Faculdade de Engenharia da Universidade do Porto

Modelização de Filtro de Trato Vocal para Reconstrução de Voz Disfónica

Marco António da Mota Oliveira

Versão Final

Mestrado Integrado em Engenharia Eletrotécnica e de Computadores

Orientador: Prof. Dr. Aníbal João de Sousa Ferreira

13 de Fevereiro de 2020

Resumo

O sussurro é, como alternativa à fala normal vozeada, um importante modo da comunicação oral, em particular quando o silêncio ou a privacidade são desejáveis ou necessários. No entanto, para aqueles que sofrem de afonia temporária ou permanente, o sussurro poderá ser a única forma de comunicação oral disponível, o que causa dificuldades adicionais com potencial impacto aos níveis sociais e profissionais. Entre as soluções clássicas para este problema, encontram-se a prótese traqueoesofágica ou soluções computacionais do tipo aplicações *text-to-speech*. Estas soluções, contudo, não são suficientemente satisfatórias, apresentando claras desvantagens como o carácter invasivo de uma ou as limitações práticas da outra. Os desenvolvimentos tecnológicos recentes nas áreas do processamento digital de sinal e do reconhecimento automático de voz suscitaram um novo interesse na procura de soluções alternativas como algoritmos de melhoria ou conversão do sinal de fala sussurrada, cujos resultados não são ainda assim os desejáveis. Esta dissertação integra o projeto DyNaVoiceR, que se propõe a desenvolver uma tecnologia assistiva, não invasiva, capaz de operar em tempo-real e que permita facilitar a comunicação de uma forma confortável e eficiente a quem sofre deste problema. Como parte desse projeto, o foco desta dissertação consistiu no desenvolvimento de modelos orientados ao orador, computacionalmente eficientes, das características espectrais de vogais sussurradas, nos domínios espectral e cepstral, e de um algoritmo expedito de identificação de vogais em contexto de fala sussurrada, baseado em correlações estatísticas. Com o propósito de avaliar a capacidade do algoritmo em classificar vogais sussurradas, conduziram-se testes objetivos que utilizaram como referência a anotação manual previamente disponível na base de dados de oradores utilizada, tendo-se verificado a viabilidade do algoritmo para operação em tempo-real. Para avaliar o efeito do vozeamento de vogais em sinal de fala sussurrada, conduziram-se experiências de vozeamento automático nas vogais correspondentes à primeira sílaba das palavras dissílabas disponíveis na base de dados, tendo-se concluído pela preservação da informação linguística e pelo impacto positivo do ponto de vista percetual do vozeamento destas vogais, na generalidade dos casos testados.

Abstract

Whispered speech, as an alternative to normal voiced speech, is an important tool in oral communication, namely when silence or privacy are required or desired. However, for those that suffer from temporary or permanent aphonia, whisper is their main form of oral communication resulting in considerable communication hurdles that can impact their social and professional life. Some solutions currently exist to help dealing with these problems like the tracheoesophageal voice prosthesis or computer-based solutions such as *text-to-speech* apps. Those solutions are not entirely satisfactory, showing clear disadvantages, due to their intrusive nature or their practical limitations. Recent technological developments in digital signal processing and speech recognition areas sparked a new interest in the search of alternative solutions such as whispered speech enhancement or whispered-to-normal speech conversion algorithms, with limited results to date though. This thesis is part of the DyNaVoiceR project, that aims at developing an assistive, non-intrusive, real-time technology that allows efficient and practical communication to those with aphonia. As part of this project, the aim of this thesis is to develop speaker oriented computational efficient models of the whispered vowels spectral characteristics, on both spectral and cepstral domains. An algorithm is also developed in order to identify whispered vowels based on statistical correlations. In order to analyse the algorithm ability to classify whispered vowels, tests were carried out that used the speaker's database manual annotation as reference. These tests confirmed the real time algorithm viability. In order to evaluate the voicing effect of whispered vowels, experiments have been conducted that consisted in the automatic voicing of first syllable vowels in the database duosyllable words. It has been concluded that for most cases the voicing of whispered vowels has been positively perceived and the linguist information retained.

Agradecimentos

Quero agradecer a todos aqueles que com o seu incentivo, apoio ou colaboração tornaram esta dissertação possível ou para ela contribuíram.

Em primeiro lugar, ao Professor Doutor Aníbal João de Sousa Ferreira, pelo convite para a sessão de lançamento do projeto DyNaVoiceR, pela oportunidade dada de ingressar no mesmo, pelo acolhimento, pelo apoio, pela orientação, pelas sugestões e correções na organização do documento e pela compreensão em relação às circunstâncias de realização deste trabalho.

Aos meus colegas no projeto. Em especial, ao meu colega João Silva, companheiro de muitas horas de trabalho, pelo acolhimento ao projeto, pela colaboração nos testes percetuais, pelos comentários críticos na organização e elaboração da dissertação, pela preciosa ajuda na revisão do documento e pela paciência para as discussões filosóficas que fomos tendo. Às colegas Francisca Brito, pela colaboração e ajuda no período de acolhimento ao projeto, e Clara Cardoso, pela ajuda na revisão da dissertação.

E, acima de tudo, à minha família. Aos meus pais António e Otília e à minha irmã Diana, pelo incentivo e apoio. À minha esposa Margarida, pelo incentivo e apoio, no meio da enorme batalha que ela própria trava e pela paciência para as horas menos humoradas, quando algum cansaço também se me fazia sentir.

Dedicado aos meus filhos, Ana e Ricardo.

Marco António da Mota Oliveira

Conteúdo

Lista de Figuras

Lista de Tabelas

Abreviaturas e Acrónimos

ASR	Automatic Speech Recognition
CELP	Code Excited Linear Prediction
DCT	Discrete Cosine Transform
DFT	Discrete Fourier Transform
EL	Electrolaringe
DyNaVoiceR	Dysphonic to Natural Voice Reconstruction
ERB	Equivalent Rectangular Bandwidth
ERBS	Equivalent Rectangular Bandwidth Scale
FEUP	Faculdade de Engenharia da Universidade do Porto
FCT	Fundação para a Ciência e Tecnologia
GMM	Gaussian Mixture Models
HMM	Hidden Markov Models
IPA	International Phonetic Alphabet
LPC	Linear Predictive Coding
MELP	Mixed Excitation Linear Prediction
MFCC	Mel Frequency Cepstral Coefficient
MOS	Mean Opinion Score
NAM	Non-Audible Murmur
ODFT	Odd-stacked Discrete Fourier Transform
PE	Português Europeu padrão
PSD	Power Spectral Density
PTE	Punção Tráqueo-Esofágica
SNR	Signal-to-Noise Ratio
STD	Standard Deviation
TTS	Text-to-Speech

Capítulo 1

Introdução

Este capítulo introdutório apresenta o enquadramento da dissertação, a motivação e objetivos subjacentes ao trabalho desenvolvido, assim como a organização deste documento.

1.1 Enquadramento

Sempre que existe a necessidade de comunicar de forma relativamente silenciosa (*e.g.*, como numa sala de biblioteca ou de cinema) ou se se deseja transmitir informação de carácter privado, o modo de fala sussurrada oferece-nos um mecanismo alternativo de comunicação oral que é particularmente útil e adequado. No entanto, o modo de fala dita vozeada, ou normalmente vozeada, garante uma comunicação mais eficiente e uma maior projeção ao sinal de voz, com superiores índices de inteligibilidade, facilitando de uma forma geral a comunicação oral [7], mais prática e indicada para a generalidade dos cenários. Porém, para os pacientes que sofrem de uma forma de afonia, temporária ou permanente, causada por condições como a paralisia bilateral das pregas vocais [8], disfonia espasmódica [9] ou que tenham sido submetidos a uma laringectomia [10], estarão eventualmente limitados a falar no modo sussurrado, o que lhes introduz dificuldades adicionais na comunicação com impactos negativos seja no plano social, seja no plano profissional.

As soluções clássicas para condições como as descritas incluem a eletrolaringe (EL) [11], o recurso à voz esofágica ou procedimentos médicos invasivos, com os riscos inerentes, como a punção traqueoesofágica com prótese (PTE) [12]. Estas soluções, além de pouco práticas, resultam num discurso menos inteligível e que não é natural do ponto de vista percetual, causando desconforto e tendendo a tornar-se cansativas para o ouvinte, dificultando de tal modo uma normal conversação. Computacionalmente, encontram-se também disponíveis interfaces de fala silenciosa [13] e aplicações do tipo *text-to-speech* (TTS) [14], porém com o ónus de não operarem em tempo-real e de imporem limitações práticas à conversação. No seu conjunto, as soluções já exploradas apresentam ainda assim significativas desvantagens, o que tem levado à investigação de métodos alternativos, tentando tirar partido da

crescente capacidade computacional e das eventuais oportunidades oferecidas pelos avanços nas tecnologias de processamento de sinal.

É neste contexto que se enquadra esta dissertação, que por seu turno se inscreve no âmbito de um projeto da Fundação para a Ciência e Tecnologia (FCT): Dysphonic to Natural Voice Reconstruction (DyNaVoiceR). Este projeto foca-se no desenvolvimento de uma tecnologia assistiva avançada por forma a ajudar pacientes afetados por disfonia da voz, em particular por afonia temporária ou permanente, a comunicar de forma eficiente e confortável. Pretende-se, com este projeto, desenvolver um protótipo capaz de converter sussurro em discurso perceptualmente natural com recurso à implantação de vozeamento sintético em segmentos selecionados do sinal, preservando a informação linguística, mantendo elementos da assinatura vocal do orador e melhorando a projeção da voz.

1.2 Motivação

A impossibilidade de comunicar num registo normalmente vozeado, estando por essa forma limitado a comunicar na forma sussurrada, é uma situação com enormes repercussões negativas quer no plano social, quer no plano profissional. Mitigar este problema através de uma solução não invasiva e prática, idealmente preservando componentes da identidade sonora do orador e capaz de operar em tempo-real, por forma a facilitar a comunicação interpessoal e homem-máquina, contribuiria indubitavelmente para uma melhoria significativa da qualidade de vida daqueles que sofrem deste tipo de condição. Com o trabalho desenvolvido nesta dissertação, pretende-se contribuir com elementos importantes para o alcance dos objetivos últimos do projeto, que são de evidente utilidade social e humana.

1.3 Objetivos

Esta dissertação foca-se na análise e modelização da envolvente espectral para as nove vogais orais do português europeu padrão nos modos de fala sussurrada e de fala vozeada, no desenvolvimento de modelos representativos das características espectrais destes fonemas, orientados ao orador, e no desenvolvimento de um algoritmo protótipo de identificação de vogais sussurradas, orientado à operação em tempo-real. Estes objetivos encontram-se enquadrados com uma das prioridades do projeto DyNaVoiceR que é a conversão de vogais sussurradas em vogais vozeadas.

Dado o requisito de operacionalidade em tempo-real, esta dissertação propõe-se a desenvolver modelos compactos no domínio espectral e no domínio cepstral para as envolventes espectrais das nove vogais orais do Português Europeu padrão e um modelo de identificação de vogais em contexto de fala sussurrada que seja computacionalmente económico e eficiente.

1.4 Estrutura do Documento

A presente dissertação encontra-se organizada da seguinte forma:

- Capítulo 1 - Presente capítulo onde se apresenta o enquadramento e a motivação do trabalho desenvolvido, os objetivos e a organização deste documento;
- Capítulo 2 - Descreve os mecanismos da produção da fala e as principais características do sistema auditivo, em especial nos aspetos mais relevantes para a dissertação, bem como os principais modelos e abordagens no processamento de sinais de voz;
- Capítulo 3 - Realiza uma revisão bibliográfica às tecnologias, resultados e conclusões disponíveis na literatura, mais pertinentes para o projeto DyNaVoiceR e para o tema da dissertação;
- Capítulo 4 - Documenta a primeira etapa desenvolvida no âmbito da dissertação envolvendo a análise e síntese de vogais sussurradas com recurso a diferentes tipos de envolvente espectral. Inclui também a caracterização de uma ferramenta crucial no projeto e na dissertação, a base de dados de oradores do DyNaVoiceR;
- Capítulo 5 - Documenta a segunda etapa desenvolvida durante a dissertação envolvendo o estudo da correlação entre envolventes espectrais para nove vogais orais do Português Europeu padrão, incluídas na base de dados DyNaVoiceR, o desenvolvimento de modelos representativos destes fonemas e de um algoritmo protótipo de identificação de vogais sussurradas;
- Capítulo 6 - Documenta as primeiras experiências de vozeamento automático de segmentos de fala sussurrada com recurso à implantação de segmentos da vogal correspondente naturalmente vozeada do mesmo orador, com vista a uma avaliação preliminar de carácter percetual;
- Capítulo 7 - Encerra a dissertação, resumindo os resultados obtidos e as principais conclusões retiradas, propondo ainda tarefas e desafios futuros dando continuidade ao trabalho realizado na dissertação.

Capítulo 2

Produção da Fala

Neste capítulo descrevem-se os mecanismos de produção da fala e a fisiologia do sistema auditivo, procurando enfatizar os aspetos mais pertinentes para o trabalho desenvolvido na dissertação, introduzindo-se também os princípios de processamento de sinais de voz que serão aqui igualmente mais relevantes.

2.1 A Fala

A comunicação desempenha um papel fundamental e praticamente indispensável nos seres humanos, enquanto espécie social, encontrando-se na fala o modo que mais privilegia para transmitir aquilo que pensa e sente. O sucesso desta tão prevalente forma de comunicação oral poderá explicar-se, na perspetiva da evolução da espécie, pelas importantes vantagens que trouxe ao libertar as mãos do orador, por não requerer a visualização direta dos interlocutores ou por facultar a comunicação mesmo no escuro, bem como até pela sua relativa eficiência, ao permitir taxas de transmissão consideravelmente elevadas de até 20 a 30 segmentos (*i.e.*, símbolos) por segundo [15]. Comunicar falando é tão natural e conveniente que se tem vindo a estender à relação homem-máquina, em especial na última década, com a multiplicação dos sistemas e aplicações capazes de interpretar comandos de voz [4]. Para os interlocutores, o recurso à voz durante a produção da fala, contendo características idiossincráticas do orador, permite não só distinguir os indivíduos como transmitir parte da sua identidade individual [16]. O físico britânico Stephen Hawking, padecendo de uma doença degenerativa (esclerose lateral amiotrófica) foi vendo diminuída a capacidade de falar de uma forma inteligível até acabar por perder a voz em virtude da traqueotomia a que foi submetido em meados da década de 80. Passou então, pouco tempo depois, a comunicar através de uma aplicação gerando voz sintetizada, com a ajuda de um computador controlado por comandos musculares. Quando, anos mais tarde, surgiu a oportunidade de atualizar o então obsoleto sintetizador de voz, recusou-a. Aquela voz caracteristicamente robótica era afinal, ao cabo desses anos, a voz com que se sentia identificar [17].

Entende-se por fala a comunicação oral no modo normalmente vozeado ou no modo sussurrado (por oposição a outros como o canto ou o grito) em que se utiliza uma combinação de sons para, com recurso a uma linguagem e de forma simbólica, transmitir a mensagem pretendida. As mais pequenas unidades de sons distintos produzidos durante a fala serão então os fonemas e dividem-se essencialmente em dois grupos: as vogais e as consoantes. A faculdade de produzir estes sons é assegurada pelo aparelho fonador humano [4, 18, 19].

2.1.1 Aparelho Fonador

Tipicamente, o aparelho fonador vem descrito na literatura como se encontrando composto por três secções: *i)* a região subglótica, constituída pela traqueia, brônquios e pulmões; *ii)* a laringe, que suporta as pregas vocais, com a glote comportando o espaço entre elas; e *iii)* a região supraglótica, o trato vocal [19, 20, 21]. Esta divisão faz de resto tanto mais sentido quanto cada uma destas regiões se encontra intimamente associada a uma função fisiológica específica na produção da fala: a respiração, a fonação e a articulação, respetivamente [19]. Controlando o volume dos pulmões, graças à atuação dos músculos do diafragma e intercostais, a respiração garante os fluxos de ar e a energia necessária para a produção do sinal acústico. A constrição a que o fluxo de ar é, entretanto, sujeito na laringe atribui-lhe características turbulentas que, dependendo do posicionamento das pregas vocais, gerará um sinal de natureza essencialmente ruidosa ou periódica, possibilitando a fonação. Finalmente, a articulação do trato vocal é crucial para a considerável variedade de sons que se produzem durante a fala, como se verá de seguida em maior detalhe.

2.1.2 Trato Vocal

O trato vocal, representado na figura 2.1, corresponde a toda a região supraglótica do aparelho fonador sendo composto pelas cavidades da faringe, a nasal e a oral. Esta última revela-se a mais importante pela sua dimensão e pela disposição de diversos articuladores que possibilitam alterar de forma substancial o seu formato ao longo do tempo. Destes, destacam-se quatro articuladores que por disporem de mobilidade própria são considerados os articuladores ativos: os lábios, que controlam a radiação do sinal; a língua, o órgão de maior mobilidade no trato vocal e que permite modular a obstrução à passagem do ar; o maxilar inferior, que controla o volume da cavidade oral; e o palato mole, ou véu palatino, que controlando o acesso à cavidade nasal permite a produção de fonemas orais ou nasais. Os dentes e o palato duro são, por seu turno, exemplos de articuladores passivos [19].

O sinal acústico que é produzido durante a fala chega ao trato vocal essencialmente numa de duas formas. Quando as pregas vocais se encontram juntas, em posição de adução, a pressão subglotal provoca a passagem ritmada do ar em rápidas rajadas sucessivas que se traduzirão em pulsos sonoros com uma dada frequência. Esta frequência, a Frequência Fundamental (F0), corresponderá do ponto de vista percetual ao tom de voz (*pitch*, no inglês), que depende em larga medida da espessura e comprimento das pregas vocais;

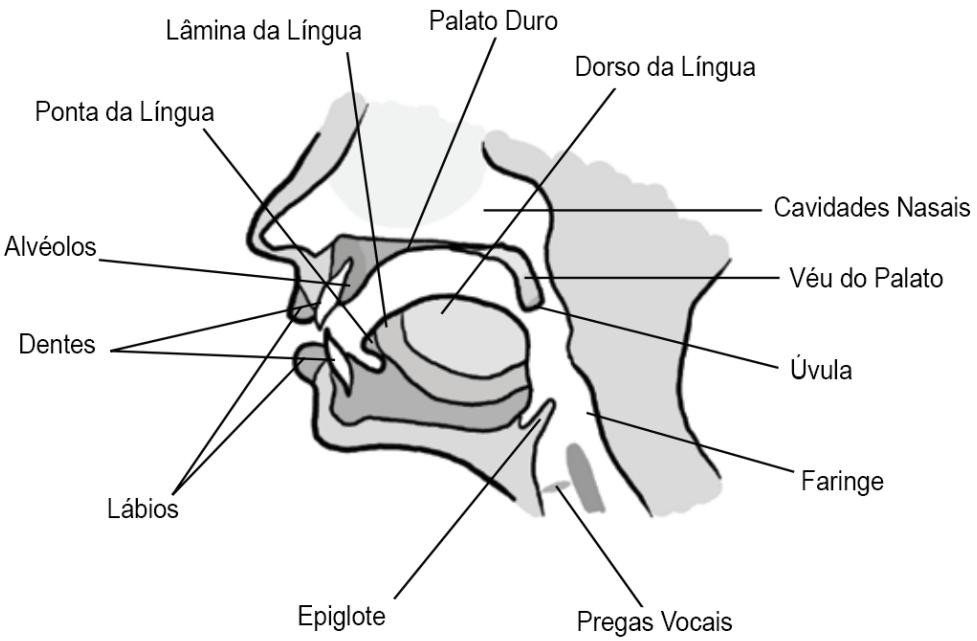


Figura 2.1: Representação da região supraglótica do aparelho fonador (*i.e.*, o trato vocal) com a indicação dos respetivos articuladores (adaptado de Seara *et al* [1]).

quando porém as pregas vocais se encontram em posição de abdução, subsiste na glote um pequeno espaço para a passagem do ar pelo que o fluxo adquire então características turbulentas que atribuem, neste caso, uma natureza ruidosa ao sinal [19, 20].

Este sinal, periódico ou ruidoso, será sujeito ao efeito do trato vocal que, operando como um tubo cuja espessura varia ao longo da sua extensão, produzirá determinadas ressonâncias e antirressonâncias naturais, atuando por essa razão como um filtro acústico capaz de moldar nas frequências o sinal que lhe chega da laringe. A figura 2.2 contém, na linha contínua a azul, uma representação no domínio das frequências de um segmento de sinal de fala de natureza periódica, como se poderá notar pelos picos espectrais correspondentes à frequência fundamental e respetivos múltiplos (os harmónicos). A linha contínua a laranja, denominada de envolvente espectral (*spectral envelope*, no inglês), corresponde a uma estimativa do efeito de filtro do trato vocal. As frequências favorecidas pelo trato vocal, refletindo as suas ressonâncias naturais, surgirão aqui como proeminências espectrais a que se dá o nome de frequências formantes, também observáveis na figura. A diversidade dos fonemas como os gerados no Português[1] é assegurada pela flexibilidade e articulação do trato vocal. O posicionamento das formantes é essencial para distinguir em particular as diferentes vogais, que são produzidas sem que o trato vocal ofereça uma obstrução signi-

[1]Nas denominadas línguas tonais, como é exemplo o Chinês, a variação da Frequência Fundamental desempenha um papel importante também na definição de diferentes fonemas e significados linguísticos [18].

