+1:msfr,nframe)];
  end
  if(frameIndex+msfr+msfs-1 > duration)
    stream = [stream; resid(msfr+1:msfs,nframe)];
  else
    overlap = resid(msfr+1:msfs,nframe).*flipud(ramp);
  end
end
stream = filter(1, [1 -preemp], stream)';
```

ANEXO D

```
function [aCoeff,resid,pitch,G,parcor,stream] =
proclpc(data,sr,L,fr,fs,preemp)

%
% This code was graciously provided by:
%   Delores Etter (University of Colorado, Boulder) and
%   Professor Geoffrey Orsak (Southern Methodist University)
% It was first published in
%   Orsak, G.C. et al. "Collaborative SP education using the Internet
and
%   MATLAB" IEEE SIGNAL PROCESSING MAGAZINE  Nov. 1995. vol.12, no.6,
pp.
%   23-32.
% Modified and debugging plots added by Kate Nguyen and Malcolm Slaney
% A more complete set of routines for LPC analysis can be found at
%   http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html
% (c) 1998 Interval Research Corporation

% Modified by Bártolo Maia - 2010

if (nargin<3), L = 13; end
if (nargin<4), fr = 20; end
if (nargin<5), fs = 30; end
if (nargin<6), preemp = .9378; end

[row col] = size(data);
if col==1 data=data'; end
nframe = 0;
msfr = round(sr/1000*fr);           % Convert ms to samples
msfs = round(sr/1000*fs);           % Convert ms to samples
duration = length(data);
speech = filter([1 -preemp], 1, data)';     % Preemphasize speech
msoverlap = msfs - msfr;
ramp = [0:1/(msoverlap-1):1]';          % Compute part of window
for frameIndex=1:msfr:duration-msfs+
    frameData = speech(frameIndex:(frameIndex+msfs-1));
    nframe = nframe+1;
    autoCor = xcorr(frameData); % Compute the cross correlation
    autoCorVec = autoCor(msfs+[0:L]);
                        % Levinson's method
    err(1) = autoCorVec(1);
    k(1) = 0;
    A = [];
    for index=1:L
        numerator = [1 A.']*autoCorVec(index+1:-1:2);
        denominator = -1*err(index);
        k(index) = numerator/denominator;       % PARCOR coeffs
        A = [A+k(index)*flipud(A); k(index)];
        err(index+1) = (1-k(index)^2)*err(index);
    end
    aCoeff(:,nframe) = [1; A];
    parcor(:,nframe) = k';
    errSig = filter([1 A'],1,frameData); % find excitation noise
    G(nframe) = sqrt(err(L+1));       % gain
    autoCorErr = xcorr(errSig); %calculate pitch & voicing information
```

```
    [B,I] = findpeaks(autoCorErr(1:msfs),'minpeakdistance',50);
    if isempty(I) == 0 && length(I) >= 2
        if abs(msfs - I(end))*2-2 <= abs(msfs - I(end-1)) && abs(msfs
- I(end))*2+2 >= abs(msfs - I(end-1))
            pitch(nframe) = abs(msfs - I(end));
        elseif length(B) >= 3 && abs(msfs - I(end-1))*2-2 <= abs(msfs
- I(end-2)) && abs(msfs - I(end-1))*2+2 >= abs(msfs - I(end-2))
            pitch(nframe) = abs(msfs - I(end-1));
        else
            pitch(nframe) = 0;
        end
    else
        pitch(nframe) = 0;
    end
    resid(:,nframe) = errSig/G(nframe);
    if pitch(nframe) == 0
        ppitch(nframe) = pitch(nframe);
    else
        ppitch(nframe) = msfs*1000/(pitch(nframe)*fs);
    end
    mascara = zeros(msfs,1);
    if(pitch(nframe) > 0)
        inc = ceil(pitch(nframe)/4);
        hil=abs(hilbert(resid(:,nframe)));
        [b,a]=butter(6,700/(sr/2));
        h=filter(b,a,hil);
        np=ceil(msfs/pitch(nframe));
        aux=zeros(msfs,1);
        aux=h;
        [B1,I1] =
findpeaks(aux,'minpeakdistance',ceil(pitch(nframe)*2/3),'npeaks',np);
        for i=1:max(size(I1))
            for j = -floor(inc/2):floor(inc/2)
                if I1(1)-floor(inc/2)<1
                    j=-I1(1)+1;
                end
                mascara(I1(i)+j,1)=1;
                if mascara(msfs) == 1
                    break;
                end
            end
        end
        resid(:,nframe) = resid(:,nframe) .* mascara;
    else
        resid(:,nframe) = resid(:,nframe);
    end
    if(frameIndex==1)                % add residual frames using a
        stream = resid(1:msfr,nframe);      % trapezoidal window
    else
        stream = [stream;
        overlap+resid(1:msoverlap,nframe).*ramp;
        resid(msoverlap+1:msfr,nframe)];
    end
    if(frameIndex+msfr+msfs-1 > duration)
        stream = [stream; resid(msfr+1:msfs,nframe)];
    else
        overlap = resid(msfr+1:msfs,nframe).*flipud(ramp);
    end
end
stream = filter(1, [1 -preemp], stream)';
```

ANEXO E