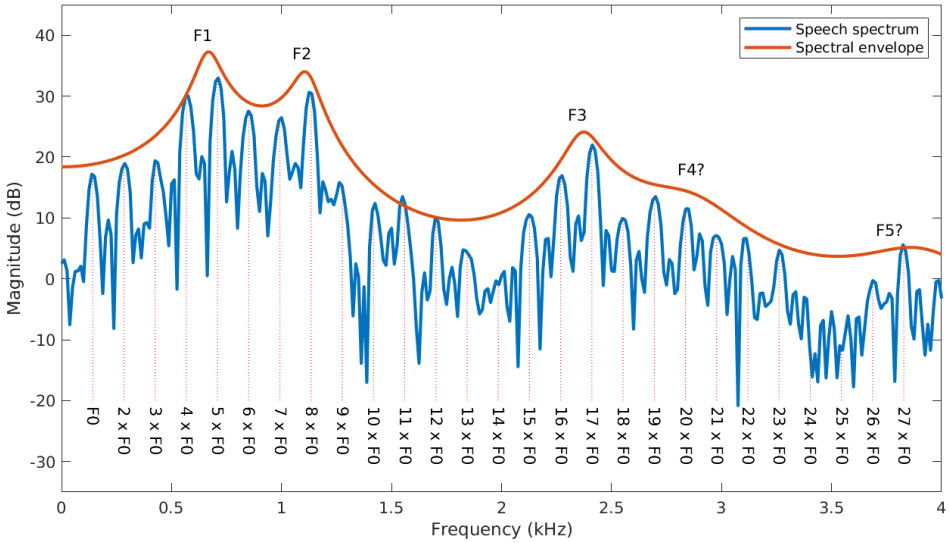


Figura 2.2: Representação de um segmento de sinal de voz no domínio das frequências, onde é visível a estrutura harmónica do sinal (linha contínua a azul) e uma estimativa da envolvente espectral (linha contínua a laranja), sendo possível identificar as principais frequências formantes (adaptado de Bäckström [2]).

ficativa à passagem do ar [22]. As formantes que mais contribuem para definir e distinguir cada uma das vogais são as duas primeiras (F1 e F2) e a terceira (F3) em menor grau [23]. As consoantes, por seu turno, caracterizam-se principalmente pela introdução de obstruções temporárias à passagem do ar, parciais ou totais, através do controlo e posicionamento dos diversos articuladores, nomeadamente a língua, os dentes e os lábios [18].

2.1.3 Triângulo Acústico das Vogais

As principais vogais do Português Europeu (PE) padrão dividem-se nas nove vogais orais, vogais estas que ocorrem quando o véu palatino bloqueia o acesso à cavidade nasal, e nas cinco vogais nasais, que são por sua vez produzidas quando este articulador baixa, facultando desse modo o acesso à cavidade nasal [19]. As nove vogais orais do PE, focadas na dissertação, encontram-se discriminadas no anexo A.1, que inclui a notação utilizada neste documento e os respetivos códigos no Alfabeto Fonético Internacional, ou *International Phonetic Alphabet* (IPA) no inglês.

O diagrama das vogais, denominado também de triângulo acústico, consiste numa conveniente representação gráfica relacionando as duas primeiras formantes, F1 e F2, as formantes mais importantes para a definição da vogal como já se viu, com outros tantos parâmetros na articulação das vogais [19]:

- a altura do dorso da língua, em que um baixo valor da formante F1 caracteriza as denominadas vogais altas e um valor elevado de F1, as vogais baixas;

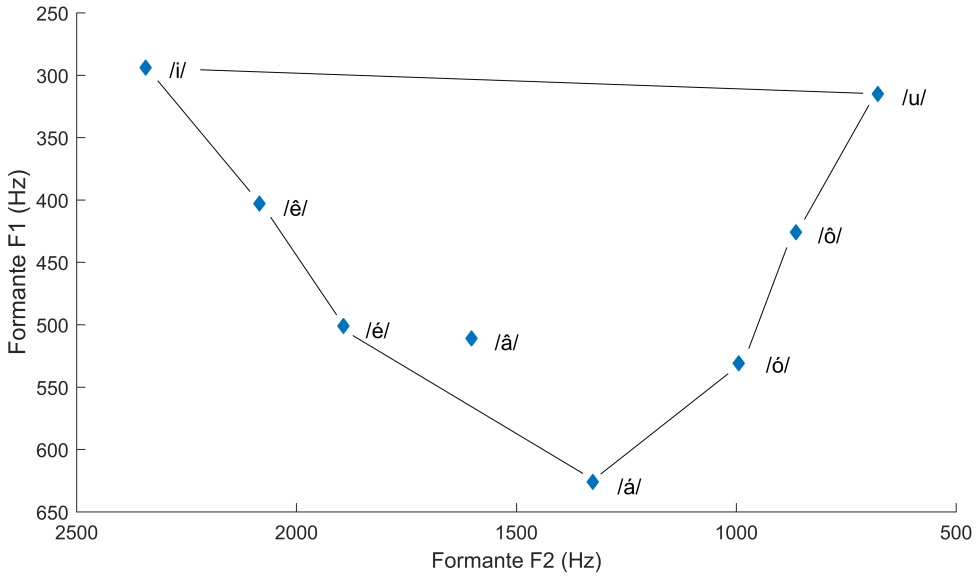


Figura 2.3: Exemplificação do triângulo acústico para as vogais orais tónicas no Português Europeu padrão (adaptado de Delgado-Martins [3]).

- o ponto de articulação, em que um baixo valor da formante F2 equivale a uma vogal posterior (*i.e.*, recuo do dorso da língua) e um valor elevado de F2 corresponde a uma vogal anterior (*i.e.*, avanço do dorso da língua).

A área formada pelos dois eixos das duas primeiras formantes é referida como o espaço das vogais. Dependendo do idioma e das vogais consideradas, se tomarmos por referência as vogais nos valores extremos das formantes F1 e F2, o diagrama corresponderá a um triângulo ou a um quadrilátero. No Português Europeu padrão poderá ser descrito como um triângulo acústico como aquele que se ilustra no exemplo da figura 2.3, onde é também possível observar a posição relativa e a distância acústica entre oito das nove vogais orais do PE no espaço das vogais[2].

2.1.4 Fala Normal e Fala Sussurrada

A fala normal (*normal speech*, no inglês), ou normalmente vozeada, corresponde à que tipicamente se utiliza na conversação normal e caracteriza-se pela presença de segmentos vozeados no sinal. Este vozeamento corresponde ao contributo das pregas vocais sempre que, como atrás se descreve, entrando num regime vibratório, atribuem características periódicas ao sinal acústico da fala. São estes segmentos em que ocorre vozeamento, identificáveis pela presença de uma estrutura harmónica como a que se vê na figura 2.2, que se denominam de vozeados. No registo de fala normal as vogais serão sempre vozeadas, podendo ocorrer também vozeamento parcial ou total durante as consoantes [18, 24, 25].

[2] A ausência da vogal /e/ de 'sede' no estudo publicado por Delgado-Martins[3], utilizado como referência, justifica-se por esta vogal nunca se encontrar em posição tónica.

A fala sussurrada (*whispered speech*, no inglês) é, por seu turno, caracterizada precisamente pela inexistência de um contributo das pregas vocais. No sussurro, a fala resulta da passagem do ar turbulento expelido dos pulmões, agora com características ruidosas apenas, pelo trato vocal que continuará a operar de forma idêntica à já descrita. Quer isto dizer que a articulação do trato vocal permitirá o estabelecimento das frequências formantes necessárias para a distinção das diferentes vogais e a introdução das obstruções necessárias à produção das diversas consoantes. Por conseguinte, se assumirmos o normal funcionamento dos articuladores do trato vocal (ver figura 2.1), o discurso sussurrado deverá estar moldado no tempo e em frequência de tal forma que é capaz, ainda assim, de transmitir a informação linguística pretendida [4, 7, 26].

Existem, porém, importantes diferenças no domínio espectral entre estes dois modos de fala e que merecem ser sublinhadas, nomeadamente na distribuição da energia nas frequências e na localização exata dos picos espectrais. A Densidade Espectral de Potência (PSD, do inglês *Power Spectral Density*) é mais plana no sussurro [27], apresentando formantes menos pronunciadas [28, 29]. As formantes tendem ainda a ocorrer, para os fonemas correspondentes, em frequências ligeiramente superiores no cenário de sussurro [30, 31, 32]. Este efeito será maior na primeira formante (F1), associada à altura do dorso da língua, mas menos pronunciado na segunda (F2), formante associada ao ponto de articulação (*i.e.*, avanço/recuo da língua), que será bastante idêntico nos dois modos de fala [31]. Outro importante aspeto a destacar é a projeção acústica que é menor no registo sussurrado, em comparação com aquela que é permitida pela fala normalmente vozeada, em larga medida pela ausência do contributo energético das pregas vocais. Como consequência destas características, o sussurro tende a ser inteligível apenas aos que se encontram mais próximos do orador, estando também mais sujeito à interferência de fontes sonoras concorrentes [7]. De facto, e como a investigação tem de resto demonstrado, a inteligibilidade do discurso sussurrado é inferior [33], ocorrendo uma maior probabilidade de confusão entre vogais próximas no espaço das vogais [34]. Estas características tornam também o reconhecimento automático de voz (ASR, do inglês *Automatic Speech Recognition*), uma tarefa substancialmente mais difícil de implementar com sucesso para a fala sussurrada, tanto mais que as técnicas de redução de ruído tendem a ser bem menos eficazes quando aplicadas ao sinal de sussurro em virtude da pior relação sinal-ruído e da distribuição mais plana da energia no espectro [35].

Mesmo tendo em conta estas características e as evidentes desvantagens do sussurro, será porém interessante observar o grau de informação que é preservada ou até inferida, por vezes de forma surpreendente. Notavelmente, diversos investigadores têm apontado a aparente perceção de um tom no sussurro, não obstante a ausência de uma componente periódica no sinal [36, 37]. Mais importante talvez, a evidência vem demonstrando que se preserva no sussurro uma substancial capacidade de identificar características específicas como o sexo do orador [33] ou mesmo o seu estado emocional [38].

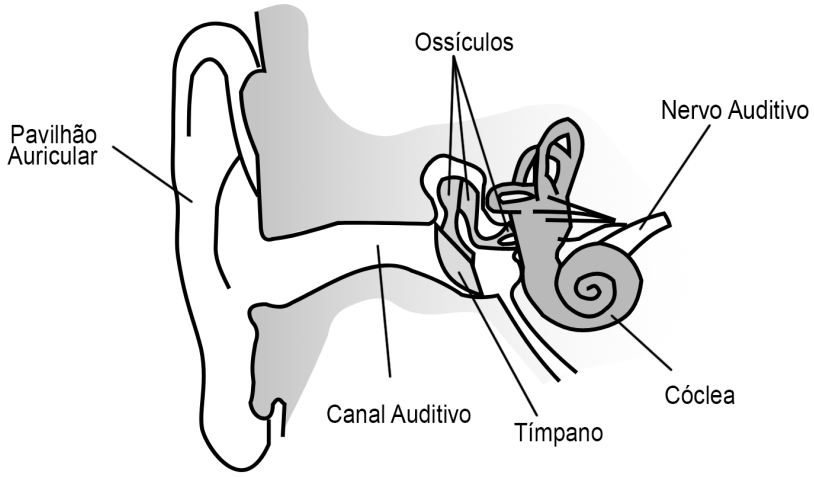


Figura 2.4: Representação do sistema auditivo periférico composto pelo ouvido externo, médio e interno, destacando os órgãos mais importantes na audição.

2.2 O Sistema Auditivo Humano

Importa igualmente descrever o sistema auditivo humano, ainda que de uma forma breve, destacando os aspetos mais relevantes para a dissertação. Este sistema poderá ser dividido essencialmente em duas partes: o sistema auditivo periférico, que é constituído pelo ouvido externo, médio e interno; e pelo sistema auditivo central, composto pelo córtex auditivo e pelos núcleos do tronco cerebral [20, 39]. O sistema periférico tem a função de converter o sinal acústico em sinais neuronais, comunicando-os pelo nervo auditivo ao sistema auditivo central onde serão processados e interpretados permitindo, por exemplo, discriminar diferentes objetos sonoros ou precisar a localização da sua fonte [40].

No que diz respeito ao sistema auditivo periférico, podem-se distinguir três regiões com funcionalidades distintas [20, 39]. O ouvido externo, formado pelo pavilhão auricular e pelo canal auditivo, termina na membrana timpânica, sendo que o pavilhão auricular desempenha o papel de coletor ao encaminhar as ondas sonoras para o canal auditivo, contribuindo ainda para a capacidade em determinar a direção da fonte sonora [41]. No ouvido médio encontram-se os ossículos martelo, bigorna e estribo, responsáveis por uma transferência eficiente do sinal do meio aéreo, oriundo do ouvido externo e que lhes é comunicado pela membrana timpânica, para os fluídos do ouvido interno, através da janela oval. O ouvido interno, que inclui para além dos órgãos relevantes para a audição o vestíbulo (responsável pelo equilíbrio), decompõe a onda mecânica e converte-a nos impulsos elétricos que serão comunicados ao sistema auditivo central [18].

2.2.1 Anatomia e Fisiologia do Ouvido Interno

No ouvido interno, encontra-se um tubo tripartido e espiralado - a cóclea - que contém ao longo da sua extensão uma membrana separando duas das suas secções. Esta membrana, a membrana basilar, aloja por sua vez o órgão de Corti, o órgão sensorial da audição, composto por milhares de células nervosas ciliadas. A membrana basilar apresenta características físicas importantes para o mecanismo da audição. Em particular, é mais rígida e menos espessa na base (*i.e.*, junto da janela oval que comunica com o ouvido médio) e mais flexível e espessa no extremo oposto, o ápex, no que pode ser visto como um sistema massa-mola com diferentes ressonâncias naturais ao longo da extensão da membrana. Desse modo, os estímulos transmitidos pelos ossículos ao ouvido interno serão decompostos em frequência, o que, ao ativar diferentes grupos de células nervosas, resultará num mapeamento espacial das componentes do sinal acústico [4, 20, 39].

Por decompor o sinal acústico nas suas componentes no domínio das frequências, desde os cerca de 20 Hz no ápex até um valor próximo dos 20 kHz na base[3], poderá entender-se que a membrana basilar desempenha um papel semelhante à Análise de Fourier, comportando-se como um banco de filtros paralelos parcialmente sobrepostos. Estes filtros vêm descritos na literatura como 'filtros auditivos' e desempenham um papel importante em diversos fenómenos psicoacústicos [18, 39, 41].

2.2.2 Psicoacústica

A psicoacústica, como disciplina, estuda os sinais acústicos numa perspetiva percetual procurando construir modelos quantitativos que relacionem as características puramente físicas do sinal com a experiência auditiva [4, 18, 42]. Nesta medida, a amplitude do sinal acústico relaciona-se com a perceção da intensidade sonora que, por seguir uma relação aproximadamente logarítmica, no processamento de sinais de som é normalmente representada numa escala em decibel (dB) para as magnitudes. A Frequência Fundamental (F0) é interpretada na audição como o tom e que em virtude também das propriedades do sistema auditivo segue uma relação não linear com a frequência. Quer isto dizer que um conjunto de frequências distribuídas uniformemente numa escala linear em Hz (*i.e.*, ciclos por segundo) não serão interpretadas como estando separadas efetivamente de forma uniforme, com o espaçamento entre os tons a parecer reduzir-se à medida que a frequência sobe. Para refletir este fenómeno e por meio de experimentação, desenvolveram-se outras escalas, de cariz percetual, como as escalas *mel* e Bark. A envolvente espectral, por seu turno, é percebida como o timbre ao passo que a fase, por exemplo, permite construir uma perceção espacial da fonte graças à audição binaural [18]. Encontram-se identificados na literatura diversos outros fenómenos psicoacústicos como os efeitos de mascaramento e da adaptabilidade auditiva, que resultam da evolução dinâmica dos limiares da perceção

[3] Este valor vai-se reduzindo com o avançar da idade à medida que as células próximas da base se vão deteriorando.

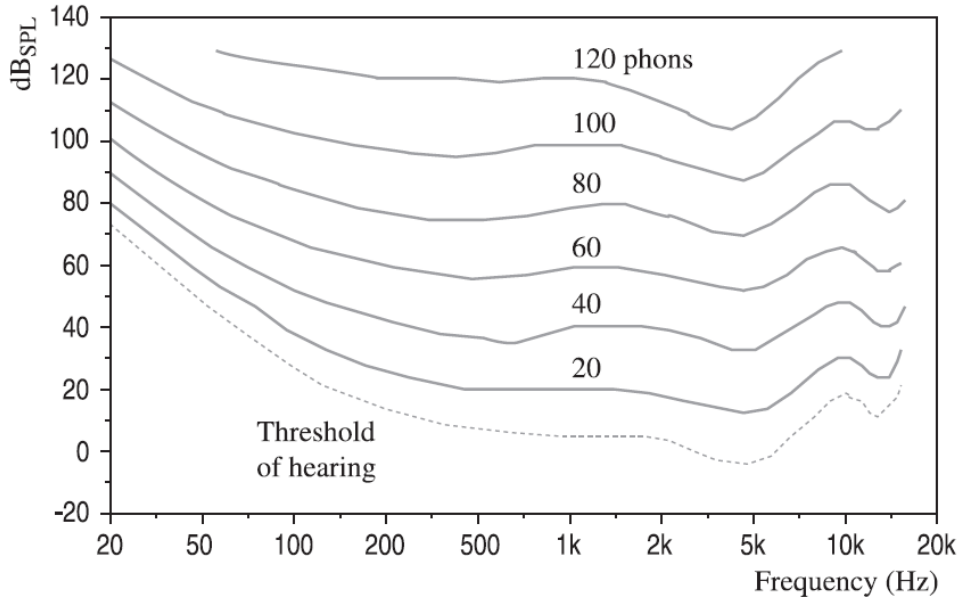


Figura 2.5: Curvas de Fletcher-Munson, ou *equal-loudness contours*, para diferentes níveis de intensidade sonora com indicação do limiar de audição, ou *threshold of hearing* (adaptado de McLoughlin [4]).

auditiva no tempo e em frequência [4, 18, 42]. A evidência empírica sugere ainda que o sistema auditivo humano processa os sinais de fala de uma forma distinta de outros tipos de fonte sonora. Notavelmente, é possível demonstrar experimentalmente que os ouvintes classificam ocorrências distintas do mesmo orador a proferir a mesma palavra sob condições diferentes como sendo similares, não obstante diferenças significativas nas propriedades físicas do sinal, nomeadamente ao nível da magnitude, tempo ou tom. Mais ainda, a fala é detetada e entendida mesmo sob relações de sinal-ruído que tornam impercetíveis outros estímulos auditivos [4].

Não se pretendendo analisar aqui exaustivamente a generalidade dos fenómenos psicoacústicos, interessará, no entanto, sublinhar a não linearidade da perceção auditiva quer em magnitude, quer em frequência e que conduziu ao desenvolvimento de modelos que melhor reflitam a fisiologia do sistema auditivo. Com efeito, não só uma escala logarítmica para as magnitudes é mais adequada para a representação da intensidade sonora, nem sempre a mesma amplitude do sinal é julgada como tendo a mesma intensidade sonora, antes dependendo a aparente intensidade da respetiva frequência. Nem tão pouco o limiar da audição é o mesmo para toda a gama de frequências da audição humana. Em particular, somos menos sensíveis às magnitudes nos limites inferiores e superiores da banda audível, com um pico de sensibilidade que se situa próximo dos 5 kHz. Este fenómeno levou ao desenvolvimento por via experimental das curvas de Fletcher-Munson, também denominadas por *equal-loudness contours* no inglês [4, 18] e que se podem observar na figura 2.5, relacionando a intensidade aparente medida em *phon* com a intensidade sonora necessária

medida em dB_{SPL} (*i.e.*, *sound pressure level* em deciBel).

Para melhor se compreender a não linearidade nas frequências, recupere-se o conceito de filtros auditivos, notando que se diferenciam da Análise de Fourier num aspeto significativo: as bandas destes filtros não se encontram distribuídas no espectro de uma forma uniformemente linear [39]. As propriedades destes filtros têm fortes implicações na seletividade e resolução do sistema auditivo em frequência e na perceção da distância entre tons estando também envolvidas na capacidade de discriminar diferentes sinais e nos já aludidos fenómenos de mascaramento em frequência. Sabendo-se de longa data que a perceção dos tons não progride de forma linear com a frequência, entre 1937 e 1940, por meio de experimentação (solicitando aos participantes para classificarem diferentes estímulos), Stevens e Volkmann introduziram a escala *mel*, uma escala percetual, pretendendo segundo os autores medir a distância psicológica entre os tons [43]. Embora seja ainda hoje bastante utilizada na acústica e no processamento de voz, a escala *mel* tem sido objeto de criticismo ao longo dos anos, com os valores obtidos em diferentes experimentações a diferirem substancialmente dos originalmente propostos, dependendo dos participantes, métodos e procedimentos utilizados [41, 43]. Também por esta razão, não existe 'uma' escala *mel*, tendo vindo a ser propostas diferentes expressões para a conversão de Hz em *mel* [42]. A versão frequentemente adoptada poderá descrever-se porém por:

$$m = 2595 \ \log_{10}\left(1 + \frac{f}{700}\right) \tag{2.1}$$

onde f corresponde ao valor da frequência dado em Hz e m ao correspondente em unidades *mel* na escala perceptual [44].

Entretanto, modelizando a fisiologia da membrana basilar para explicar fenómenos como o mascaramento, Fletcher introduz em 1940 o conceito de 'bandas críticas' (*critical bands*, no inglês) [41, 45], e que se relaciona com duas escalas psicoacústicas: Bark e ERB. Sucintamente, a banda crítica corresponde à gama de frequências a que cada segmento da membrana basilar (*i.e.*, filtro auditivo) responde [43]. Estimando a largura de cada uma destas bandas, Zwicker introduziu em 1961 a escala Bark, uma escala psicoacústica em que cada banda corresponde à largura de banda estimada para o filtro auditivo correspondente e cujos valores se encontram na tabela 2.1. Um Bark corresponderá ainda neste caso à distância entre a frequência central de filtros auditivos consecutivos [6, 39, 43].

Outra abordagem adotada para a modelização das bandas críticas passa pela denominada *equivalent rectangular bandwith* (ERB), *i.e.*, o correspondente filtro passa-banda retangular com a mesma largura de banda da banda crítica centrada numa dada frequência. A aproximação linear para esta largura de banda é dada pela expressão:

$$ERB = 24.7 \ \left(4.37 \ F_c + 1\right) \tag{2.2}$$

onde ERB corresponde à largura de banda (em Hz) centrada na frequência central F_c (em KHz) [39], proposta por Glasberg e Moore [46]. A largura de banda estimada para os

Tabela 2.1: Escala Bark com indicação da frequência central e a largura de cada uma das bandas (adaptado de Zwicker [6]).

Banda	Freq.Central (Hz)	Larg.Banda (Hz)
1	50	100
2	150	100
3	250	100
4	350	100
5	450	110
6	570	120
7	700	140
8	840	150
9	1000	160
10	1170	190
11	1370	210
12	1600	240
13	1850	280
14	2150	320
15	2500	380
16	2900	450
17	3400	550
18	4000	700
19	4800	900
20	5800	1100
21	7000	1300
22	8500	1800
23	10500	2500
24	13500	3500

filtros auditivos pelo método ERB poderá ser utilizada para produzir uma escala psicoacústica alternativa à escala *mel*. Para converter uma frequência em Hz para a escala ERB (ERBS, *i.e.*, *ERB scale* no inglês) poderá recorrer-se à expressão derivada da aproximação anterior:

$$ERBS = 21.4 \ \log_{10}\left(4.37 \ F + 1\right) \tag{2.3}$$

onde $ERBS$ denota o correspondente valor da frequência F (em KHz) na escala psicoacústica de ERBs [39]. Uma diferença relevante entre a escala ERB e Bark é a primeira poder ser descrita simplesmente numa forma analítica (contínua) tornando-a mais prática do ponto de vista computacional, não obstante em qualquer um dos casos ser sempre possível: *a)* converter o valor em Hz na escala linear no correspondente valor na escala psicoacústica; e *b)* fazer representar o espectro num dado número bandas críticas (discretas).

Finalmente, uma outra escala logarítmica descrita na literatura será a escala de 1/3 de oitava, onde cada banda corresponderá a um terço de uma oitava e que, pela sua relação com a escala musical igualmente temperada, se utiliza no meio musical [47, 48].

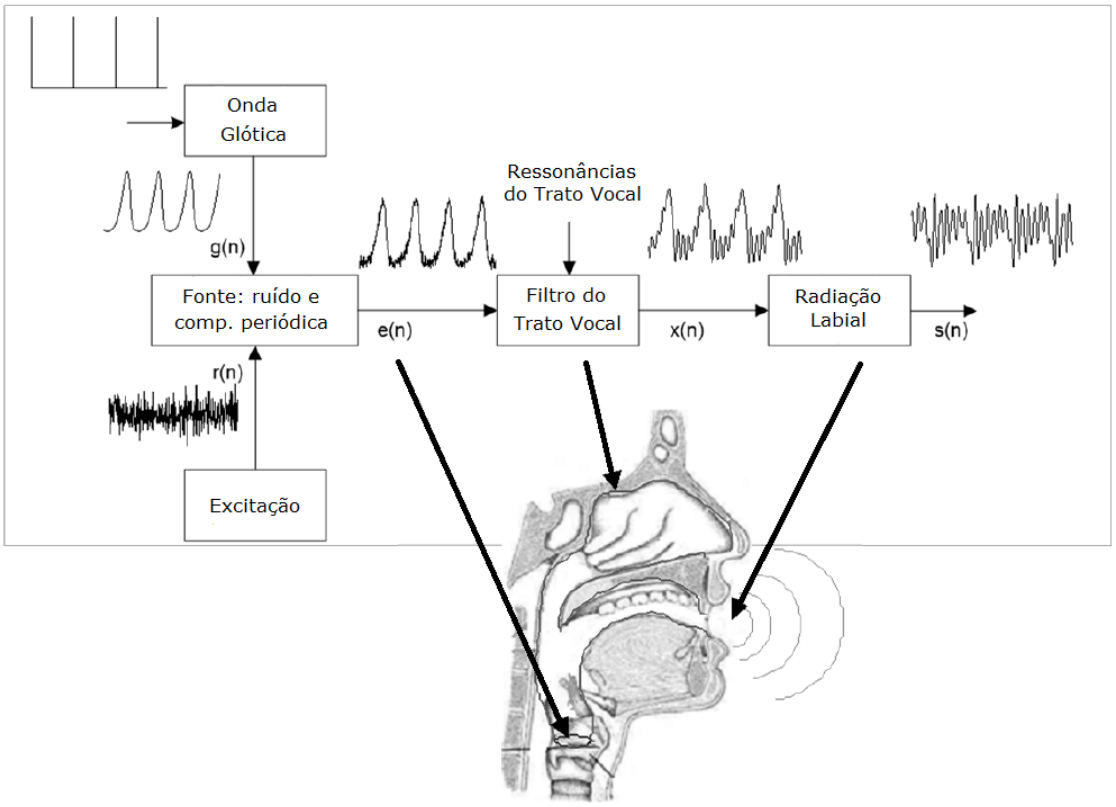


Figura 2.6: Representação do modelo simplificado fonte-filtro através de um diagrama de blocos e a sua relação com aparelho fonador na produção da fala.

2.3 Princípios de Processamento do Sinal de Voz

Uma vez caracterizados os mecanismos fisiológicos da produção da fala e o essencial do sistema auditivo, na perspetiva do trabalho desenvolvido na dissertação, será oportuno então analisar os principais fundamentos da teoria de processamento de sinais de voz, introduzindo agora o paradigma do modelo simplificado fonte-filtro da produção da fala, assim como algumas das mais importantes ferramentas utilizadas em processamento de sinal na estimação e caracterização do filtro do trato vocal.

2.3.1 O Modelo Fonte-Filtro

Com a publicação em 1960 do seu marcante trabalho *"Acoustic Theory of Speech Production"*, Gunnar Fant estabeleceu uma sólida fundação para as décadas seguintes no que diz respeito ao estudo da produção da fala e que é conhecida pelo modelo fonte-filtro. Este modelo descreve o aparelho fonador humano como uma combinação de fontes sonoras e de filtros acústicos, assumindo a independência entre estas duas componentes [49]. A fonte poderá ser de dois tipos: *a)* periódica, correspondendo portanto ao contributo das pregas vocais e sendo responsável pelo vozeamento dos sons; e *b)* não periódica, correspondendo

ao ar turbulento que caracteriza os segmentos de fala não vozeados, como ocorre durante consoantes não vozeadas ou no sussurro de uma forma geral, e que se poderá aproximar por ruído branco. Por seu lado, o filtro representa o contributo de todo o sistema supraglótico que, tal como já descrevemos, atua como um filtro acústico, definindo a envolvente espectral dos sons emitidos na produção da fala [50]. A figura 2.6 ilustra esta relação entre o modelo fonte-filtro e a produção da fala.

Ainda que se trate de uma aproximação, este modelo notavelmente simples constitui uma conveniente representação da produção da fala em geral, permitindo modelizar a fala quer nos cenários de fala sussurrada, quer nos cenários de fala normalmente vozeada [50]. Este paradigma serve também de base quer à denominada modelização linear preditiva, ou *Linear Predictive Coding* (LPC) no inglês [51], quer à Análise Cepstral [2], ambas amplamente utilizadas no processamento de sinais de voz, nomeadamente para a estimação da componente do filtro. Estas técnicas serão objeto de exposição e caraterização nas próximas secções.

2.3.2 Modelização do Filtro via LPC

A modelização LPC é um dos métodos mais utilizados para estimar a envolvente espectral na análise de sinais de voz, sendo também uma técnica aplicada à síntese ou reconstrução dos sinais [22, 51]. Sendo de implementação computacional simples, eficiente e bastante direta, permite também alcançar um interessante compromisso entre qualidade e a taxa de *'bitrate'* na codificação/compressão de sinais de voz quando comparada, por exemplo, com os denominados codificadores de forma de onda [51]. O primeiro passo no processo de codificação e descodificação pela modelização LPC consiste em estimar a envolvente espectral do filtro, correspondente ao efeito do trato vocal e que contem, nomeadamente, a informação respeitante às formantes, permitindo também a remoção dessa componente - a envolvente espectral - do sinal. Este processo vem também denominado na literatura por *'inverse filtering'* (figura 2.7a). O resultado desta operação será uma versão 'branqueada' do sinal, ou resíduo (*'residue'* no inglês), que corresponderá à componente da fonte. A informação relativa ao filtro poderá então ser parametrizada e representada por um conjunto de coeficientes que serão armazenados ou transmitidos a um descodificador remoto para reconstrução. No Capítulo 4 descrever-se-á uma metodologia para a computação da envolvente. Este processo é tipicamente repetido sobre pequenos segmentos de amostras sucessivas do sinal, por norma com sobreposição (*'overlapping'*). Cada um destes segmentos contendo uma pequena porção do sinal denomina-se de *'frame'* [52]. No descodificador, será possível reconstruir o sinal de voz revertendo o processo, ou seja, utilizando o resíduo como excitação (a 'fonte'), aplica-se-lhe o filtro correspondente ao efeito do trato vocal. Tipicamente, este filtro é assumido como um filtro do tipo *'all-pole'*, com uma dada ordem p [52]. A ordem corresponderá, por conseguinte, ao número de pólos do filtro e que

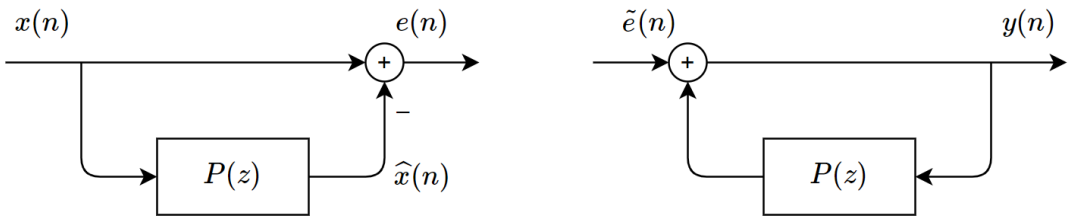


Figura 2.7: Diagrama de blocos de um codificador e de um descodificador LPC (da esquerda para a direita, respetivamente).

é tipicamente de pelo menos dez: dois por cada formante e pelo menos dois adicionais, por forma a compensar a inexistência de zeros, o que permite replicar até quatro formantes. Esta modelização LPC poderá representar-se, no domínio Z, pela seguinte expressão:

$$H(z) = \frac{1}{1 - \sum_{i=1}^{p} a_i z^{-i}} = \frac{1}{P(z)} \tag{2.4}$$

onde p corresponde ao número de polos do filtro *'all-pole'* $P(z)$ representativo do trato vocal [50]. A envolvente espectral assim estimada, ainda que não seja fisiologicamente exata, constitui uma boa aproximação do efeito do filtro do trato vocal [22]. Os diagramas de blocos da figura 2.7 ilustram a aplicação destes princípios, mostrando também como uma imagem do resíduo *e(n)* obtido no codificador em conjunto com a reconstrução do filtro *P(z)* permite ao descodificador reproduzir uma aproximação do sinal original.

2.3.3 Modelização do Filtro via Análise Cepstral

A denominada análise cepstral oferece uma forma alternativa de estimar a envolvente espectral e de separar as componentes da fonte e do filtro, recorrendo neste caso a um artifício curioso que passa por considerar a representação do sinal no domínio das frequências como se de um sinal no tempo se tratasse. Recorrendo a uma escala logarítmica para a representação da magnitude espectral do sinal, aplica-se-lhe a Transformada de Fourier Inversa, o que transporta o sinal para um novo domínio, o domínio *cepstral*, ou seja, como que o espectro do espectro. O cepstrum de potência, de particular interesse para o processamento de sinais de voz, poderá descrever-se pela seguinte expressão:

$$power\ cepstrum = |\mathcal{F}^{-1}\{\log\left(|\mathcal{F}\{x(t)\}|^2\right)\}|^2 \tag{2.5}$$

onde *x(t)* diz respeito ao sinal original no domínio dos tempos, $\mathcal{F}$ à Transformada de Fourier e $\mathcal{F}^{-1}$ à sua inversa [2].

Pretende-se com este artifício capturar a estrutura da envolvente espectral e decompô-la nas suas componentes. O padrão repetitivo formado pela Frequência Fundamental e os

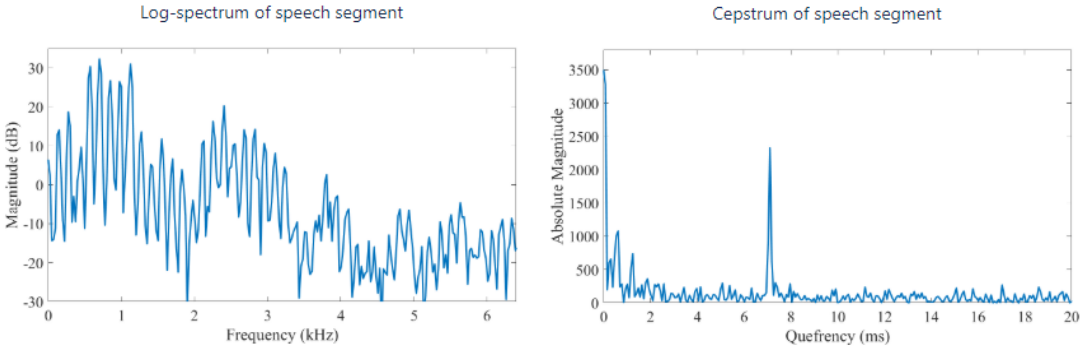


Figura 2.8: Espectro de um segmento de sinal (à esquerda) e respetivo cepstrum (à direita), onde os coeficientes relativos às *quefrencies* correspondem às formantes e o pico isolado nas altas *quefrencies* ao Período Fundamental (adaptado de Bäckström [2]).

respetivos harmónicos (de natureza periódica no domínio espectral) corresponderá a um pico nas *quefrencies* (dadas em unidade de pseudo-tempo) enquanto que o conjunto das formantes, de evolução mais lenta e suave no espectro, será composto por várias *quefrencies* de baixo valor. Dessa forma, como se observa na figura 2.8, será possível determinar o Período Fundamental T0 através do pico isolado nas altas *quefrency*, permitindo assim estimar o valor de F0 (*i.e.*, 1/T0). Por sua vez, o conjunto dos coeficientes *quefrency* de baixo valor dizem respeito à envolvente espectral, que se poderá estimar transportando de novo estes coeficientes para o domínio espectral, obtendo uma envolvente com tanto mais detalhe espectral quantos mais coeficientes se considerarem (*liftering*) [2].

2.3.4 Mel-Frequency Cepstral Coefficients

Pela relativa simplicidade e eficiência computacional, os *Mel Frequency Cepstral Coefficients* (MFCCs) encontram-se entre as características mais prevalentes no processamento de sinais de fala [53]. Recorrendo a uma análise do tipo cepstral por forma a capturar e quantificar as características da envolvente espectral do sinal, incorporam informação relativa à perceção auditiva humana, nomeadamente através da representação das magnitudes numa escala logarítmica e da representação segundo uma escala percetual do tipo *mel* para as frequências [53]. A computação destes coeficientes poderá sintetizar-se nos seguintes passos: começar por segmentar o sinal em pequenas *frames*; estimar o espectro de potência de cada uma delas; aplicar ao espectro de potência um banco de filtros triangulares uniformemente distribuídos na escala *mel*; apurar a energia contida em cada filtro; obter o logaritmo de cada um desses valores; aplicar a *Discrete Cosine Transform* (DCT) ao conjunto de coeficientes obtido, transportando o sinal para o domínio cepstral e, finalmente, truncar os coeficientes (tipicamente) do 2º ao 13º (*liftering*), eliminando desta forma informação redundante ou pouco relevante como a componente DC ou o detalhe fino do espectro do sinal [54].

O recurso à DCT explica-se pela eficiência computacional, pelo facto de operar apenas no domínio real e pela capacidade em descorrelacionar os coeficientes obtidos com o banco de filtros triangulares, sujeitos a uma elevada covariância [53]. Deste modo, os MFCCs caracterizam a envolvente espectral que é tão importante para a identificação de fonemas, ainda que de uma forma compacta e abstrata, com especial apreço no caso das vogais. Adicionalmente e como forma de capturar as características não estacionárias do sinal, são consideradas as primeiras e segundas diferenças dos próprios MFCCs, denominadas de *delta* e *delta-delta*, as quais caracterizarem a taxa de variação e de aceleração dos coeficientes, o que permite aumentar o número de características extraídas com reduzido esforço computacional adicional [2].

Não obstante o assinalável sucesso dos MFCCs no processamento e análise de sinais de fala, podem apontar-se-lhes ainda assim algumas das suas deficiências como a limitada robustez na presença de ruído branco aditivo, a escolha pela escala percetual *mel* que poderá não ser ótima ou a sua inadequação para a síntese, dada a dificuldade em obter a partir dos MFCCs o espectro de potência de forma precisa e congruente [2].

2.4 Síntese do Capítulo

Neste capítulo descreveram-se os mecanismos de produção da fala, destacando o papel de filtro acústico desempenhado pelo trato vocal, crucial na produção dos fonemas (vogais e consoantes) seja na fala normal, seja na fala sussurrada. Discutiram-se as principais diferenças do ponto de vista percetual e do ponto de vista do processamento de sinal entre os dois modos de fala, sublinhando-se o maior desafio que o sussurro introduz.

Analisaram-se também, de forma breve, a anatomia e fisiologia do sistema auditivo, com maior ênfase nos fenómenos psicoacústicos que levaram à derivação de diversas escalas alternativas para a análise e representação dos sinais de voz no domínio das frequências e que se revelarão importantes no capítulo 5, dedicado à identificação de vogais sussurradas.

Resumiram-se ainda os aspetos fundamentais do processamento de sinais de voz, descrevendo o modelo simplificado fonte-filtro da produção da fala e duas das principais técnicas utilizadas no processamento de sinal para estimar o filtro do trato vocal, baseadas na codificação linear preditiva (LPC) e na análise cepstral.

No próximo capítulo apresenta-se uma revisão bibliográfica focada na investigação e nos resultados obtidos em anos recentes, conforme disponíveis na literatura, na melhoria e conversão de sinais de fala sussurrada.

Capítulo 3

Melhoria e Conversão de Sussurro - Uma Revisão Bibliográfica

Com o objetivo de avaliar o que de melhor se tem explorado e conseguido na melhoria e conversão de sinais de fala sussurrada, realiza-se neste capítulo uma breve revisão bibliográfica às técnicas previamente utilizados e aos resultados e conclusões disponíveis na literatura.

3.1 Abordagem MELP

Aponta-se, em primeiro lugar, o trabalho pioneiro levado a cabo por Morris e Clements [55], baseado numa abordagem Mixed-Excited Linear Prediction (MELP) [52]. Este codificador permite manipular separadamente o sinal de excitação da fonte e o modelo de filtro, dado por um LPC de 10ª ordem (ver 2.3.2).

Neste caso, a injeção de uma onda periódica no sinal de excitação da fonte permite adicionar vozeamento enquanto a implementação de um filtro de fase mínima, para efeitos de compensação das diferenças espectrais entre o discurso vozeado e não vozeado, modifica o filtro LPC supramencionado. Adicionalmente, é efetuada uma correção às frequências das formantes e bandas das envolventes espectrais transpondo as formantes para frequências mais baixas e estreitando as suas larguras de banda, por forma a refletir as características distintas típicas das correspondentes versões vozeadas. Ainda, segundo os autores, para permitir o controlo do *'pitch'* em tempo-real por parte do utilizador, o período deste é controlado via intensidade sonora, algo que a documentação é lacónica em detalhar. Destaca-se neste processo, porém, a ausência de segmentação, o que significa que a banda de 0-3 kHz é permanente e forçosamente vozeada enquanto que a banda de 3-4 kHz é tratada como não vozeada. Aníbal Ferreira em [7] observa também, em análise, que o algoritmo poderá necessitar de treino prévio recorrendo a discurso vozeado antes de se conseguir a conversão de sussurro para vozeado.

3.2 Abordagem CELP

Encontra-se outro exemplo de conversão de sussurro para vozeado no trabalho desenvolvido por McLoughlin *et al* [56, 57, 58], desta vez assente num codificador do tipo Code Excited Linear Prediction (CELP) [52].

Neste caso, o codificador implementa a modificação das formantes e a modulação glótica artificial sem requerer informação *a priori* que dependa do orador. Em concreto, e dada a natureza ruidosa do sussurro, o codificador estima primeiramente a frequência e amplitude de até quatro formantes, de forma robusta, evitando formantes não genuínas. Estas formantes assim estimadas são modificadas tendo em consideração as diferenças características entre as correspondentes versões vozeadas e não vozeadas, de forma análoga ao indicado na abordagem anterior. Estas formantes, já ajustadas, são então moduladas via um modelo de cosseno elevado cuja Frequência Fundamental está dependente da frequência da primeira formante, mais precisamente em cerca de 10 % desse valor, sendo aparente que o vozeamento é essencialmente determinado pelo processo de estimação das formantes. Foi também tomada a decisão de não incluir neste processo uma distinção entre *'frames'* vozeadas e *'frames'* não vozeadas, o que os autores sustentam com a observação de que decisões rígidas derivadas a partir do sussurro não funcionarão satisfatoriamente.

Ainda que os resultados dos testes realizados sugiram um impacto positivo em termos percetuais, valores de 3.0 a 3.6 [56, 58] na escala MOS (*Mean Opinion Score*) em diferentes iterações apontam no sentido de uma qualidade aquém do desejável.