```
function [cepstra,aspectrum,pspectrum] = melfcc(samples, sr, varargin)

http://www.ee.columbia.edu/~dpwe/resources/matlab/rastamat/mfccs.html
%
% 2005-04-19 dpwe@ee.columbia.edu after rastaplp.m.
% Uses Mark Paskin's process_options.m from KPMtools

if nargin < 2;    sr = 16000;     end

% Parse out the optional arguments
[wintime, hoptime, numcep, lifterexp, sumpower, preemph, dither, ...
 minfreq, maxfreq, nbands, bwidth, dcttype, fbtype, usecmp,
modelorder] = ...
    process_options(varargin, 'wintime', 0.025, 'hoptime', 0.010, ...
          'numcep', 13, 'lifterexp', 0.6, 'sumpower', 1, 'preemph',
0.97, ...
      'dither', 0, 'minfreq', 0, 'maxfreq', 4000, ...
      'nbands', 40, 'bwidth', 1.0, 'dcttype', 2, ...
      'fbtype', 'mel', 'usecmp', 0, 'modelorder', 0);
 if preemph ~= 0
  samples = filter([1 -preemph], 1, samples);
end
 % Compute FFT power spectrum
pspectrum = powspec(samples, sr, wintime, hoptime, dither);

aspectrum = audspec(pspectrum, sr, nbands, fbtype, minfreq, maxfreq,
sumpower, bwidth);

if (usecmp)
  % PLP-like weighting/compression
  aspectrum = postaud(aspectrum, maxfreq, fbtype);
end

if modelorder > 0

  if (dcttype ~= 1)
    disp(['warning: plp cepstra are implicitly dcttype 1 (not ',
num2str(dcttype), ')']);
  end
   % LPC analysis
  lpcas = dolpc(aspectrum, modelorder);
   % convert lpc to cepstra
  cepstra = lpc2cep(lpcas, numcep);
else
    % Convert to cepstra via DCT
  cepstra = spec2cep(aspectrum, numcep, dcttype);

end

cepstra = lifter(cepstra, lifterexp);
```

ANEXO F

```
%-------------------------------------------------------------------------
%
% This MATLAB implements the proof-of-concept of a frequency domain
pitch
% shifter
%
% This code is property of SEEGNAL Research, Lda.
% (c)2007 SEEGNAL
%
% Anibal Ferreira, University of Porto, PORTUGAL / SEEGNAL Research
% April 2nd, 2009
%
%-------------------------------------------------------------------------
close all;
clear all;

% flag indicating there is no frame history

sync=0;
%
% input audio file (raw PCM)
%
inpfile = 'audio/ricardo24/ricardo24.wav';
outfile = 'audio/ricardo24/mag/ricardo24_mag_1-5_02.pcm';
%
% N        = size of the ODFT and MDCT transforms
% win      = sine window
%
N=1024; N2=N/2; N4=N/4;
win=sin(pi/N*([0:N-1]+0.5));
FS=32000;
order = 18; % order of LPC / cepstrum spectral envelope model

data   = zeros(1,N);     % input data
idata   = zeros(1,N);    % input (int) data
odata   = zeros(1,N);    % output (int) data
osdata  = zeros(1, N);   % buffered data
tmpdata = zeros(1, N2);
fdata   = zeros(1,N);    % (float complex) data
ofdata  = zeros(1,N);    % (float complex) data
ntonaltrack = zeros(1,1000);    % MAX is 1000 frames
tonaltrack  = zeros(1000,300); % MAX is 1000 frame segments x 300
partials in each frame

% number of side lobes to be synthesized
sidelobes = 3; % three sidelobes (9 bins per sinusoid in total)

nmaxpartials = 0;

% number of partials that are synthesized on a voiced frame
numpartials=30; % 30 should be sufficient
change  = zeros(1,numpartials);
% specify here which partials will have synthetic phase, e.g.
```