3.3 Abordagem NAM

Toda *et al* propõem em [59] uma outra abordagem ainda onde se recorre a um microfone do tipo NAM (*non-audible murmur*) que é colocado sobre a pele, atrás da orelha. Distinguindo-se de um microfone convencional, que captura sussurros audíveis, este dispositivo captura fala não vozeada que é conduzida pelo corpo humano. Este processo favorece desde logo a robustez contra sinais sonoros concorrentes como ruídos externos. Por outro lado, não beneficia dos efeitos de radiação dos lábios e narinas, perdendo-se ainda bastante informação contida nas altas frequências devido a um forte efeito de filtro passa-baixo. Esta metodologia resulta igualmente em formantes menos definidas e pronunciadas com prejuízo para a audibilidade e inteligibilidade. Ainda assim, os autores mostram que uma conversão de carácter estatístico é capaz de converter sinais NAM em sussurro audível ou em funcionalidades de fala vozeada. Os autores reconhecem também que a conversão obtida na sua metodologia sofre de características não naturais de prosódia como resultado da dificuldade em estimar a frequência fundamental da fala normal. Mais importante, uma análise mais detalhada do método indicia não ser adequado para operação em tempo-real dado que requer alinhamentos precisos entre sinais de entrada e sinais de referência.

3.4 Investigação na FEUP

A conversão de sussurro tem sido, em anos recentes, também já objeto de investigação na Faculdade de Engenharia da Universidade do Porto. Em 2015, Patrícia Oliveira desenvolveu na sua dissertação de mestrado [60] a ideia base de um algoritmo para vozeamento artificial de fala sussurrada adotando estruturas e aspetos funcionais de um codificador percetual de alta qualidade [61] como: o processamento de amostras de sinais de fala a 22050 Hz acomodando banda suficiente para sinal de fala de alta qualidade; processos de análise e síntese baseados em transformadas; e permitindo a implantação de um sinal periódico controlado parametricamente no domínio das frequências em termos de magnitude e fase. A ideia central é a de que, uma vez que as regiões não vozeadas da fala normal e da fala sussurrada são essencialmente as mesmas, estas não devem ser modificadas pelo algoritmo, preservando a inteligibilidade ou mesmo informação de carácter idiossincrático. Deste modo, o algoritmo proposto deverá primeiramente localizar as regiões do discurso sussurrado a ser vozeadas e, feito isto, implantar nestas regiões um substituto para a componente periódica do sinal, em falta. Apesar da simplicidade desta abordagem, os testes subjetivos revelaram que, em termos de inteligibilidade, os sinais de voz convertidos apresentam resultados ligeiramente melhores que os sussurrados originais.

Aníbal Ferreira em [7], tirando partido das pistas e conclusões obtidas naquele trabalho, e propondo melhorias, observa que:

- sendo a segmentação temporal corretamente implementada, a ideia de implantar uma componente sintética vozeada é simples e bastante efetiva;
- a segmentação fonética conduzida estatisticamente tende a originar erros de classificação com impactos negativos do ponto de vista percetual;
- para obter resultados que soem naturais, a segmentação deverá ser implementada utilizando uma resolução de pelo menos 10 ms;
- a envolvente espectral da versão vozeada estimada apresenta-se bastante plana herdando esta característica da versão sussurrada e em contraste com as versões reais e naturais da fala vozeada que contêm picos pronunciados e uma assimetria significativa entre a energia às baixas e às altas frequências;
- na conversão espectral de sussurrado para vozeado, a coarticulação pode dar origem a uma envolvente espectral que difere significativamente da mesma vogal quando pronunciada de forma sustentada, nomeadamente no número e magnitude das formantes;
- a variação da frequência fundamental é um dos fatores mais importantes no que diz respeito a fornecer um sentido de aparente naturalidade à versão sintetizada;

- micro-variações da frequência fundamental afetam a perceção e podem conter informação idiossincrática relevante.

As melhorias propostas dizem respeito à componente periódica a implantar, à segmentação e à estimação da envolvente espectral. Uma vez que as duas primeiras dizem respeito a outros tantos módulos do projeto DyNaVoiceR e extravasam o foco específico desta dissertação, destaca-se as que dizem respeito à última.

Para que a correção do ponto de vista linguístico seja atingida na conversão, o autor propõe que o primeiro passo deva consistir em encontrar um modelo da envolvente espectral de cada *'frame'* do sinal sussurrado utilizando uma modelização LPC. Para estimar a envolvente espectral da versão vozeada sintética testaram-se dois métodos: um primeiro baseado numa manipulação e projeção das raízes do polinómio LPC no plano Z com o objetivo de estreitar a largura de banda das formantes, tornando-as mais pronunciadas; um segundo método consistindo numa compensação simples da envolvente sussurrada semelhante à abordada em 3.1. Em concreto, a compensação é obtida através da diferença média entre os espectros da versão sussurrada e da versão vozeada para as diversas vogais. Em ambos os casos se concluiu que resultam em vogais sintéticas que soam nasaladas e não naturais. O autor indica que os espectros estimados preservam ainda muito da natureza plana do espectro da versão sussurrada original, sugerindo que uma conversão convincente requer o recurso a envolventes espectrais protótipo que sejam realistas, concluindo com o que deverá ser o foco de próximos desenvolvimentos:

- reforçar a naturalidade dos sinais sintetizados, nomeadamente enriquecendo a base de dados de envolventes protótipo promovendo a eficácia da componente de análise e a naturalidade na componente de síntese;
- avaliar formalmente a qualidade subjetiva do discurso sintetizado;
- implementar o algoritmo numa plataforma que opere em tempo-real;
- adaptar o algoritmo a condições acústicas diversas;

O projeto DyNaVoiceR, inspirando-se e beneficiando da experiência e conhecimento adquiridos nestes trabalhos, poderá em boa medida ser visto como uma sequência natural deles.

3.5 Síntese do Capítulo

A revisão bibliográfica conduzida neste capítulo permitiu concluir que o objetivo que se pretende atingir no projeto DyNaVoiceR não foi alcançado ainda de forma satisfatória pela investigação e pelos métodos já explorados. Subsistem importantes desafios por resolver, nomeadamente a implementação de um sistema viável, capaz de converter em tempo-real um sinal original de fala sussurrada num sinal de fala vozeado e que ofereça elevados índices de inteligibilidade e de naturalidade.

Capítulo 4

Análise e Modelização da Envolvente Espectral

Tendo em consideração a complexidade do problema a solucionar, conforme já foi sendo descrito ao longo dos capítulos anteriores, estabeleceram-se três etapas para o trabalho a desenvolver nesta dissertação. Este capítulo documenta a primeira destas etapas, que consiste na análise e modelização das envolventes espectrais das vogais orais do Português Europeu, utilizando a base de dados de oradores disponível no projeto, incluindo ainda a implementação de um sistema de síntese de vogais sussurradas com recurso a diferentes tipos de envolvente, por forma a avaliar perceptualmente as diversas envolventes espectrais.

4.1 Base de Dados de Oradores

A base de dados de oradores do DyNaVoiceR é uma ferramenta do projeto contendo um conjunto de exercícios vocais realizados por 30 informantes (oradores), 15 femininos e 15 masculinos, gravados em ficheiros áudio no formato *.wav* amostrados a 44100 Hz, em que cada exercício está disponível na versão normal e na correspondente sussurrada. A base de dados inclui a anotação fonética manual destas gravações, constituída por um conjunto de ficheiros formatados que permitem localizar e identificar cada fonema. Esta base de dados foi criada para suprir as necessidades específicas do projeto tendo sido implementada em colaboração com a equipa da Universidade de Aveiro que integra o DyNaVoiceR. Para melhor adaptar a base de dados às necessidades do projeto, todos os ficheiros áudio foram previamente convertidos para uma frequência de amostragem de 22050 Hz. Cada um destes ficheiros de áudio contém três repetições do mesmo exercício pelo mesmo orador. De entre os exercícios disponíveis na base de dados, foram utilizados na dissertação os seguintes:

- Nove vogais orais utilizadas no Português Europeu, na forma sustentada (A.1);
- Vinte e oito dissílabos, contendo várias daquelas vogais em contexto de palavra (A.2);
- Seis pequenas frases, contendo também alguns exemplos de vogais nasais (A.3).

Tabela 4.1: As nove vogais orais do Português Europeu padrão disponíveis na base de dados DyNaVoiceR, no modo sustentado.

Código	Vogal	Exemplo
01	/i/	i̲lha
02	/ê/	pe̲so
03	/é/	e̲la
04	/á/	á̲gua
05	/â/	a̲marelo
06	/ó/	ó̲culos
07	/ô/	o̲vo
08	/u/	u̲va
09	/e/	sede̲

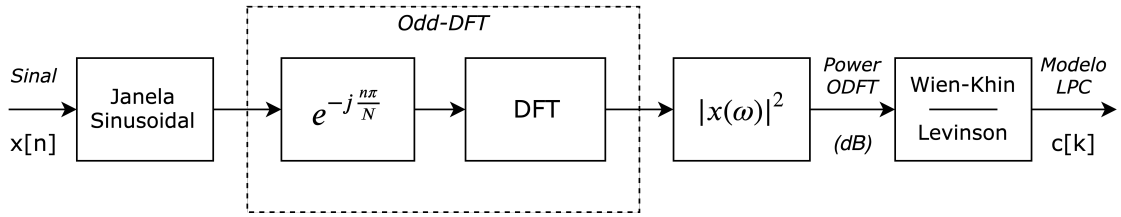
As 9 vogais orais que constituem o foco de estudo da dissertação encontram-se discriminadas na tabela 4.1. Importa notar que a frequência de ocorrência na base de dados de cada uma das vogais varia significativamente. A isto soma-se também o facto de o número de ocorrências variar de orador para orador, existindo cenários em que uma dada vogal não está disponível na base de dados para determinado orador. As vogais mais frequentes na base de dados incluem o /i/ (de 'i̲lha'), o /á/ (de 'á̲gua'), o /â/ (de 'a̲marelo') e o /u/ (de 'u̲va'). Devido a estas e outras inconsistências (como, a título de exemplo, a não correspondência entre o fonema na forma normal e na sussurrada para determinados oradores) optou-se por realizar uma pré-seleção de 20 oradores, 10 femininos e 10 masculinos, a serem utilizados nos trabalhos que aqui se descrevem. Com o objetivo de agilizar as etapas previstas, consolidou-se a base de dados com um terceiro conjunto de ficheiros contendo representações espectrais pré-processadas de cada segmento de sinal contendo vogais e que foram obtidas conforme os procedimentos que se descrevem de seguida.

4.2 Modelização LPC de Vogal Sussurrada

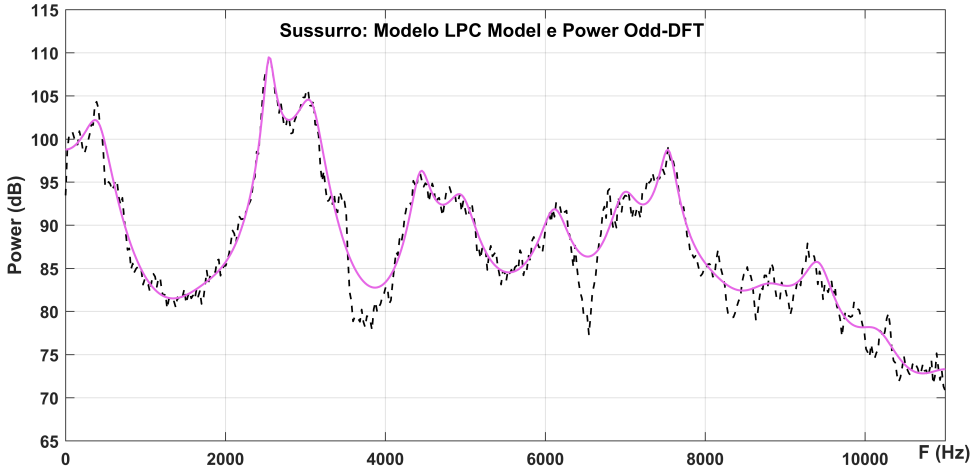
A extração e modelização da envolvente espectral das vogais orais sussurradas presentes na base de dados incorporou parte do trabalho previamente desenvolvido no projeto. Esta operação segue um percurso geral típico no processamento de sinais de voz e que se encontra sintetizado no diagrama de blocos da figura 4.1a. Na implementação do DyNaVoiceR, o sinal de voz é primeiramente sujeito a uma operação de janelamento em segmentos de 1024 amostras de comprimento e 50 % de sobreposição, ou seja, com passos de 512 amostras, sendo cada um destes segmentos (*frames*) multiplicado por uma janela sinusoidal, correspondente à raiz quadrada da janela de Hanning deslocada, conforme a expressão:

$$w_n = sin\left(\frac{\pi\ (n+0.5)}{N}\right) \tag{4.1}$$

onde n ($0<=n<=N$-1) representa o índice da amostra e N o comprimento da *frame*.



(a) Diagrama de blocos da extração da envolvente espectral segundo o modelo LPC.



(b) Exemplo da Power ODFT média de um segmento de sinal (linha a tracejado) e o correspondente modelo LPC de 22ª ordem (linha contínua).

Figura 4.1: Diagrama de blocos da extração de envolventes (a) acompanhado de um exemplo da envolventes espectral estimada para um segmento de fala sussurrada (b).

A opção por esta janela sinusoidal justifica-se pelas suas propriedades e conveniência analítica, cumprindo requisitos para uma reconstrução perfeita que outras janelas alternativas tipicamente utilizadas na codificação de sinais de áudio, como as janelas de Hamming ou de Hanning, não cumprem [62, 63].

A passagem para o domínio das frequências, realizada através da *Odd Discrete Fourier Transform* (ODFT) [64], é assegurada pela multiplicação de cada *frame* por uma exponencial conforme:

$$exp_n = e^{-j\frac{n\pi}{N}} \tag{4.2}$$

em que n corresponde ao índice da amostra na *frame* de comprimento N, seguida da aplicação da *Discrete Fourier Transform* (DFT). A ODFT traduz-se numa variante da DFT em que a frequência de cada amostra no domínio das frequências (*i.e.*, cada *bin*) se encontra deslocada para a direita de $\frac{\pi}{N}$ (frequência normalizada). Esta alternativa oferece um conjunto de vantagens como a estimação mais precisa nos limites superior e inferior do espectro, graças à nova localização dos *bin* nos extremos do espectro [65], e a faculdade de representar sinais reais no domínio das frequências com apenas $\frac{N}{2}$ coeficientes únicos

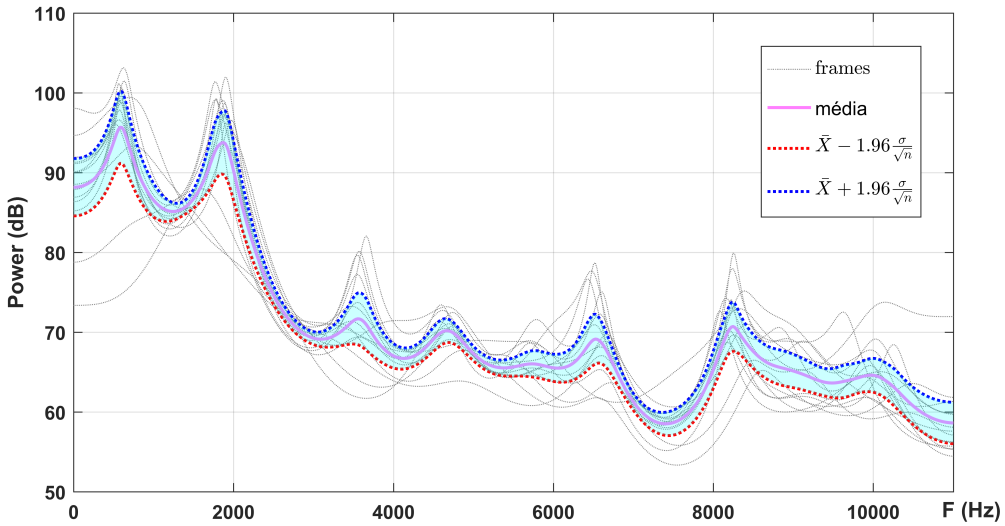


Figura 4.2: Média da envolvente espectral e respetivo intervalo de confiança de 95% (envolventes processadas *frame* a *frame* representadas a tracejado).

em lugar de $1 + \frac{N}{2}$, com os benefícios computacionais que daqui advêm [63, 64]. Será desta transformada do sinal que se retirará a *Power Spectral Density* (PSD) da magnitude, em dB, doravante denominada de Power ODFT do sinal (ou de cada *frame*). Desta são computados os respetivos coeficientes de autocorrelação, tirando partido do Teorema de Wiener-Khintchine. Finalmente, os coeficientes do modelo LPC *all-pole* da ordem pretendida são obtidos com recurso ao algoritmo de Levinson-Durbin [66, 67].

Nesta etapa foram obtidas diversas variantes de envolventes LPC de vogais sussurradas, sempre tomando por referência para a localização dos fonemas a anotação contida na base de dados:

- a envolvente LPC de 22ª ordem obtida da média simples da Power ODFT do segmento de sinal correspondente a cada vogal, com exclusão da primeira e última *frame*, resultando na variante mais plana das consideradas;
- a envolvente LPC de 22ª ordem obtida da média simples da Power ODFT dos 60% das *frames* mais centrais desses segmentos de sinal, beneficiando de uma maior estabilidade das características espectrais nesta porção do sinal e resultando em formantes mais definidas;
- variações das versões anteriores, como considerar o topo do intervalo de confiança de 95 % da envolvente média, por forma a obter formantes mais definidas e pronunciadas como no exemplo da figura 4.2 ou utilizar um LPC de ordem superior.

Para complementar este estudo e permitir a análise comparativa das diferentes envolventes espectrais, nomeadamente através da síntese de sussurro, foram extraídas também as envolventes LPC das vogais correspondentes normalmente vozeadas dos mesmos oradores.

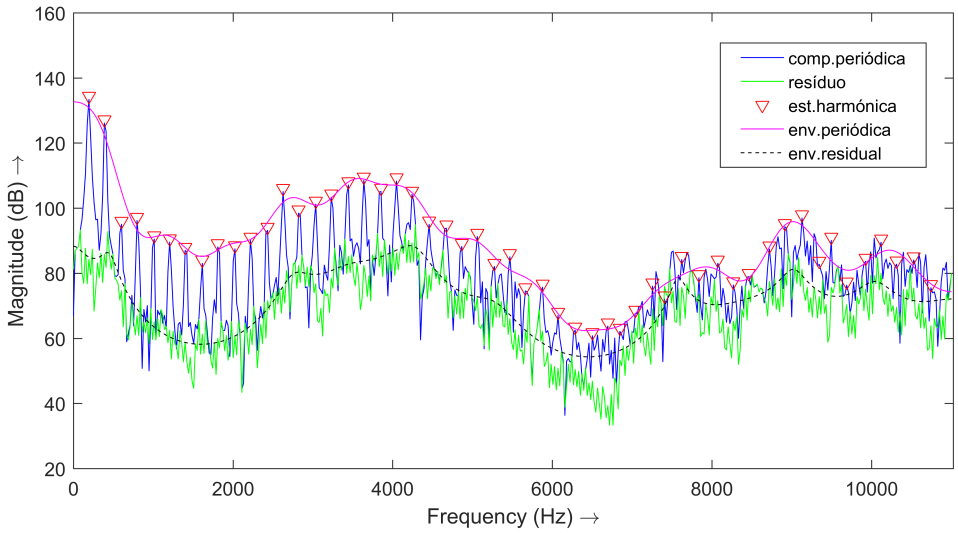


Figura 4.3: Representação de uma *frame* de vogal vozeada no domínio das frequências, revelando a respetiva estrutura harmónica, o resíduo e as correspondentes estimativas de envolvente espectral.

4.3 Modelização LPC de Vogal Vozeada

A extração das envolventes de vogais vozeadas beneficiou de igual modo da plataforma e procedimentos já implementados no projeto DyNaVoiceR, seguindo um percurso e metodologia idênticos ao descrito na extração das envolventes de vogal sussurrada. No cenário de vogal vozeada, consideraram-se, no entanto, dois tipos distintos de envolvente espectral: *a)* a envolvente obtida a partir da componente periódica do sinal e *b)* uma segunda envolvente correspondente à componente de ruído do sinal, ou resíduo.

4.3.1 Componente Periódica

Partindo neste caso também da Power ODFT do sinal, obtida segundo o mesmo procedimento já descrito na secção anterior mas agora aplicado a segmentos de sinal de fala normalmente vozeada, a plataforma do DyNaVoiceR estima de forma independente a Frequência Fundamental (F0) e todos os picos relevantes do espectro em magnitude, recorrendo ao algoritmo *SearchTonal* que, assentando num banco de heurísticas, permite derivar de forma precisa toda a estrutura harmónica do sinal [68]. A envolvente LPC da componente periódica é então estimada a partir da interpolação linear com base nos picos espectrais desta estrutura harmónica (em dB), mais uma vez tirando partido do Teorema de Wiener-Khintchine e aplicando de seguida a recursão de Levinson-Durbin. Esta envolvente pode ser observada no exemplo da figura 4.3, na linha contínua que acompanha os picos relativos aos harmónicos, demarcados na mesma figura pelos triângulos a vermelho.

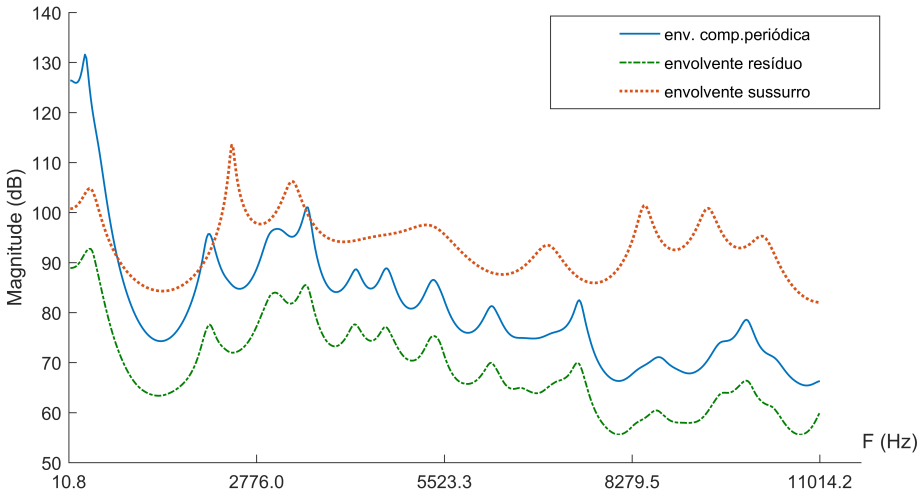


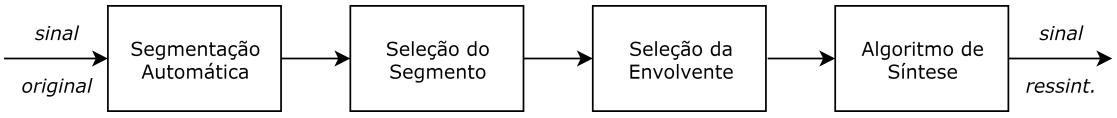
Figura 4.4: Envolvente espectral de uma vogal sussurrada e da mesma vogal vozeada para o mesmo orador, incluindo a componente periódica e resíduo.

4.3.2 Componente de Ruído

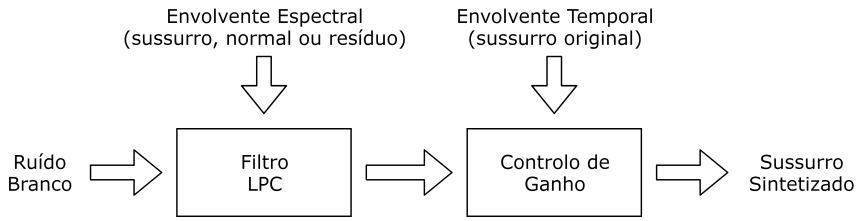
Para estimar a componente de ruído do sinal, o resíduo, a plataforma do DyNaVoiceR sintetiza a estrutura harmónica estimada pelo *SearchTonal* e subtrai esta estrutura harmónica sintética ao espectro original (complexo), *i.e.*, a ODFT do sinal, gerando com esta operação uma estimativa da componente ruidosa. O resíduo assim obtido é suavizado preenchendo com ruído adequado os vales pronunciados no espectro resultante, que surgem como artefacto da subtração da estrutura harmónica. Será desse resíduo suavizado, o qual se poderá observar na figura 4.3 representado a verde, que será derivada a envolvente residual, utilizando procedimentos idênticos aos descritos para as versões anteriores. Esta envolvente residual encontra-se na mesma figura representada pela linha a tracejado acompanhando o resíduo.

Foram extraídas pelos processos descritos as envolventes espectrais para cada uma das 9 vogais orais sustentadas, no modo normal e no modo sussurrado, para o conjunto pré-selecionado de 20 oradores (10 femininos e 10 masculinos) assim como das vogais que se encontram nos dissílabos disponíveis na base de dados, na forma isolada.

Numa primeira análise, por inspeção visual, foi desde logo possível constatar que a envolvente residual apresenta sensivelmente as mesmas frequências formantes que a envolvente normal (componente periódica) correspondente, mas caracterizando-se por uma PSD mais plana, no que se assemelha ao sinal de sussurro. Verificou-se também, de uma forma geral, a tendência para a subida, da versão vozeada para a versão sussurrada, no valor das primeiras formantes, em especial na primeira formante (F1), conforme se encontra descrito na literatura [32, 69] e será possível observar no exemplo da figura 4.4.



(a) Diagrama de blocos do sistema de síntese de sussurro.



(b) Esquema seguido pelo algoritmo de síntese de sussurro.

Figura 4.5: Diagrama de blocos do sistema de ressíntese implementado (a) e o esquema seguido pelo algoritmo de síntese (b), utilizando ruído branco como a componente de fonte no sussurro numa aplicação do modelo fonte-filtro.

4.4 Síntese de Sussurro

Com o objetivo de avaliar, do ponto de vista percetual, o efeito de cada um daqueles tipos de envolvente, implementou-se um sistema de síntese de sussurro baseado num esquema relativamente simples, como o que ilustra o diagrama de blocos da figura 4.5a. Este sistema permite ao utilizador selecionar qualquer tarefa previamente gravada e disponível na base de dados (incluindo vogais sustentadas e palavras isoladas), selecionar um segmento desse sinal para ressíntese, uma envolvente espectral pré-processada e, com base nela, reconstruir o sinal selecionado com sussurro sintetizado. Esta ressíntese de sussurro permitirá avaliar perceptualmente a qualidade e representatividade dos modelos médios obtidos de vogais sussurradas. Uma vez que o utilizador poderá selecionar também uma envolvente espectral obtida de vogal vozeada e utilizá-la para sintetizar sussurro, poderá em boa medida entender-se esse processo como uma inversão do objetivo do projeto (*i.e.*, a conversão de fala sussurrada em fala vozeada) com o intuito de obter pistas e revelar possíveis oportunidades na conversão de fala sussurrada em fala normal vozeada, nos processos de síntese. Procurava-se, com particular interesse, perceber a relação entre a envolvente do sussurro e a envolvente do resíduo para os fonemas correspondentes.

O sistema de síntese de sussurro segue um processo de concatenação em que se mantêm os segmentos originais intactos com exceção daquele que o utilizador definir para substituição (*e.g.*, o segmento correspondente ao /i/ em 'pica'). Isto tanto permite ao sistema a reprodução de uma vogal sustentada como a implantação de sussurro sintético num dos fonemas de uma determinada palavra. A localização de cada um dos segmentos no sinal é determinada de forma automatizada com recurso à anotação disponível na base de dados.

A ressíntese do segmento selecionado é processada *frame* a *frame*, fazendo aplicar o filtro LPC definido pelo utilizador sobre a fonte, a qual é aproximada por ruído branco. Sendo esta síntese de sussurro gerada no domínio dos tempos, para assegurar uma reconstrução consistente e uma implantação graciosa do segmento sintético, é gerado previamente um segmento de ruído branco com a duração adequada e introduzido um controlo de ganho em cada *frame*, durante a reconstrução, por forma a que o sinal ressintetizado replique o envelope temporal do sinal original. Esta precaução, além de mitigar artefactos no implante, contribui também de forma substancial para uma perceção de maior naturalidade no resultado final, que é especialmente evidente na ressíntese de vogais sustentadas.

4.5 Análise Subjetiva dos Resultados

Utilizando o sistema de síntese de sussurro implementado, conduziram-se diversos testes com caráter informal, com a participação de um conjunto reduzido de participantes, nomeadamente elementos do projeto. Estes testes abrangeram a ressíntese de sussurro utilizando diferentes variantes de envolventes médias de vogais sussurradas (incluindo também os modelos médios obtidos de vogais em contexto de palavra, por forma a avaliar a representatividade desses modelos), a síntese de sussurro com recurso a envolventes obtidas das correspondentes vogais vozeadas e a implantação de vogais sussurradas sintéticas em palavras sussurradas, avaliando a naturalidade, inteligibilidade e correção linguística dos sinais obtidos.

Da ressíntese de vogais sussurradas sustentadas com recurso a envolventes obtidas de sussurro, destacam-se as seguintes conclusões:

- A envolvente LPC de 22ª ordem obtida a partir da média simples da Power ODFT (excluindo a primeira e última *frame*) de vogais extraídas de palavras produziu resultados de qualidade sofrível; verificou-se com certa frequência ser difícil reconhecer a vogal, especialmente quando ouvida isoladamente e desprovida de contexto;
- A envolvente LPC de 22ª ordem obtida a partir dos 60% das *frames* centrais dessas mesmas vogais conduziu a resultados comparativamente melhores, verificando-se uma melhoria sensível em termos de inteligibilidade, o que se sugere dever-se ao facto destas *frames* corresponderem de forma mais precisa à vogal, estando menos sujeitos a efeitos de coarticulação com os fonemas adjacentes;
- A utilização de envolventes LPC de ordem superior (*e.g.* 33ª ou 44ª ordem) melhoram de forma sensível os resultados em termos de naturalidade e inteligibilidade, graças à capacidade superior do LPC em seguir de forma precisa a curva média da PSD; a substituição da média simples pelo topo do intervalo de confiança de 95%, que resulta em formantes mais pronunciadas, conduziu a conclusões idênticas; estas diferenças são menos evidentes porém quando comparado com a envolvente obtida a partir dos 60% das *frames* mais centrais.

A síntese de sussurro com base na envolvente residual obtida da vogal vozeada correspondente produziu resultados equiparáveis aos da síntese com base em envolventes da vogal sussurrada. A localização das formantes nas envolventes obtidas do resíduo do sinal (de vogal vozeada) tende a coincidir com a da correspondente componente periódica, conforme já tinha sido observado, o que é de resto expectável uma vez que qualquer uma destas componentes do sinal será sujeita aos mesmos efeitos de modulação nas frequências, decorrentes da articulação do trato vocal. Isto é, ambas as componentes encontram-se sujeitas ao mesmo filtro de trato vocal. A síntese de sussurro com base na envolvente residual conduziu à conclusão de que se preserva no resíduo informação linguística, não obstante a natureza mais plana da sua PSD. A síntese de sussurro com recurso à envolvente obtida da componente periódica, ainda que soasse diferente, especialmente devido à maior concentração da energia nas baixas frequências, não indicou apresentar índices de inteligibilidade ou naturalidade superiores aos obtidos com a envolvente residual.

Finalmente, o implantação de sussurro na vogal situada na primeira sílaba de um pequeno conjunto de palavras de teste, que incluíram as vogais /a,e,i,o,u/, quer utilizando a envolvente média obtida da vogal sussurrada sustentada, quer utilizando a envolvente obtida dos modelos médios (melhorados) das mesmas vogais mas extraídas de outras palavras, conduziu a resultados perceptualmente comparáveis, com índices de inteligibilidade e correção linguística próximos dos sinais sussurrados originais e semelhantes entre si.

4.6 Síntese do Capítulo

Em súmula, a representatividade da envolvente obtida de uma média simples de vogais extraídas de palavra mostrou-se sofrível devido à má relação sinal-ruído e à fraca definição das formantes. A implementação de estratégias de melhoria do contorno da envolvente, promovendo formantes mais definidas, registou efeitos positivos, melhorando a inteligibilidade e aproximando a vogal sussurrada sintética da vogal sussurrada original. Concluiu-se também pela adequação do modelo obtido via LPC de 22ª ordem, desde que implementadas estratégias de descarte das *frames* menos representativas da vogal ou de melhorias do contorno da envolvente.

A síntese de sussurro com recurso a envolventes com origem na vogal vozeada correspondente permitiu concluir que a envolvente obtida do resíduo preserva a informação linguística e produz resultados equiparáveis aos de envolventes de sussurro. A envolvente resíduo partilha também características com a envolvente da componente periódica (localização das formantes) e com a envolvente da versão sussurrada (natureza plana da PSD), o que poderá revelar-se útil nos processos de vozeamento sintético de sinais sussurrados.

O próximo capítulo será dedicado à implementação de um algoritmo protótipo para identificação de vogais sussurradas, desenvolvido em ambiente MATLAB.

Capítulo 5

Identificação de Vogais em Fala Sussurada

Neste capítulo propõe-se uma modelização compacta e computacionalmente económica das características espectrais de vogais sussurradas para fins de classificação e identificação de fonemas, em particular para as nove vogais orais indicadas no capítulo anterior. Muito embora a abordagem proposta para a extração de características se assemelhe a outras técnicas de processamento de sinais de fala já existentes, a versão desenvolvida nesta dissertação foi conceptualizada tendo em conta as características e requisitos específicos do projeto DyNaVoiceR.

5.1 A Metodologia para a Identificação de Vogais

O reconhecimento de voz no contexto de fala sussurrada representa um desafio mais complexo do que o que ocorre com a fala normalmente vozeada. Esta dificuldade maior fica essencialmente a dever-se às características energéticas do sussurro que resultam numa pior relação sinal-ruído e à ausência de uma importante componente do sinal de voz, a componente periódica derivada do contributo das pregas vocais. A fasquia eleva-se ainda mais no caso presente porquanto se pretende que a aplicação que o projeto DyNaVoiceR se propõe a desenvolver opere em tempo-real. Isto significa que os algoritmos de classificação de fonemas terão de ser eficientes, económicos e expeditos do ponto de vista computacional. Por outro lado, tendo em consideração o tipo de aplicação que é visada, não se entende como pertinente que a capacidade de identificação de fonemas seja independente do orador, admitindo-se a calibração da aplicação para cada utilizador. Por esta ordem de razões, idealizou-se um sistema de identificação de vogais sussurradas, assente num princípio relativamente simples em que cada segmento de sinal a classificar é comparado com uma biblioteca de fonemas que é específica para cada utilizador, composta pelos modelos de referência de cada uma das vogais. No sistema proposto, tanto o segmento a classificar como os modelos de referência constituem representações das características do sinal,

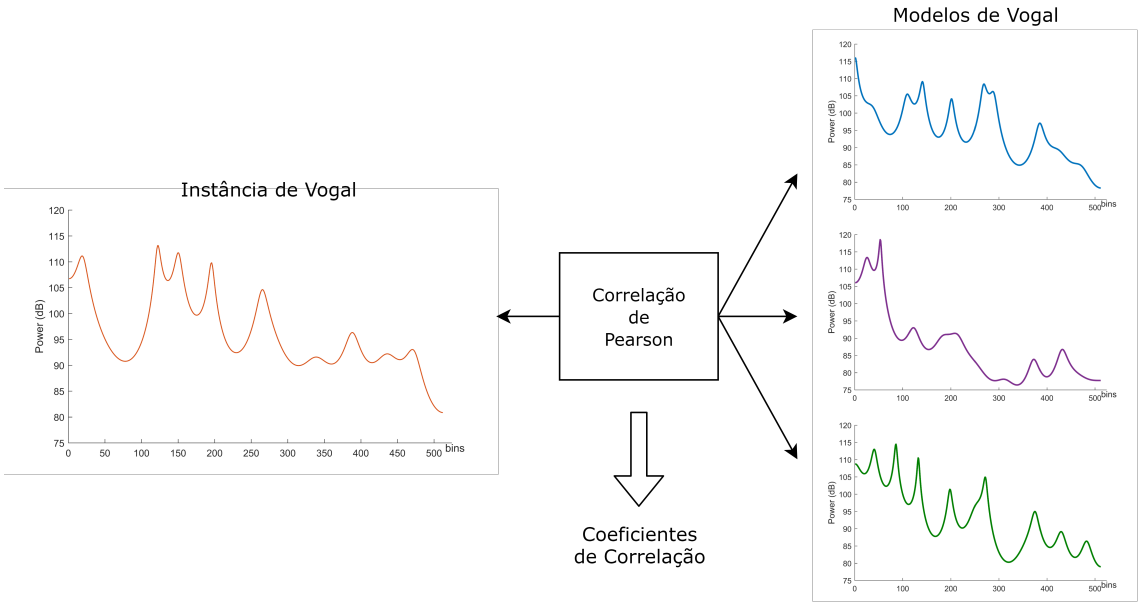


Figura 5.1: Abordagem proposta para a classificação dos segmentos de fala sussurrada: o modelo representativo de cada segmento é comparado com uma biblioteca de referência do orador (constituída no exemplo, para fins ilustrativos, por três fonemas diferentes) gerando um vetor com os respetivos coeficientes de correlação.

obtidas da sua envolvente espectral. A discriminação da vogal será inferida da correlação estatística entre as características espectrais desse segmento de sinal e cada um dos fonemas de referência.

O esquema de classificação proposto, ilustrado na figura 5.1, permitirá gerar um vetor composto pelos 9 coeficientes de correlação, respeitante à correlação entre o segmento de sinal e cada uma das 9 vogais orais de referência que constituem a biblioteca de vogais do orador. O método utilizado para determinar estes coeficientes, assim como as razões que justificaram essa escolha, encontram-se detalhados de seguida.

5.1.1 O Coeficiente de Correlação de Pearson

O coeficiente de correlação de Pearson, desenvolvido na sua forma atual por Karl Pearson, permite medir a correlação linear entre duas séries de variáveis, A e B [70]. Segundo este critério, o coeficiente de correlação ρ entre as duas séries pode ser obtido através da expressão:

$$\rho(A,B) = \frac{1}{N-1}\sum_{i=1}^{N}\left(\overline{\frac{A_i-\mu_A}{\sigma_A}}\right)\left(\frac{B_i-\mu_B}{\sigma_B}\right) \tag{5.1}$$

onde μ_A e σ_A representam a média e desvio padrão de A, e μ_B e σ_B representam a média e desvio padrão de B. Em alternativa, facultando possivelmente uma descrição mais intuitiva do significado estatístico do coeficiente, a correlação de Pearson poderá ser

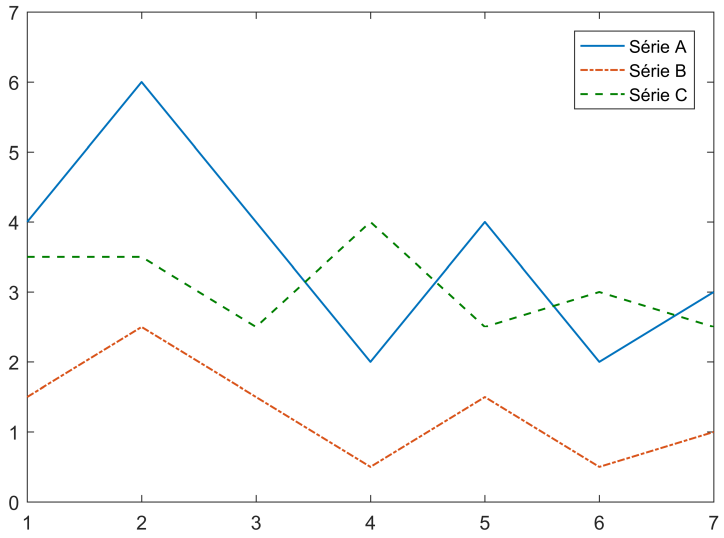


Figura 5.2: Ilustração de 3 séries representando simplificadamente 3 envolventes espectrais, para a análise da aplicação da correlação de Pearson à comparação de envolventes.

entendida como a razão entre a covariância (entre as duas séries) e o produto dos respetivos desvios padrão e por essa razão, como corolário da desigualdade de Cauchy-Schwarz, o valor do coeficiente encontra-se intrinsecamente normalizado entre -1 e 1. Um valor de ρ igual a 1 representa uma correlação linear positiva perfeita, um valor igual a 0 significa que não existe correlação linear entre as séries enquanto que um valor de -1 quererá dizer que as séries apresentam uma correlação linear negativa perfeita.

Em termos práticos, se as duas séries constituírem representações de envolventes espectrais e apresentarem uma determinada formante na mesma região do espectro, essa correspondência irá contribuir para um valor positivo de ρ. No limite, se as duas séries apresentam as várias formantes e depressões no espectro aproximadamente nas mesmas regiões, o valor da correlação tenderá a aproximar-se de 1. Dado ainda que a cada série é subtraída a respetiva média e este resultado dividido pelo desvio padrão da série correspondente, o coeficiente de correlação será por consequência insensível a qualquer deslocação vertical (*i.e.*, em magnitude) de qualquer uma das envolventes e a qualquer fator de escala, sendo para todos os efeitos práticos afetado essencialmente pela localização das formantes e das depressões no espectro, sempre com respeito ao valor médio da envolvente. Este comportamento é ilustrado pelas séries A, B e C representadas na figura 5.2. Neste exemplo, para fins meramente ilustrativos, consideremos:

- A=[4.0 6.0 4.0 2.0 4.0 2.0 3.0]
- B=(A-1)/2
- C=[3.5 3.5 2.5 4.0 2.5 3.0 2.5]

Para estes valores, o coeficiente de correlação entre A e B será obviamente de 1 (*i.e.*, correlação perfeita), o que contrasta com a correlação entre A e C (ou entre B e C) que será aproximadamente nula (-0.0561). Se assumirmos que as séries representam as envolventes espectrais de vogais, o coeficiente de correlação expressaria claramente que A e B dizem respeito à mesma vogal, enquanto que C deverá dizer respeito a uma outra vogal diferente, não obstante as visíveis diferenças de magnitude e de taxas de variação entre A e B ou a maior proximidade entre os valores médios de A e C. Finalmente, é importante sublinhar que alteração de um determinado valor nestas séries (*e.g.*, o valor de B_7) poderá alterar drasticamente estes resultados em função do seu eventual impacto na média e desvio padrão da respetiva série.

O cálculo expedito do coeficiente, a normalização intrínseca e o comportamento geral descrito, confere ao coeficiente de correlação características adequadas para o propósito de identificação de vogais, no contexto de um algoritmo que se pretende capaz de operar em tempo-real. Foram igualmente considerados, e testados em estágios subsequentes, métodos alternativos como o coeficiente de correlação tau de Kendall e uma versão modificada da autocorrelação, tendo-se concluído pelo maior potencial do coeficiente de correlação de Pearson, uma vez que conduziu aos resultados mais consistentes e robustos.

5.1.2 Uma Primeira Avaliação da Metodologia

Procurando-se, numa primeira instância, avaliar a metodologia proposta, em particular a capacidade do coeficiente de correlação de Pearson de identificar a vogal correta, procedeu-se a um estudo preliminar de correlações cruzadas em que se fez uso das envolventes espetrais LPC de 22ª ordem das vogais sussurradas sustentadas e extraídas de palavra, já disponíveis. Cada fonema será representado por um vetor com 512 coeficientes, relativos à envolvente LPC obtida da Power ODFT média do segmento correspondendo a cada fonema, considerando os 60% das *frames* mais centrais de cada segmento de sinal a analisar (*i.e.*, são descartados os 20% das *frames* iniciais e os 20% das *frames* finais), localizando o fonema com base na anotação fonética disponível na base de dados. Nesta primeira análise consideraram-se 4 oradores (2 masculinos, 2 femininos) e as 3 vogais mais frequentes na base de dados, as quais se encontram situadas nos extremos do triângulo acústico das vogais: o /á/ (de 'água'), o /i/ (de 'ilha') e o /u/ (de 'uva'). Para construir o banco de vogais de referência, utilizaram-se as segundas repetições de cada uma das 9 vogais sustentadas, do mesmo orador, no modo sussurrado.