```
 change(1)=1;
 change(2)=1;
 change(3)=1;
 change(4)=1;
 change(5)=1;

decaydB = 10.0; % dB decay of vanishing sinusoids

% OLDPHI retains the phase of the previous ODFT frame (partials only)
oldphi    = zeros(1,N2);
% OLDMAG retains the magnitude of the previous ODFT frame (partials
only)
oldmag    = zeros(1,N2);

% vectores de memória
oldell   = zeros(1,N2);
olddeltaell   = zeros(1,N2);

%
%  complex vectors for the non-optimal computation of the ODFT, and
IODFT transforms
%
direxp = exp(-i*pi*[0:(N-1)]/N);
invexp = exp( i*pi*[0:(N-1)]/N);
%-------------------------------------------------------------
%
% read audio file
%
fidr = fopen(inpfile,'r');
fidw = fopen(outfile,'w');
%
% begin overlap-add
%
[tmpdata, nread] = fread(fidr, N2, 'short');
data(1,N2+1:N)=tmpdata(1:N2,1).';

k=0; % set frame counter

while(nread==N2),
%
% overlap/add analysis, ODFT
%
    k = k+1; % frame counter
    figure(1); % displays time signal before windowing
    plot([0:N-1]*1000.0/FS, data(1:N));
    xlabel('Time (ms) \rightarrow');
    ylabel('Amplitude \rightarrow');
    idata=data.*win;
    fdata = idata.*direxp; % this is sub-optimal ODFT computation
    odft=fft(fdata);       % this is sub-optimal ODFT computation
    phaseodft = angle(odft(1:N2));
    powerodft = 1E-6+abs(odft).^2;
    envodft=10*log10(powerodft);
%
% looks in the spectrum for the most prominent harmonic structure, if
any
%
% NOTE 1: searchtonal() is a compiled (MEX) C function
% NOTE 2: althougth interface is prepared for two harmonic structures,
only the first is effectively used
```

```
%
     % estranhamente esta rotina searchtonal() tem que ser colocada
aqui senão
    % dá erro, parece ser erro do Matlab que não quer uso de gráficos
antes
    [npartials0 missing0 posstop0 f0pitch0 npartials1 missing1
posstop1 f0pitch1] = searchtonal(FS, N2, envodft, powerodft); % N
resolution
    figure(2);
    plot([0:N2-1]*FS/N, envodft(1:N2),'b'); % displays original
spectrum
    xlabel('Frequency (Hz) \rightarrow');
    Ylabel('Magnitude (dB) \rightarrow');
    axis([0 N2*FS/N 0 130])
    %
% ODFT smooth spectral envelope model based on cepstrum
% ODFT -> ABS -> LOG -> IODFT -> ShortPassLifter -> ODFT
%
    cepstrum=ifft(envodft);        % this is sub-optimal IODFT
computation
    cepstrum=cepstrum.*invexp;     % quefrency domain
    limite=order;
    cepstrum(1+limite:N-limite+1)=0;  % ideal short-pass lifter
    envceps=cepstrum.*direxp;      % this is sub-optimal ODFT
computation
    envceps=real(fft(envceps));    % this is sub-optimal ODFT
computation
    figure(2);
    hold on;
    plot([0:N2-1]*FS/N, envceps(1:N2),'k');
    hold off;

    edge=0.5; % 5 dB % modficado em 2set07 para aumentar muito a
sensibilidade
    maxima = []; minima = [];
    %
    % finds all relevant peaks and valleys in the spectrum
    %
    [maxima, minima] = peaksvalleys(envodft, N2, edge, edge);
    %
    % find delta ell of all maxima
    %
    deltaell = [];
    if (length(maxima) >1) deltaell=getdeltaell2(powerodft, maxima);
end
    disp('Number of partials of the harmonic structure');
    npartials0
    disp('Number of missing partials in the harmonic structure');
    missing0
    if (missing0>0)
        disp('Position of first missing partial');
        posstop0
    end
    disp('PITCH of the harmonic structure');
    f0pitch0

    if (npartials0 >0)
        nmaxpartials = floor((N2-sidelobes)/f0pitch0); % maximum
number of partials in the full spectrum
    else
        nmaxpartials = 0;
```