O estudo de correlações, realizado individualmente por orador e para cada uma das 3 vogais selecionadas, incluiu o cálculo da correlação entre cada ocorrência da vogal nos dissílabos (isolados) disponíveis na base de dados e cada uma das 9 vogais utilizadas para referência. No final, determinaram-se as médias e desvio padrão registados por cada modelo de referência. Adicionalmente, com o objetivo de melhor avaliar a viabilidade e adequação de se utilizar a vogal sustentada como fonte dos modelos de referência, calcularam-se, de forma semelhante, os coeficientes de correlação entre cada ocorrência da vogal em palavra

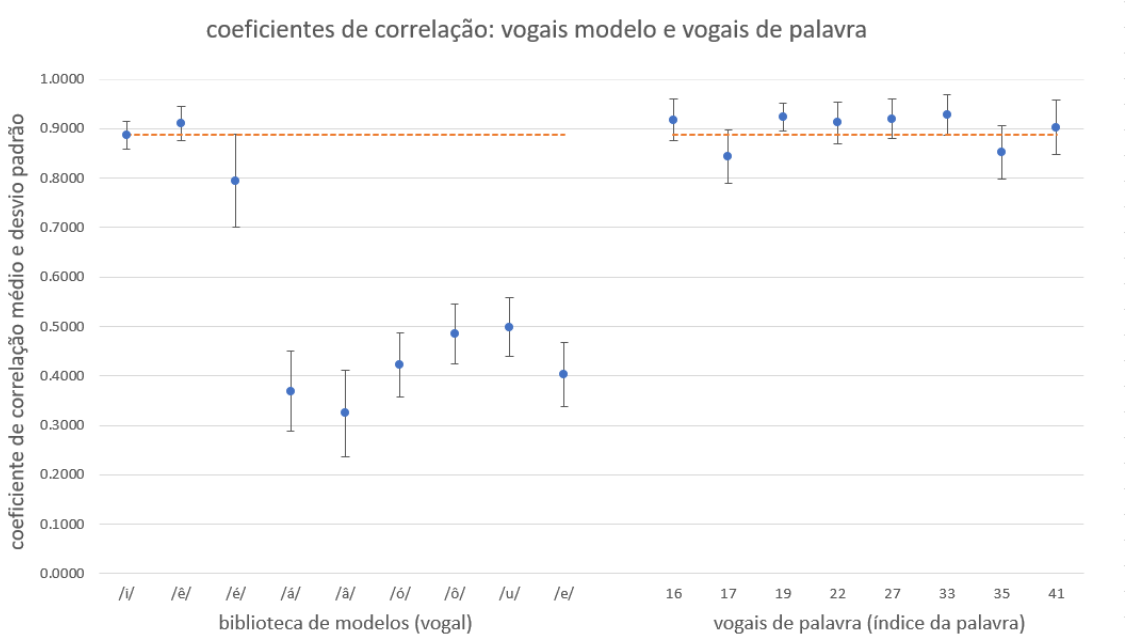


Figura 5.3: Identificação da vogal /i/ para o orador F07; desempenho dos modelos de referência das vogais (à esquerda) e das restantes vogais /i/ em palavras (à direita).

e todas as restantes ocorrências dessa mesma vogal nas restantes palavras. Este cenário equivale a considerar que o modelo de referência é constituído por uma instância dessa vogal ocorrida em contexto de palavra. Dois fatores, com potencial impacto na qualidade ou representatividade dos modelos de referência, motivaram esta avaliação adicional: por um lado, os segmentos contendo vogais obtidas de palavra contêm um número médio de *frames* substancialmente menor (com frequência, 10 *frames* ou menos) quando comparado com as vogais sustentadas (formadas tipicamente, em virtude da sua prolongada duração, por várias dezenas de *frames*); por outro, o modo como a vogal é gerada (sustentado ou em contexto de palavra) poderá condicionar a forma como a vogal é articulada pelo orador.

A Fig. 5.3 exemplifica, num cenário típico para os testes conduzidos, os resultados consolidados para a vogal /i/ para um dos oradores femininos. Neste caso, o modelo de referência da vogal /i/ desse orador (retirado de vogal sustentada) registou o segundo valor mais elevado das médias das correlações e o menor desvio padrão entre as nove vogais de referência. Também é possível observar, graças à linha horizontal auxiliar a tracejado, que 6 das 8 vogais /i/ retiradas de palavra registam melhor desempenho que o /i/ sustentado (nota: a correlação de uma ocorrência da vogal com ela própria, que será necessariamente 1, nunca é considerada na média). As duas outras vogais de referência com valores médios elevados são neste caso, conforme se poderá observar, o /ê/ (de 'p̲eso') e o /é/ (de 'e̲la'), duas das vogais mais próximas no espaço das vogais.

Este primeiro estudo de correlações serviu também para experimentar com diferentes critérios e concluir numa primeira análise o impacto de diversas variações no procedimento,

do que se destaca:

- A substituição da escala logarítmica (em dB) para as magnitudes por uma raiz de potência, uma alternativa utilizada em diversas técnicas de processamento de sinais acústicos [71, 72], não indicou trazer benefícios ao passo que a substituição por uma escala linear indicou ser completamente inadequada;
- Compararam-se as escalas linear e Bark para as frequências, tendo a escala psicoacústica, de uma forma geral, revelado um melhor desempenho;
- No que diz respeito ao critério de escolha das vogais de referência, a análise preliminar sugere que as vogais obtidas de palavra serão melhores candidatas para a construção de modelos de referência do que as vogais sustentadas;

Uma outra observação, consistente nas 3 vogais analisadas e para o conjunto dos oradores testados, é a de que os modelos de referência com classificação mais elevada segundo o coeficiente de correlação de Pearson, dizem precisamente respeito às vogais mais próximas da correta no espaço das vogais (ver figura 5.4), revelando uma relação próxima entre a correlação estatística e a proximidade fonética:

- para o /i/ (de 'ilha'), o /é/ (de 'ela') e o /ê/ (de 'peso');
- para o /á/ (de 'água'), o /â/ (de 'amarelo') e o /ó/ (de 'óculos');
- para o /u/ (de 'uva'), o /ô/ (de 'ovo') e o /ó/ (de 'óculos').

Esta observação sugere duas implicações: por um lado, a proximidade fonética das vogais poderá causar dificuldades na identificação precisa da vogal, em linha com o que ocorre na identificação de fonemas quando realizada por ouvintes humanos (ver 2.2.2); por outro lado, as situações de identificação incorreta poderão eventualmente ser mitigadas pela própria proximidade fonética das vogais selecionadas ou através de estratégias de interpolação, sendo plausível que se consiga preservar o significado linguístico pretendido, do ponto de vista percetual. Finalmente, nesta primeira implementação, verificou-se também que as vogais de referência correspondentes (*i.e.*, as foneticamente corretas) obtêm, de forma consistente, dos coeficientes de correlação mais elevados entre as nove candidatas.

Considerando como critério de acerto simplesmente o modelo de referência com o coeficiente de correlação mais elevado, obtiveram-se nesta avaliação preliminar, para o conjunto dos 4 oradores, taxas de até 94% para o /á/, 81% para o /i/ e 42% para o /u/, dependendo das diversas variações e opções testadas. Estes resultados, reforçados pelo facto de a generalidade das identificações incorretas terem ocorrido por confusão com vogais foneticamente próximas e pela vogal correta se apresentar de forma sistemática entre os coeficientes de correlação mais elevados, conduziram ao desenvolvimento de modelos compactos para a representação dos fonemas e de um algoritmo protótipo de identificação de vogais sussurradas, onde se fez uso das conclusões obtidas até este ponto e se contemplaram melhorias e otimizações adicionais.

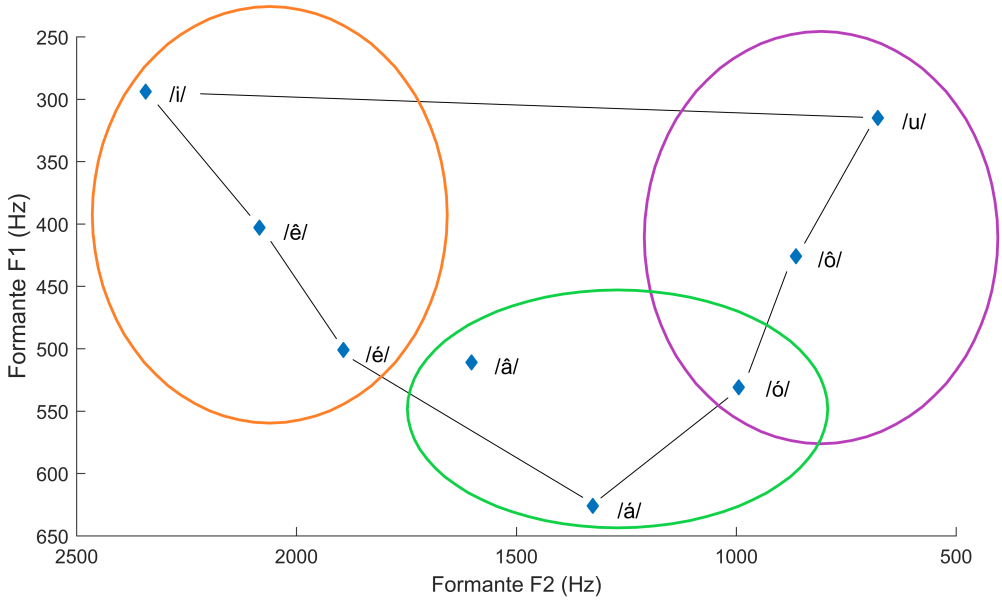


Figura 5.4: Regiões de confusão das vogais testadas no triângulo acústico das vogais: região do /u/ a roxo; região do /á/ a verde; região do /i/ a laranja.

5.2 Algoritmo de Identificação de Vogais Sussurradas

O estágio seguinte consistiu em implementar um algoritmo protótipo de identificação de vogais sussurradas baseado na abordagem geral apresentada na secção anterior e que se encontra sintetizado no diagrama de blocos da figura 5.5. O algoritmo agora implementado, para além de sistematizar os procedimentos atrás descritos, substitui a envolvente LPC por um modelo compacto (*i.e.*, formado por um número reduzido de coeficientes) contendo as características espectrais de cada segmento de sinal. Este algoritmo permite ao utilizador selecionar os segmentos de sinal a analisar, de entre os disponíveis na base de dados, e a manipulação de parâmetros como o tipo de escala em frequência, controlar a resolução espectral a utilizar ou selecionar diferentes modelos de referência, facultando, deste modo, a realização de testes exaustivos sobre a base de dados e a avaliação do desempenho das diferentes alternativas.

5.2.1 Biblioteca de Modelos de Referência das Vogais

Tendo por base a análise realizada em 5.1.2 e a sugestão de que as versões sustentadas não deverão constituir as melhores candidatas para vogais de referência, optou-se por construir duas versões alternativas da biblioteca de modelos de referência das nove vogais orais, sempre por orador. A primeira versão é construída exclusivamente com base nas instâncias sustentadas das vogais. A segunda versão inclui as versões sustentadas e as ocorrências em palavras. Ainda que fosse desejável construir modelos (médios) alternativos baseados exclusivamente em vogais obtidas de palavras, tal mostrou não ser viável em

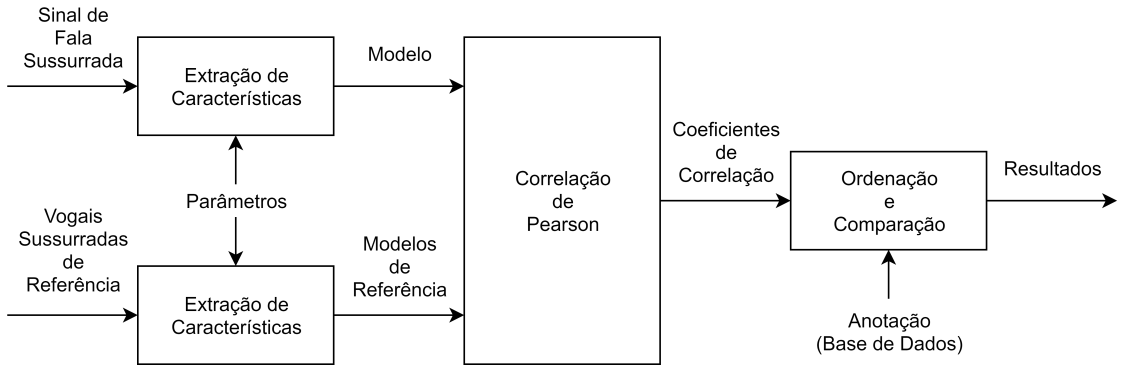


Figura 5.5: Diagrama de blocos do algoritmo protótipo de identificação de vogais sussurradas.

virtude da enorme variação na frequência de ocorrência de diferentes vogais na base de dados e, inclusivamente, da indisponibilidade de determinadas vogais para determinados oradores. Deste modo, a segunda versão representa uma solução de compromisso, permitindo inferir o impacto de se utilizarem vogais extraídas de palavra na construção dos modelos de referência, por comparação com os modelos obtidos exclusivamente das vogais sustentadas.

5.2.2 Extração de Características

Na implementação do algoritmo protótipo de identificação de vogais sussurradas, as envolventes espectrais da figura 5.1, obtidas por modelização LPC, são substituídas agora por novos modelos compactos, formados por um número reduzido de coeficientes representando as características espectrais dos segmentos de sinal a analisar. Os modelos, a serem construídos durante a execução do algoritmo, foram concebidos em duas variantes: uma primeira versão representando o sinal no domínio espectral e uma segunda em que o modelo representa as características do sinal no domínio cepstral, seguindo um procedimento que se assemelha ao utilizado no cálculo de MFCCs. O procedimento de extração de características, sintetizado no diagrama de blocos da figura 5.6, aplicado quer aos segmentos de sinal a analisar, quer às vogais de referência, será descrito em detalhe de seguida.

5.2.2.1 Domínio das Frequências e Banco de Filtros

O sinal a processar, seguindo um procedimento semelhante ao descrito no capítulo anterior e ilustrado na figura 4.1a, é primeiramente segmentado em janelas de 1024 amostras, com passos de 512 amostras (*i.e.*, 50% de sobreposição) e transportado para o domínio das frequências pela ODFT, uma variante da *Discrete Fourier Transform* em que cada *bin* se encontra posicionado nos múltiplos ímpar de $\frac{\pi}{N}$, com as implicações já atrás descritas, da qual se derivará a Power ODFT de cada *frame* (ver 4.2).

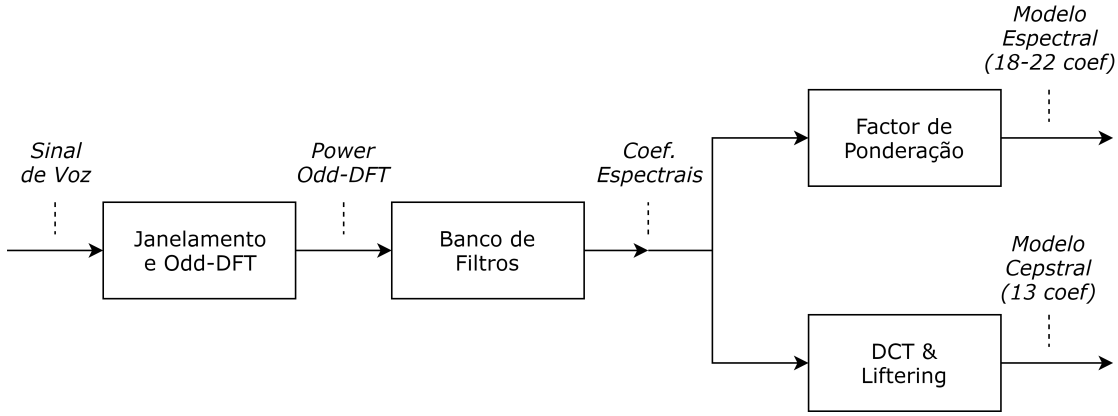


Figura 5.6: Diagrama de blocos da extração das características da vogal sussurrada.

O sinal, agora representado no domínio espectral pelos 512 coeficientes únicos da Power ODFT, será sujeito à aplicação de um banco de filtros paralelos triangulares e parametrizável. Este banco de filtros cumpre dois objetivos. Por um lado, graças à definição de diferentes filtros, permite controlar a resolução espectral e a banda útil utilizada para representar as características espectrais do sinal, possibilitando desde logo o descarte de regiões ou detalhe espectral pouco informativos. Por outro, controlando o posicionamento e distribuição das bandas de cada filtro, permite obter representações espectrais seguindo diferentes lógicas não lineares, aplicando diferentes escalas em frequência que possam refletir melhor as características percetuais do sistema auditivo. Este banco de filtros poderá ser expresso por:

$$H_i(k) = \begin{cases} 0 & , \ k \leq f_{i-1} \\ \dfrac{k - f_{i-1}}{f_i - f_{i-1}} & , \ f_{i-1} < k < f_i \\ 1 & , \ k = f_i \\ \dfrac{f_{i+1} - k}{f_{i+1} - f_i} & , \ f_i < k < f_{i+1} \\ 0 & , \ k \geq f_{i+1} \end{cases} \tag{5.2}$$

onde i se refere ao índice do filtro, k ao *bin* correspondente na representação espectral original (*i.e.*, a Power ODFT) e fi à frequência de índice i na escala escolhida. Estas frequências fi encontram-se uniformemente distribuídas em cada uma das quatro escalas disponíveis no algoritmo: a escala linear, a escala *mel*, a escala Bark e a escala ERB. Cada uma destas escalas está disponível sobre a forma de vetores contendo as frequências fi calculadas de acordo com as expressões e tabelas apresentadas no capítulo 2 (ver 2.2.2). Adicionalmente, o ganho de cada filtro é manipulado por forma a assegurar uma resposta plana no espectro, o que é conseguido garantindo que todos os triângulos têm a mesma área, resultando num banco de filtros como o ilustrado, para a escala Bark, na figura 5.7. Esta

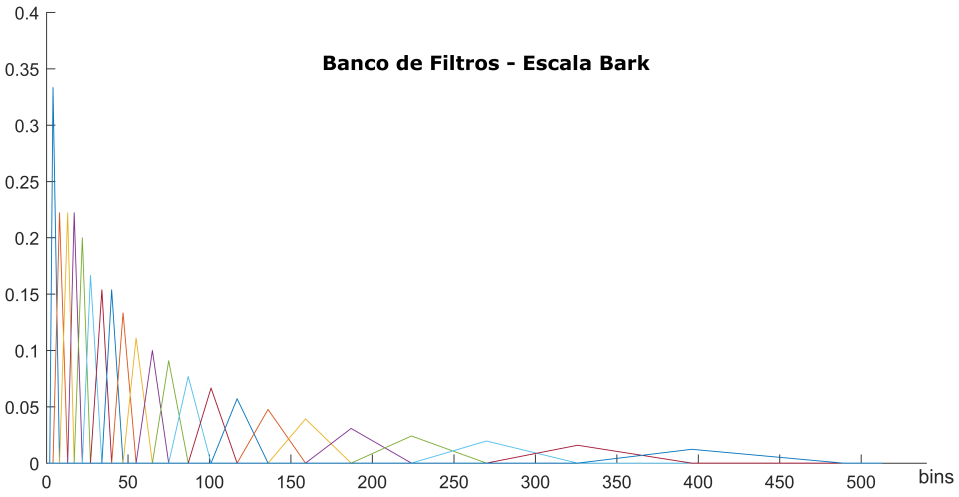


Figura 5.7: Banco de filtros triangulares uniformemente espaçados na escala Bark (inclui ajuste de ganho que garante que o banco de filtros tenha uma resposta plana em frequência).

flexibilidade do banco de filtros permitirá comparar o desempenho de diversas alternativas à escala *mel*, que apesar de comummente utilizada (nomeadamente nos MFCCs) tem sido objeto de algum criticismo ao longo dos anos (ver 2.2.2).

O algoritmo permite também controlar o número de bandas utilizadas para o banco de filtros. É possível, por um lado, controlar a resolução espectral do banco de filtros (*i.e.*, a densidade dos bancos) e, por outro, definir a banda útil estipulando o número máximo de bandas, o que poderá permitir uma redução da complexidade computacional dos modelos sem prejuízo do desempenho do algoritmo de identificação. Uma vez que a frequência de amostragem utilizada é de 22050 Hz, a largura de banda total do sinal (11025 Hz) corresponde a 22 bandas Bark completas (o limite superior da 23ª banda Bark excede o espectro disponível). Optou-se por essa razão e por uma questão de conveniência utilizar as alternativas de um máximo de 22 bandas (frequências centrais espaçadas de 1 Bark) ou de 44 semi-bandas (frequências centrais espaçadas de 0.5 Bark). Por forma a garantir que o desempenho das diferentes escalas é comparado sob as mesmas condições de resolução/espectro, tomou-se por referência para qualquer uma das alternativas (linear, *mel*, ERB) o extremo superior do espectro na configuração Bark.

5.2.2.2 Coeficientes no Domínio Espectral

Na sua versão mais simples, as características espectrais do sinal de voz são representadas pelos coeficientes espectrais obtidos do banco de filtros e sujeitos a uma última transformação. Esta modificação consiste na aplicação de uma curva de ponderação (empírica) cujos pesos enfatizam especialmente os coeficientes correspondem às duas primeiras

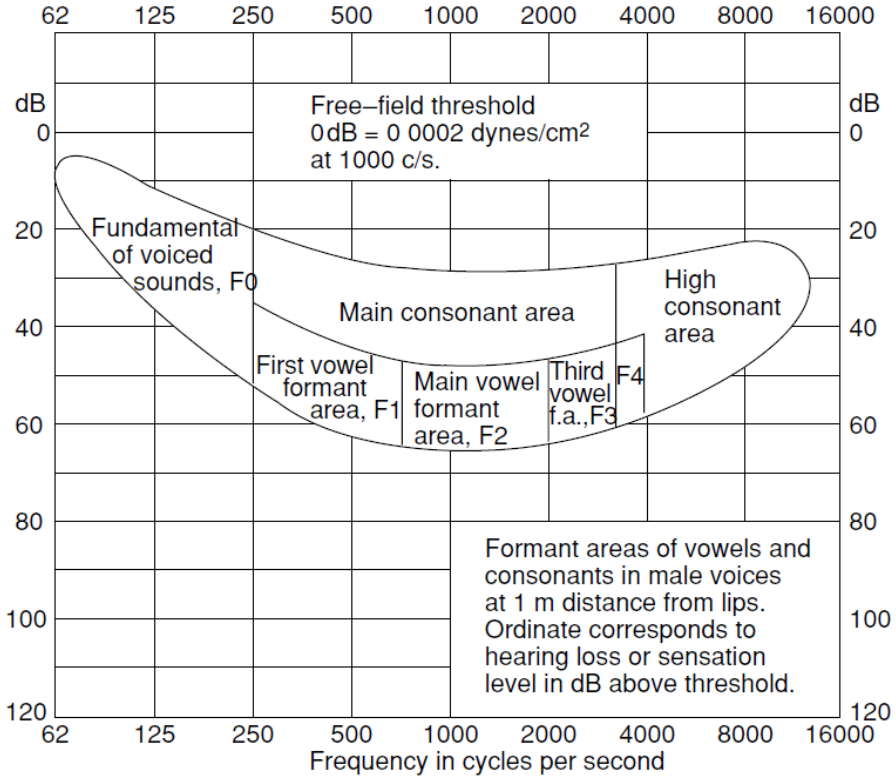
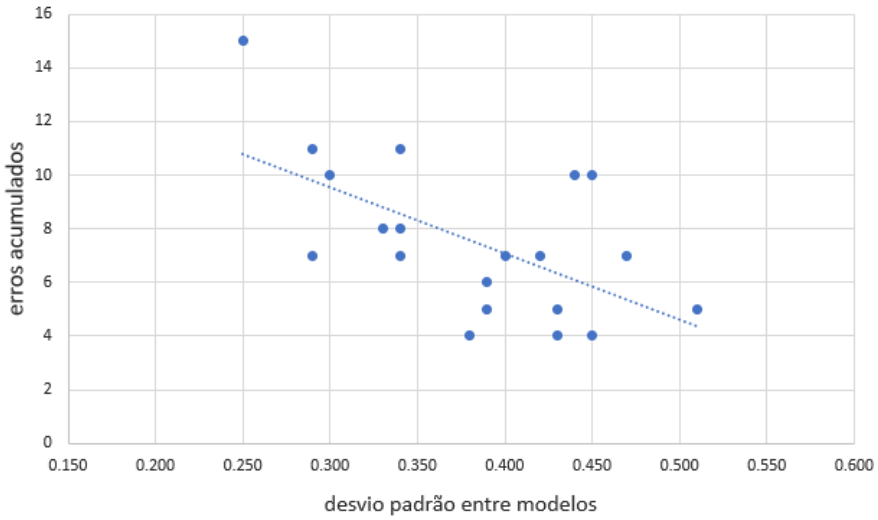


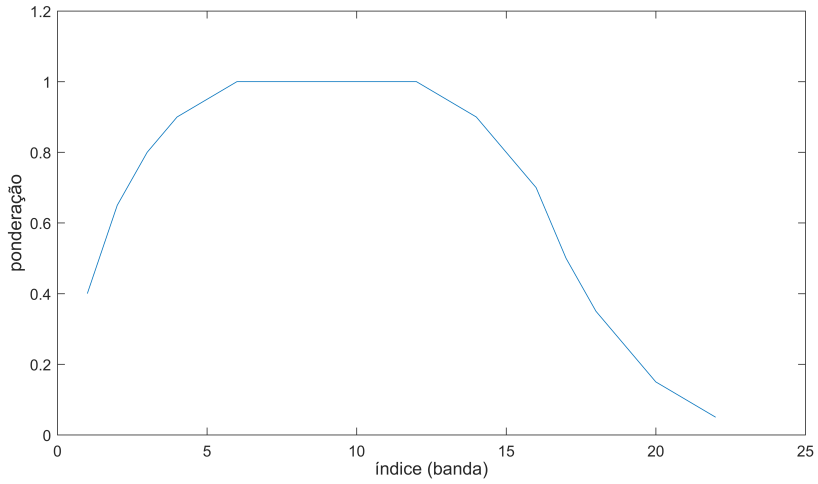
Figura 5.8: Região das formantes, ao centro no gráfico (adaptado de G. Fant [5]).

formantes (F1 e F2), sensivelmente entre a 5ª e 13ª banda na escala Bark, e a terceira formante (F3) em menor extensão. Também para assegurar, neste processo, uma modificação conveniente dos modelos, estes são em primeiro lugar centrados em torno do seu valor médio. Esta manipulação, que não tem *per se* qualquer impacto no cálculo dos coeficientes de correlação, uma vez que o coeficiente de Pearson é imune às deslocações no eixo vertical como já se viu, garante porém que a curva de ponderação tenha o efeito pretendido no modelo qualquer que seja a curva espectral original, nomeadamente, enfatizar as depressões e formantes na região de interesse do espectro em detrimento das restantes.

A curva de ponderação proposta, para além de enfatizar a gama de frequências de maior interesse para a identificação de vogais, tem ainda uma segunda importante motivação: promover uma maior diferenciação entre os modelos das vogais de referência de cada orador, com o objetivo de melhorar a capacidade do algoritmo distinguir e identificar corretamente as vogais nos segmentos de sinal em análise. Com efeito, verificou-se que o desempenho do algoritmo tende a degradar-se quando o desvio padrão entre os modelos de referência do orador decresce, conforme se ilustra na figura 5.9a onde se compara o desempenho do algoritmo (número de erros de identificação acumulados) e o desvio-padrão entre os modelos de referência. A promoção, pela curva de ponderação, de uma maior diferenciação entre modelos, explica-se por garantir que os coeficientes nos extremos superior e inferior do espectro têm valores próximos do valor médio, não condicionando o valor da correlação.



(a) Erros de identificação acumulados em função do desvio padrão entre os 9 modelos de referência de vogais, por orador.



(b) Curva de ponderação aplicada aos coeficientes espectrais, enfatizando as duas primeiras formantes e atenuando o peso dos coeficientes nos extremos do espectro (bandas na escala Bark).

Figura 5.9: Relação entre erros acumulados e desvio padrão entre modelos de referência (a) e curva de ponderação empírica implementada (b).

A existência de valores sistematicamente bastante desviados da média, para um dado orador, fora das zonas de interesse do espectro, teria como corolário o aplanamento dos modelos na região de interesse por via do aumento do desvio padrão de cada modelo e como resultado, uma maior semelhança entre modelos de vogais distintas (ver análise em 5.1.1). Deste modo, atenuando o peso dos coeficientes nas regiões de menor interesse, garante-se por um lado que é a região das formantes a que mais pesa no cálculo da correlação entre o segmento de sinal a analisar e os modelos de referência das vogais, prevenindo em simultâneo que a presença de valores idiossincraticamente baixos/elevados em regiões do

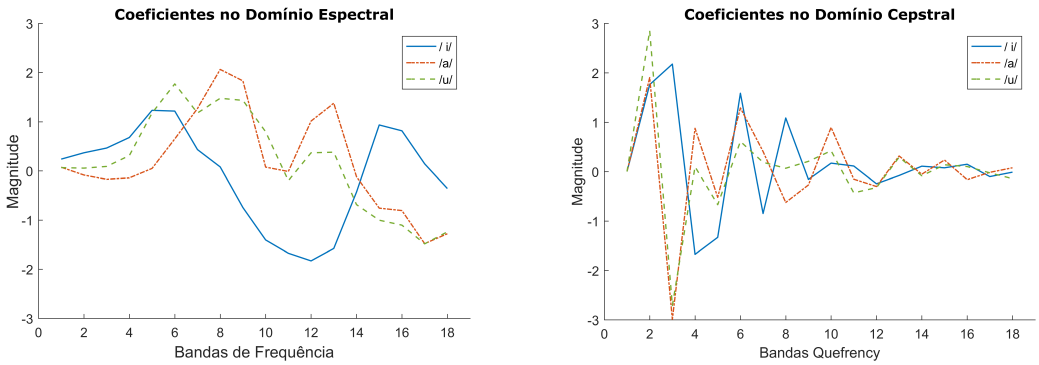


Figura 5.10: Exemplos dos modelos das vogais /i/, /a/ e /u/ nos domínios espectral (à esquerda) e cepstral (à direita).

espectro pouco relevantes para a identificação de vogais afete negativamente o desempenho do algoritmo ao conduzirem a modelos de referência que se assemelham demasiadamente entre si. Os valores da curva de ponderação utilizada, exemplificada na figura 5.9b, foram obtidos experimentalmente, tendo-se tomado por referência a região tipicamente ocupada pelas primeiras formantes, a destacar, conforme disponíveis na literatura (ver figura 5.8). Na figura 5.10 encontram-se os exemplos das vogais /a/, /i/ e /u/ modelizadas por este método, no domínio espectral, para um dos oradores disponível na base de dados.

5.2.2.3 Coeficientes no Domínio Cepstral

Alternativamente, o modelo da vogal poderá ser obtido com base numa análise do tipo cepstral seguindo um procedimento idêntico ao aplicado nos MFCCs em que a passagem para o domínio cepstral se efetua por meio da *Discrete Cosine Transform* (DCT). Esta transformação aplica-se aos coeficientes obtidos do banco de filtros anterior (ver figura 5.6) que contêm a energia na escala logarítmica (em dB) apurada para cada filtro. Os coeficientes c_j desta DCT poderão ser obtidos pela expressão:

$$c_j = \sum_{i=1}^{N} E_i \cos\left(\frac{j\pi}{N}\left(i - 0.5\right)\right) \tag{5.3}$$

onde j representa o índice do coeficiente da DCT, N corresponde ao número de canais da transformada e E_i à energia apurada para o filtro de índice i, à saída do banco de filtros.

A DCT, nas suas diversas variantes, é frequentemente adotada em processamento de sinal em virtude da respetiva eficiência computacional e pelo facto de operar apenas no domínio real, razões estas que estão por detrás da sua utilização nos MFCCs em particular e da adoção no presente algoritmo também. Outra importante vantagem da utilização da DCT, que pode ser inferida dos exemplos da figura 5.10, está na sua capacidade em descorrelacionar os coeficientes quando comparado com a representação espectral correspondente. Com efeito, pode observar-se facilmente que na versão espectral os coeficientes tendem a

apresentar covariâncias elevadas. Ou seja, os coeficientes correlacionam-se bastante bem com os seus vizinhos, o que se entende dado que cada formante não é constituída por um único coeficiente, mas antes por vários coeficientes vizinhos entre si, o que implica que exista informação redundante nos coeficientes espectrais. Ao aplicar a DCT, observa-se que a informação relevante fica concentrada nos primeiros coeficientes, bastante descorrelacionados entre si, enquanto a cauda da DCT (*i.e.*, nos valores de *quefrency* mais elevados) contém relativamente pouca informação, podendo ser descartada. Esta operação que consiste em truncar a um determinado número de coeficientes no domínio cepstral é conhecida por *liftering*. A DCT desempenha, portanto, um papel de compressão também, permitindo compactar as características espectrais num menor número de coeficientes, com as decorrentes vantagens computacionais.

Optou-se, na versão cepstral dos modelos de vogal, por testar o algoritmo de identificação de vogais sussurradas utilizando os 13 primeiros coeficientes cepstrais, sensivelmente em linha com a prática corrente em MFCCs. Se o passo adicional da transformada DCT e a redução do número de coeficientes se justifica, dependerá de um compromisso entre *a)* o desempenho do algoritmo com cada uma das versões e *b)* a complexidade computacional das operações a que os coeficientes são sujeitos, o que só será possível aferir plenamente numa fase posterior do projeto.

5.2.3 Análise *off-line*

O algoritmo de identificação de vogais implementado foi testado em duas variantes distintas permitindo dois modos de análise. A primeira teve como objetivo uma análise *off-line* exaustiva sobre os 20 oradores da base de dados que foram previamente selecionados, por forma a avaliar a capacidade do algoritmo em identificar a vogal correta e computar as respetivas taxas de acerto. Nesta aplicação do algoritmo, a anotação manual contida na base de dados é utilizada para determinar a localização das vogais na base de dados e para comparar a previsão do algoritmo (com base no vetor de coeficientes de correlação) com a vogal identificada de acordo com a anotação, que é utilizada como base de referência. Isto é, se o valor mais elevado no vetor dos coeficientes de correlação diz respeito a uma vogal de referência que condiz com a anotação manual para o correspondente segmento de sinal, considera-se então que o algoritmo identificou corretamente a vogal (acerto total). O algoritmo determina ainda se a vogal de referência correta se encontra entre os 3 valores mais elevados no vetor (sucesso relativo), uma indicação do potencial do algoritmo, que poderá vir a incorporar processamento adicional de natureza estatística (*e.g.*, Hidden Markov Models) e/ou mitigação dos erros de identificação via métodos de interpolação entre fonemas próximos no espaço das vogais. Finalmente, os resultados são consolidados, segundo as duas métricas indicadas, para todas as ocorrências de vogal nos segmentos de sinal testados. O resultado desta análise (cujos valores se encontram sintetizados no anexo B) será discutido em detalhe na secção 5.4.

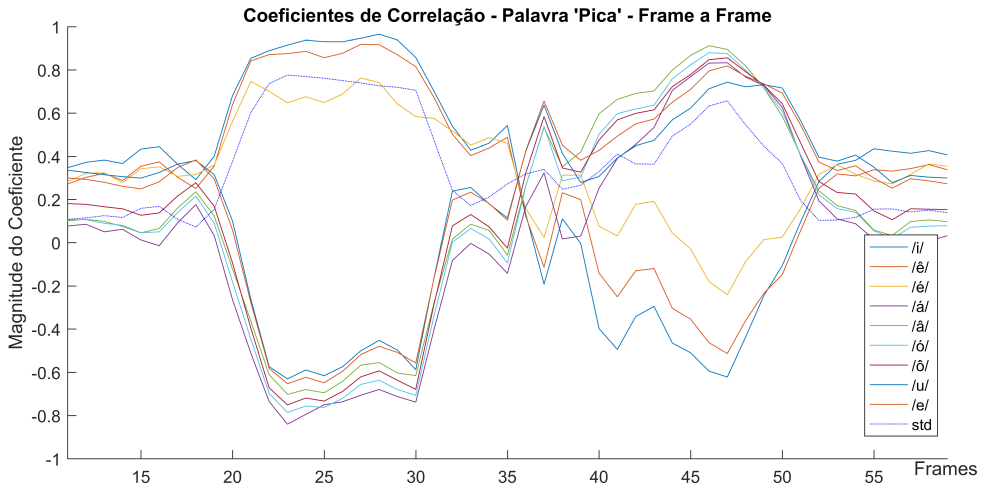


Figura 5.11: Desempenho dos 9 coeficientes de correlação relativos às vogais de referência ao longo da palavra portuguesa *'pica'* (média das últimos duas *frames*).

5.2.4 Análise *on-the-fly*

A segunda variante do algoritmo permite analisar qualquer segmento de sinal da base de dados *frame* a *frame*, apurando a média dos coeficientes de correlação para as duas *frames* mais recentes, à medida que o segmento escolhido é percorrido. Esta versão permite, por conseguinte, perceber sobre a praticabilidade de efetuar o reconhecimento das vogais, por estes métodos, em tempo-real e num contexto de fala corrida, ainda que utilizando os estímulos disponíveis na base de dados como teste. Esta variante do algoritmo gera uma sequência temporal dos vetores de correlação que é possível visualizar também graficamente, como ilustra o exemplo da figura 5.11, permitindo realizar uma análise visual do comportamento dinâmico dos coeficientes de correlação.

5.3 Teste do Algoritmo

O protótipo, desenvolvido em ambiente MATLAB, foi testado num computador portátil com processador intel i5-5300U a 2.3 GHz. O tempo de execução por *frame*, sem otimizações e num ambiente de desenvolvimento, é inferior a 5 ms. Este valor é bastante inferior à cadência das *frames*, que à taxa de amostragem de 22050 Hz é de 23 ms, sustentando a viabilidade do algoritmo para operação em tempo-real.

O comportamento típico dos coeficientes de correlação com as 9 vogais orais de referência pode ser observado na figura 5.11, neste exemplo com a palavra portuguesa 'pica' (análise *on-the-fly*, ver 5.2.4). De uma forma geral, verifica-se, que os coeficientes tendem a divergir de forma evidente durante os períodos correspondentes a vogais, tendendo a convergir nos restantes períodos, em especial aqueles que dizem respeito a segmentos de silêncio (na imagem, o período inicial e o período final). Observa-se também no caso em

exemplo (e com certa frequência em geral) que a vogal da última sílaba apresenta características menos estáveis e mais assimétricas no tempo (entre o seu início e fim), denunciando a progressiva modificação das características espectrais do fonema à medida que se aproxima da terminação da palavra, o que poderá servir de pista para efeitos de controlo prosódico.

Com o objetivo de avaliar o desempenho do algoritmo protótipo implementado, testou-se a capacidade do algoritmo em identificar as 4 vogais mais frequentes na base de dados - o /i/ de 'ilha' (com 475 ocorrências), o /á/ de 'água' (com 647 ocorrências), o /â/ de 'amarelo' (com 1365 ocorrências) e o /u/ de 'uva' (com 459 ocorrências) - no conjunto dos 20 oradores previamente selecionados, recorrendo à variante e métricas já discutidas em 5.2.3. Esta avaliação terá um carácter objetivo se assumirmos que a identificação da vogal foi corretamente realizada durante o processo de anotação fonética, utilizada como base de referência. Os resultados consolidados das 4 vogais testadas para a totalidade dos 20 oradores, segundo as duas métricas consideradas (acerto total e sucesso relativo), discutidos em detalhe de seguida, encontram-se disponíveis no Anexo B.

5.4 Discussão de Resultados

Pretendendo-se facilitar a discussão dos resultados assim como a avaliação do impacto das diversas alternativas consideradas, toma-se como referência (ou ponto de partida para a comparação das diversas alternativas) a utilização de modelos no domínio espectral e como vogais de referência do orador, os modelos médios obtidos de ocorrências sustentadas e de palavras (ver 5.2.1). Os resultados consolidados correspondentes estão disponíveis nas tabelas B.1, B.2, B.3 e B.4.

A primeira importante observação é a de que o desempenho do algoritmo varia consideravelmente com a vogal, o que desde logo suscita a necessidade de validar o algoritmo num contexto mais alargado, idealmente incluindo pelo menos as 14 principais vogais do PE (9 vogais orais e 5 vogais nasais). De entre as 4 vogais testadas, a vogal /á/ de 'água' apresenta o melhor desempenho com taxas de acerto de cerca de 90% para as escalas não lineares, em contraste com a vogal /u/ de 'uva' (85%), /i/ de 'ilha' (70% de taxa de acerto, muito embora com taxas de sucesso de 100%) e o /á/ de 'amarelo' (inferior a 50%). O desempenho substancialmente inferior nesta última vogal deverá justificar-se pelo facto desta vogal se encontrar sempre na última sílaba na base de dados DyNaVoiceR, pelo que as suas características espectrais serão mais sensíveis a efeitos de natureza prosódica, como a figura 5.11 sugere. Por outro lado, a vogal /á/ é cerca de 50% mais longa, em média, que o /i/ e /u/ na base de dados, consistente com observações idênticas na literatura [73], o que poderá explicar o seu superior desempenho.

5.4.1 Alternativas de Escala nas Frequências

No que respeita ao impacto de diferentes escolhas para a escala nas frequências utilizada no banco de filtros, verificou-se que de uma forma geral a escala linear apresenta um

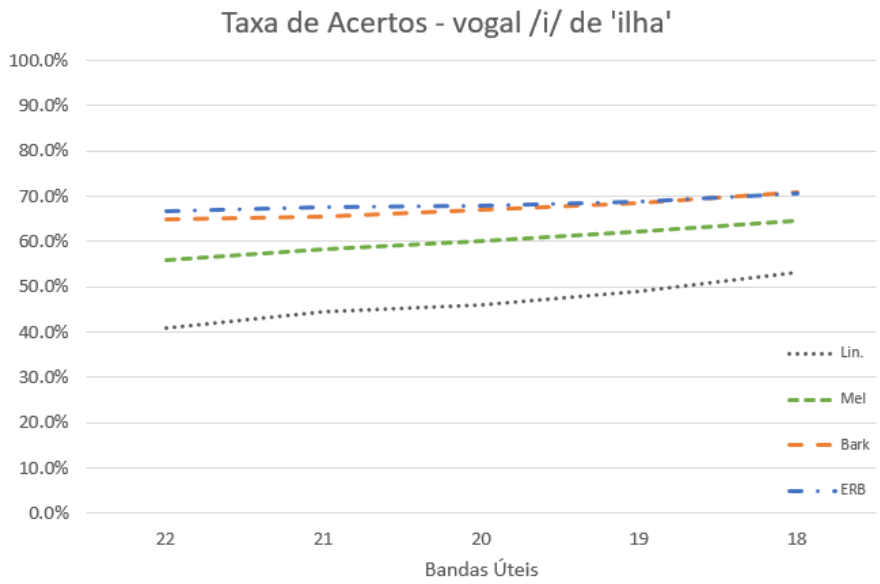


Figura 5.12: Taxas de acerto consolidadas para a vogal /i/ em função do tipo de escala em frequência e do número de bandas utilizadas (dados extraídos de B.1).

desempenho inferior a qualquer uma das escalas não lineares.

Os resultados obtidos para as escalas *mel*, Bark e ERB registam diferenças pouco expressivas entre si para 3 das 4 vogais testadas (/á/ de 'água', /â/ de 'amarelo' e /u/ de 'uva'). Globalmente, porém, a escala *mel* registou um desempenho ligeiramente inferior às duas restantes, devido essencialmente à prestação inferior com a vogal /i/. Ainda que estes resultados sugiram a maior adequação das escalas Bark e ERB, havendo ainda um conjunto considerável de vogais por testar, será prematuro descartar a escala *mel*, tanto mais que a alteração da escala não apresenta qualquer dificuldade ou implicação em termos de custos computacionais.