```
    end
    %
% now identify the natural partials that are closest to the ideal
harmonic
% alignement **   this is based on maxima() and deltaell()   ****
%
     trueell = []; truedeltaell = []; truemag = []; truephi = [];
    if (npartials0>0)
        truepeaks = min(nmaxpartials, min(length(maxima)-
2,numpartials)); % 30 should be sufficient
    else
        truepeaks = 0;
    end
    % esta rotina tem o problema de poder não encontrar facilmente o
    % verdadeiro pico quando ele está mergulhado em ruído, como tornar
mais
    % robusto ?
    pointer=1;
    for m=1:truepeaks,
        dtmp = m * f0pitch0;
        % tenta colocar pointer no parcial real à direita do parcial
ideal
        % aqui em searchtonal pode-se aproveitar o vector dos parciais
        % já pesquisados
        while (pointer < length(maxima) && maxima(pointer)-
1+deltaell(pointer) <= dtmp),
            pointer=pointer+1;
        end
        difaft = abs(maxima(pointer)-1+deltaell(pointer)-dtmp);
        if (pointer>1)
            difbef = abs(dtmp-maxima(pointer-1)+1-deltaell(pointer-
1)); % HAVIA ERRO !
        else
            difbef = 1E4; % just a large number
        end
        if ( difaft < difbef )
             trueell(m) = maxima(pointer);
             truedeltaell(m) = deltaell(pointer);
%              pointer = pointer + 1;
        else
             trueell(m) = maxima(pointer-1);
             truedeltaell(m) = deltaell(pointer-1);
        end
    end
    % fase diferencial
    difphi = zeros(200,2);
    if (truepeaks>0)
         truephi = getphase2(phaseodft, trueell, N);
         truemag = getPSD(envodft, trueell, truedeltaell, N);
         % só se calcula o diferencial de fase para antes e depois,
será
         % necessário para cacular para os 4 bins antes e 4 bins
depois ?
         % NAO!, induz maior erro nos bins mais laterais
         difphi = difphase2(phaseodft, trueell, N);
         figure(2);
         hold on;
         plot( (trueell-1)*FS/N, 1+envodft(trueell),'rv'); %
identifies peaks graphically
         hold off;
    end
```

```
    truedata  = zeros(1,N); % make sure we start with an empty buffer
    if (truepeaks>0)
        if (sync==1) % there is history
            for s=1:truepeaks, % for all tonals in the current frame
                if (change(s)==1)

dtmp=0.5*(oldell(s)+olddeltaell(s)+trueell(s)+truedeltaell(s))-1.0;
                    tmpphi=mainarg(oldphi(s)+2*pi/N*dtmp*N/2); % to
facilitate readability
                    truemag(s)=truemag(s)*0.2;
                else
                    tmpphi=truephi(s);
                end
                tmpdata = syntonal2(truedata, trueell(s),
truedeltaell(s), truemag(s), tmpphi, difphi(s,:), N, sidelobes);
                truedata = tmpdata;
                oldphi(s)=tmpphi;
                oldmag(s)=truemag(s);
                oldell(s)=trueell(s);
                olddeltaell(s)=truedeltaell(s);
            end
        else %there is no history
            for s=1:truepeaks, % for all tonals in the current frame
                tmpphi=truephi(s); % we may reset this value at
sinusoidal birth
%                 tmpphi=pi/4*s;
                truemag(s)=truemag(s)*0.2;
                tmpdata = syntonal2(truedata, trueell(s),
truedeltaell(s), truemag(s), tmpphi, difphi(s,:), N, sidelobes);
                truedata = tmpdata;
                oldphi(s)=tmpphi;
                oldmag(s)=truemag(s);
                oldell(s)=trueell(s);
                olddeltaell(s)=truedeltaell(s);
            end
        end
        sync=1;
        sussurro = odft - truedata; % sinal diferença (i.e. ruído)
        figure(2);
        hold on
        plot([0:N2-1]*FS/N, 20*log10(abs(sussurro(1:N2))),'g'); %
displays residual spectrum
        axis([0 N2*FS/N 0 130])
        hold off
    else
        sync=0;
        truedata = odft;
        oldphi   = zeros(1,N2);
        oldmag   = zeros(1,N2);
        oldell   = zeros(1,N2);
        olddeltaell   = zeros(1,N2);
    end
    %
    % output data, pick one: the signal or the residual
    %
    fdata(1:N2) = truedata(1:N2); % este é o sintetizado harmónico
%     fdata(1:N2) = sussurro(1:N2); % este é o sinal resíduo
%
% For visualization purposes
%
%
```

```
    graphell = []; graphdeltaell = [];
    ntonaltrack(k) = truepeaks; % stores number of spectral peaks
found in current frame
    for m=1:truepeaks,
        graphell(m) = trueell(m);
        graphdeltaell(m) = truedeltaell(m);
    end

%-------------------------------------------------------------------
------------------------------
% Depicts detected partials
%-------------------------------------------------------------------
------------------------------

    if (ntonaltrack(k) > 0) % if there are any relevant maxima
            tonaltrack(k,1:ntonaltrack(k)) = (graphell+graphdeltaell-
1.0); % stores accurate bin frequency of each peak
            figure(3);
            hold on;
            plot(k,tonaltrack(k,1:ntonaltrack(k)),'*'); % depicts all
tonals found in current frame
            axis([0 k+1 0 200]) % 200 is arbitrary, only chosen to
represent low frequency tonals
            xlabel('Time (frames) \rightarrow');
            Ylabel('Frequency (accurate bin scale) \rightarrow');
            hold off;
     end
%
% IODFT, overlap/add reconstruction
%
    fdata(N:-1:N2+1)=conj(fdata(1:N2));
    ofdata=ifft(fdata);        % this is sub-optimal IODFT computation
    ofdata=ofdata.*invexp;     % this is sub-optimal IODFT computation
    odata=real(ofdata);
    odata=odata.*win;

    tmpdata(1,1:N2)=floor(0.5+osdata(1,1:N2)+odata(1:N2));
    osdata=odata(N2+1:N);
    fwrite(fidw, tmpdata(1,1:N2), 'short');
    data(1,1:N2)=data(1,1+N2:N);
    [tmpdata, nread] = fread(fidr, N2, 'short'); % reads new half-
segment
    if nread<N2,
        tmpdata(nread+1:N2,1)=zeros(N2-nread,1);
    end
    data(1,N2+1:N)=tmpdata(1:N2,1).';
%       pause;
end

fclose(fidr);
fclose(fidw);
disp('END of processing !');
```

Referências

[1] Apontamentos da disciplina processamento da fala do mestrado integrado em engenharia electrotécnica e de computadores, ano lectivo 2008/2009

[2] Guimarães, I. "A Ciência e a Arte da Voz Humana", ESSA – Escola Superior de Saúde do Alcoitão, 2007

[3] Teixeira, J. P. "Modelização Paramétrica de Sinais para Aplicação em Sistemas de Conversão Texto-Fala", 1995

[4] Arthur C. Guyton, John E. Hall, "Textbook of Medical Physiology", Eleventh Edition, Elsevier Saunders

[5] Aníbal Ferreira, Carlos Salema, Fernando Pereira, Isabel Trancoso, Paulo Lobato Correia, Pedro Assunção, Sérgio Faria, Comunicações Audiovisuais: Tecnologias, Normas e Aplicações, IST Press, Julho de 2009

[6] Andreas Spanias, Ted Painter, Venkatraman Atti, "Audio Signal Processing and Coding", John Wiley & Sons, Inc., Hoboken, New Jersey, 2007

[7] John R. Deller Jr., John H. L. Hansen, John G. Proakis, "Discrete-Time Processing of Speech Signals", New York: IEEE, 2000

[8] http://www.langsci.ucl.ac.uk/ipa/vowels.html. Último acesso a 30/Janeiro/2010

[9] Lawrence R. Rabiner, Ronald W. Schafer, "Digital Processing os Speech Signals", Englewodd Cliffs: Prentice-Hall, 1978

[10]Malcolm Slaney , Auditory Toolbox version 2. Disponível em http://cobweb.ecn.purdue.edu/~malcolm/interval/1998-010/. Último acesso a 30/Janeiro/2010

[11]Dan Ellis - http://labrosa.ee.columbia.edu/matlab/rastamat/. Acesso a 30/Janeiro/2010

[12] Hermansky, H. *Perceptual Linear Predictive (PLP) Analysis of Speech*, J. Acoust. Soc. Am., Abril 1990

[13]Aníbal J. S. Ferreira, "Accurate estimation in the ODFT domain of the frequency, phase and magnitude of stationary sinusoids"

[14]http://www.l2f.inesc-id.pt/~lco/ptsam/ptsam.pdf. Último acesso a 30/Janeiro/2010

[15]Florian Hönig, Georg Stemmer, Christian Hacker, Fabio Brugnara, "Revising Perceptual Linear Prediction (PLP)"

[16]Help do Matlab 2007b

[17]Melvyn J. Hunt, *Spectral Signal Processing for ASR*

[18]João Canas Ferreira, João Correia Lopes José Machado da Silva, "Norma de Formatação e Orientações para a Escrita de Dissertações ou Relatórios de Projecto do MIEEC", Maio de 2008