5.4.2 Opções de Resolução Espectral

Pretendendo avaliar se o incremento da resolução espectral dos modelos, ao reproduzir com um maior índice de detalhe as características espectrais do sinal, contribuiria favoravelmente e de forma expressiva para o desempenho do algoritmo, duplicou-se a resolução utilizando o dobro dos filtros triangulares, com bandas uniformemente espaçadas de metade da distância (*e.g.* 0.5 Bark). Os resultados consolidados para a versão de alta resolução, disponíveis nas tabelas B.9, B.10, B.11 e B.12, permitem concluir que o aumento da resolução espectral não contribui significativamente para a capacidade de identificar corretamente a vogal sussurrada, não se justificando o custo computacional acrescido.

Por seu turno, a redução do número de bandas, descartando as bandas relativas ao extremo superior do espectro, conduz a melhorias de desempenho para a vogal /i/, como a figura 5.12 ilustra. O impacto nas 3 restantes vogais é, de uma forma geral, marginal.

Ou seja, a utilização de 18 ou 19 bandas conduziu a resultados comparáveis, quando não superiores, aos obtidos com a utilização de toda a banda disponível no sinal original.

A redução e o descarte de um determinado número de bandas significa não só simplificação e redução de cálculo no banco de filtros (os filtros descartados não necessitam de ser calculados) mas também a redução do número de coeficientes nos modelos e a consequente redução do custo computacional de qualquer operação a que venham a ser sujeitos (*e.g.*, o cálculo da correlação de Pearson).

5.4.3 Modelos de Referência das Vogais

A biblioteca de modelos de referência obtida exclusivamente de vogais sustentadas resultou em taxas de acerto consistentemente inferiores, reforçando a indicação de que a produção de vogais sustentadas não constituirá a melhor metodologia para a construção de modelos de referência (ver também 5.1.2). A análise destes resultados, disponíveis nas tabelas B.13, B.14, B.15 e B.16, sugere que a articulação das vogais fora do seu contexto habitual em palavra compromete a representatividade dos modelos obtidos.

Uma vez que os modelos de referência alternativos utilizados nestes testes foram obtidos de um compósito de vogais sustentadas e de vogais retiradas de palavra, dadas as limitações da base de dados indicadas em 5.2.1, interessará complementar a base de dados com novas gravações de modo a permitir a construção e o teste de modelos mais robustos, obtidos exclusivamente com base em vogais extraídas de palavras.

5.4.4 Alternativa de Modelização no Domínio Cepstral

Os resultados obtidos nos testes do algoritmo utilizando os modelos alternativos no domínio cepstral (cujos valores consolidados estão disponíveis em B.5, B.6, B.7 e B.8) apontam para taxas de acerto ligeiramente inferiores aos registados nos modelos espectrais. Porém, é importante notar que a aplicação da curva de ponderação (como a utilizada nos modelos no domínio espectral) aos modelos cepstral conduziu a resultados virtualmente idênticos aos que se encontram nas tabelas B.1, B.2, B.3 e B.4, o que significa que se poderá utilizar qualquer um destes modelos com resultados semelhantes.

Caso se verifique, em estágios mais avançados do projeto, a necessidade de extrair outras características ou introduzir andares adicionais de processamento, poderá revelar-se pertinente simplificar os modelos por forma a reduzir o custo computacional das diversas operações a que serão sujeitos. A variante cepstral dos modelos, seguindo uma abordagem semelhante aos MFCCs, permite uma substancial redução do número de coeficientes sendo também adequada a técnicas de aprendizagem automática, como de resto são comummente aplicadas aos MFCCs, na eventualidade de estas virem a ser incorporadas no algoritmo. Esta abordagem permite também, conforme aponta a literatura, lidar com os aspetos transitórios do sinal com um reduzido custo computacional adicional (ver 2.3.4).

5.5 Síntese do Capítulo

Neste capítulo propôs-se uma abordagem de identificação de vogais sussurradas, baseada em estudos de correlação estatística das características espectrais dos segmentos a analisar com um conjunto de vogais de referência, orientado ao orador. Após a avaliação preliminar da abordagem proposta, em que se fez uso das modelizações LPC de envolventes espectrais de vogais pré-processadas, desenvolveram-se novos modelos compactos no domínio espectral e no domínio cepstral, adequados à plataforma DyNaVoiceR e aos sinais de fala sussurrada, tendo-se testado o algoritmo protótipo com as 4 vogais mais frequentes na base de dados DyNaVoiceR.

Os testes realizados permitiram concluir sobre a razoabilidade da abordagem e a adequação do algoritmo para operação em tempo-real, sendo que o tempo de processamento de cada *frame* é substancialmente inferior à cadência a que as *frames* serão geradas numa utilização em tempo-real. Registaram-se, porém, desempenhos significativamente diferentes para o conjunto das vogais testadas, sublinhando-se a necessidade de testar o algoritmo num contexto mais alargado de vogais. Outras importantes observações e conclusões incluem:

- A comparação dos resultados obtidos com cada uma das duas bibliotecas de referência construídas aponta no sentido da necessidade de se obterem modelos mais robustos e de se testarem modelos médios obtidos exclusivamente de vogais extraídas de palavra;
- Concluiu-se pela maior viabilidade das escalas não lineares e pela adequação das escalas percetuais Bark e ERB em alternativa à escala *mel*;
- Verificou-se que o aumento da resolução espectral não se traduz numa melhoria relevante da capacidade de identificar vogais sussurradas;

Finalmente, constatou-se que as abordagens espectral e cepstral conduziram a resultados similares entre si, desde que aplicada a curva de ponderação em ambos os cenários. Sendo que os modelos cepstral implicam a aplicação de uma transformada adicional de cálculo eficiente e expedito (a DCT), a compactação e maior versatilidade permitida pelos modelos cepstral poderá, no entanto, vir a revelar-se determinante num estágio mais avançado do projeto.

Os testes realizados neste capítulo tiveram um carácter objetivo, em que se tomou por base de referência a anotação fonética manual prévia, disponível na base de dados. No próximo capítulo, documentam-se as experiências preliminares de vozeamento automatizado de secções selecionadas do sinal, utilizando o algoritmo de identificação de vogais desenvolvido, com vista à realização de testes subjetivos informais, de carácter percetual.

Capítulo 6

Experiências Preliminares com Vozeamento

A última etapa deste trabalho consistiu numa primeira experiência de vozeamento em que, aplicando o algoritmo de identificação de vogais sussurradas apresentado no capítulo anterior, se realizou a implantação automatizada de segmentos de fala natural vozeada em regiões selecionadas do sinal. Pretendeu-se avaliar, do ponto de vista percetual, o impacto do vozeamento das vogais selecionadas, analisando-se com particular interesse o resultado da substituição da vogal por uma diferente da pretendida, como simulação de uma identificação incorreta.

6.1 Implantação de Vogal Natural Vozeada

A experiência de vozeamento, que consistiu na implantação de sinal de fala vozeada natural em segmentos de vogais originalmente sussurradas, baseou-se no algoritmo de identificação de vogais sussurradas desenvolvido na dissertação e recorreu, mais uma vez, à anotação fonética disponível na base de dados para realizar a localização automática dos segmentos candidatos a substituição. Nestas operações de vozeamento, a vogal a substituir correspondeu à primeira vogal na segunda repetição das palavras dissílabas disponíveis na base de dados DyNaVoiceR. Utilizou-se, para a identificação das vogais, os modelos espectrais com 18 bandas na escala Bark. Consideraram-se, para vogais de referência, os modelos médios compósitos de vogais sustentadas e de palavras. Uma vez identificada a vogal, o algoritmo procede de forma automatizada à substituição dessa vogal por um segmento de comprimento adequado extraído da vogal sustentada vozeada correspondente, do mesmo orador, introduzindo um ajuste de ganho por forma a garantir a implantação graciosa do segmento no sinal original. Este teste de vozeamento teve carácter abrangente tendo sido nele utilizados os 20 oradores previamente selecionados e cada um dos 28 dissílabos disponíveis isoladamente na base de dados (ver mapa da Fig. 6.1).

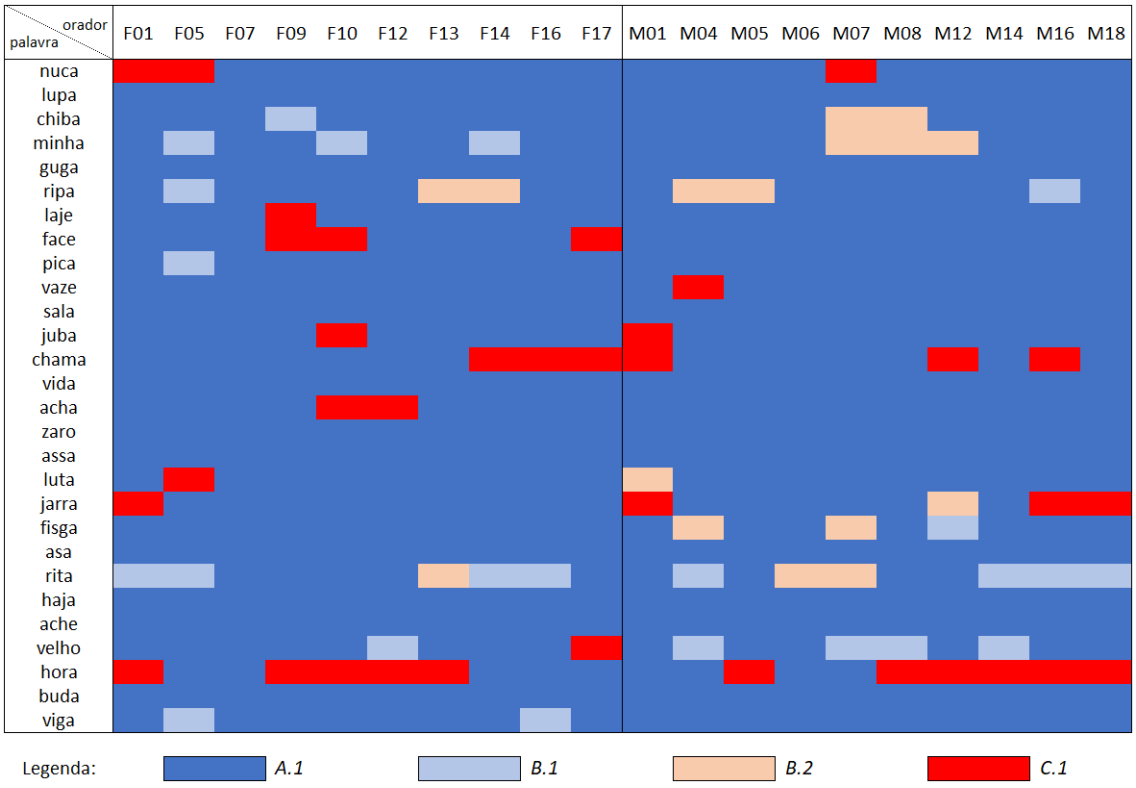


Figura 6.1: Mapa do vozeamento automatizado da primeira vogal de todas as palavras dissílabas isoladas da base de dados, para os 20 oradores previamente selecionados, recorrendo ao algoritmo de identificação de vogais sussurradas desenvolvido. (Legenda: consultar 6.2)

6.2 Análise Subjectiva dos Resultados

A avaliação dos novos sinais obtidos após a implantação das vogais vozeadas, realizada por elementos do projeto, contemplou as seguintes classificações subjetivas, de carácter percetual:

- vogal corretamente identificada e substituída, informação linguística preservada (A.1);
- a identificação incorreta da vogal e substituição por uma vogal diferente da pretendida é pouco percetível, preservando-se a informação linguística (B.1);
- a identificação incorreta da vogal e substituição por uma vogal diferente da pretendida resultou na adulteração percetível e evidente da palavra, sendo contudo possível inferir a informação linguística original (B.2);
- a identificação incorreta da vogal e substituição por uma vogal diferente da pretendida resultou na destruição da informação linguística (C.1).

6.2.1 Síntese dos Resultados

A experiência de vozeamento incluiu as 28 palavras dissílabas sussurradas disponíveis na forma isolada na base de dados e os 20 oradores previamente selecionados (10 masculinos e 10 femininos) num total de 560 palavras. A vogal selecionada para substituição foi sempre a vogal respeitante à primeira sílaba, incluindo:

- 12 casos com o /á/ (de 'água');
- 8 casos com o /i/ (de 'ilha');
- 6 casos com o /u/ (de 'uva');
- 1 caso com o /é/ (de 'ela');
- 1 caso com o /ó/ (de 'óculos').

O algoritmo identificou corretamente a vogal sussurrada num total de 486 casos do conjunto das 560 palavras, correspondendo a uma taxa global de acerto de aproximadamente 86.8 %. Dos 74 casos de identificação incorreta, o vozeamento automático resultou na destruição da informação linguística em 35 casos, aproximadamente metade, gerando palavras inexistentes no vocabulário Português Europeu (*e.g.*, 'hora' transformou-se em 'ura') ou existentes mas claramente distintas da palavra original sussurrada (*e.g.*, 'rita' transformou-se em 'reta'), o que corresponde a uma taxa de 6.25 %. Nos restantes 39 casos de identificação incorreta, verificou-se ser possível perceber ou deduzir com relativa facilidade a informação linguística pretendida, sendo no entanto de admitir que uma parte destes seja suscetível de gerar ambiguidade. Será necessário, em estágios posteriores do projeto, confirmar estas conclusões com testes formais, realizados com um conjunto alargado de participantes e seguindo as recomendações adequadas a este tipo de testes.

6.2.2 Análise Pontual de Vogais

A experimentação com vozeamento incluiu duas vogais, não abrangidas pelos testes do capítulo anterior, por se encontrarem disponíveis unicamente numa palavra por orador: o /ó/ em 'hora' e o /é/ em 'velho'.

A vogal /é/ em 'velho' registou uma taxa de acerto de 70 % (*i.e.*, 14 das 20 situações possíveis), sensivelmente em linha com as vogais já testadas no capítulo anterior. Entre as identificações incorretas, apenas uma conduziu a um implante com carácter destrutivo do ponto de vista da informação linguística (identificada como o /u/ de 'uva', tendo resultado em 'vulho'). As restantes identificações incorretas conduziram à substituição da vogal por 'â' de /amarelo/ (uma das vogais mais próximas no espaço das vogais). O resultado obtido nestes casos poderá considerar-se aceitável do ponto de vista percetual. Quer isto dizer que se subentende a palavra 'velho', confundindo-se com uma pronunciação estranha desta palavra.

A vogal /ó/ em 'hora' gerou 11 identificações incorretas em 20 situações possíveis, o que corresponde a uma taxa de acerto de 45 %. Todas as identificações incorretas conduziram a implantes de vogal vozeada com carácter destrutivo, do ponto de vista da informação linguística. Destas identificações incorretas, 8 (cerca de 73 %) foram identificadas como sendo o /á/ de 'água' (uma das vogais mais próximas do /ó/ no espaço das vogais, ver Fig. 2.3), e 3 (cerca de 27 %) como o /u/ de 'uva'. A vogal de referência correta registou a segunda melhor classificação em 8 destes casos e a terceira nos 3 restantes, encontrando-se sempre entre as 3 primeiras segundo a classificação do algoritmo. Para despiste das possíveis causas desta reduzida taxa de acerto, procedeu-se à audição das vogais sussurradas originais destes oradores, tanto das versões sussurradas como das versões extraídas da palavra 'hora', tendo-se identificado artefactos num conjunto limitado de casos (estalidos da língua no orador masculino M12 e ligeiro vozeamento na oradora feminina F09). Verificou-se também, nos restantes oradores, que algumas vogais sussurradas apresentavam uma qualidade perceptualmente sofrível. Esta análise não é, porém, conclusiva, devido à limitação do material disponível, sugerindo-se a construção de novos modelos de referência com recurso a um complemento da base de dados e posterior reavaliação.

6.2.3 Impacto do Vozeamento

A audição dos sinais obtidos, após a implantação das vogais vozeadas sobre a palavra sussurrada original com recurso ao algoritmo de identificação de vogais sussurradas desenvolvido, permitiu concluir que a substituição da primeira vogal nos dissílabos permite preservar a informação linguística na generalidade das palavras testadas. A audição destes sinais, não obstante a simplicidade do processo utilizado, sugere também que o vozeamento da primeira vogal na palavra originalmente sussurrada promove, do ponto de vista percetual, a aproximação da palavra à correspondente na fala normalmente vozeada, criando a sugestão de toda a palavra ter sido gerada no modo vozeado.

6.3 Síntese do Capítulo

As experiências preliminares de vozeamento conduzidas na dissertação tiveram como objetivo avaliar o impacto do vozeamento de segmentos selecionados do sinal de fala sussurrada, com recurso ao algoritmo de identificação de vogais desenvolvido. O algoritmo identificou corretamente a vogal sussurrada em 86.8 % das palavras testadas. Em 52.7 % das identificações corretas, é possível perceber ou inferir a palavra pretendida. O procedimento de vozeamento automático resultou na perda da informação linguística em 6.25 % do total das palavras testadas. As vogais /é/ e /ó/, disponíveis numa única palavra por orador, registaram taxas de acerto de 70 e 45 %, respetivamente, tendo sido confundidas com vogais próximas no espaço das vogais nos casos de identificação incorreta.

Capítulo 7

Conclusão

A revisão bibliográfica realizada no estágio de preparação da dissertação permitiu identificar a dificuldade subjacente à melhoria ou conversão de sinais de fala sussurrada, em virtude das características inerentes a este tipo de sinal, revelando também as deficiências das soluções que têm sido desenvolvidas ou exploradas no campo da investigação, nomeadamente pelas suas limitações práticas ou pelos resultados obtidos, pouco convincentes. O projeto DyNaVoiceR propõe-se a desenvolver uma solução prática, de carácter não intrusivo, que sendo capaz de operar em tempo-real, converte o sinal de fala sussurrada em sinal de fala perceptualmente natural, implantando vozeamento sintético em regiões selecionadas do sinal de fala, acentuando a informação linguística e transmitindo elementos da assinatura sonora do orador, assegurando também, neste processo, melhor projeção vocal.

Nesta dissertação:

- analisaram-se as características das vogais orais em cenário de fala sussurrada e de fala normalmente vozeada, comparando-se do ponto de vista percetual as envolventes espectrais destas vogais no sussurro e das componentes periódica e de ruído na fala normalmente vozeada;
- propuseram-se modelos compactos das características espectrais das vogais sussurradas e desenvolveu-se um algoritmo protótipo de identificação de vogais sussurradas, adequado aos requisitos e objetivos do projeto DyNaVoiceR, nomeadamente a operacionalidade em tempo-real;
- realizaram-se experiências preliminares de vozeamento de vogais originalmente sussurrada nos dissílabos disponíveis na base de dados DyNaVoiceR.

Apresenta-se de seguida uma síntese dos resultados obtidos e conclusões alcançadas, na sequência destes trabalhos.

7.1 Síntese das Conclusões

A análise de envolventes espectrais de sinais de fala sussurrada e dos correspondentes sinais de fala vozeada, apresentada no capítulo 4, permitiu concluir sobre a naturalidade, do ponto de vista percetual e da preservação da informação linguística, dos sinais de fala sussurrada sintética com base em modelos espectrais médios, obtidos do mesmo orador. A audição de sussurro sintético, comparando a envolvente espectral residual obtida da versão vozeada com o fonema correspondente na versão sussurrada, indicou a preservação da informação linguística no resíduo (a componente de ruído do sinal na fala normalmente vozeada), o qual partilha características espectrais com as envolventes da componente periódica (formantes sensivelmente nas mesmas regiões) e com a envolvente sussurrada correspondente (o mesmo tipo de distribuição de energia no espectro). Estas conclusões terão eventual interesse, em fases posteriores do projeto, no âmbito dos processos de ressíntese de sinal (*e.g.*, interpolação de envolventes espectrais por forma a atribuir naturalidade ou melhores índices de inteligibilidade).

No capítulo 5, propôs-se uma metodologia de identificação de vogais sussurradas baseada em correlações estatísticas, utilizando o coeficiente de correlação de Pearson, entre os fonemas a analisar/classificar e um banco de vogais de referência do mesmo orador. Esta metodologia foi primeiramente avaliada utilizando envolventes LPC pré-processadas, tendo-se posteriormente desenvolvido um algoritmo protótipo parametrizável de identificação de vogais sussurradas, idealizado para operação em tempo-real, fazendo-se representar os fonemas por modelos compactos das características espectrais, no domínio espectral e no domínio cepstral, constituídos por um número reduzido de coeficientes.

A condução de testes de desempenho sobre a capacidade de identificação de vogais sussurradas, contemplando diferentes alternativas e opções do algoritmo, permitiu concluir que:

- o algoritmo é viável para operação em tempo-real;
- a taxa de acerto, segundo o critério simples do coeficiente de correlação mais elevada, varia substancialmente entre as vogais testadas;
- a taxa de sucesso relativo do algoritmo, definida como a frequência com que a vogal correta se encontra entre as três mais bem classificadas pelo algoritmo, é bastante elevada para todas as vogais testadas, indicando um potencial superior do algoritmo que poderá ser utilizado com recurso a técnicas de natureza estatística (*e.g.*, Hidden Markov Models / Gaussian Mixture Models) ou a possibilidade de mitigação dos erros de identificação através da interpolação entre fonemas próximos entre si no espaço das vogais;
- os modelos de referência das vogais deverão ser obtidos com base em vogais extraídas em contexto de palavra;

- o algoritmo beneficia da utilização de uma escala não linear para as frequências (banco de filtros, na extração de características) que melhor se adeque às características fisiológicas do ouvido humano como as escalas Bark e ERB;
- a redução do espectro útil, limitando o número de bandas por forma a abranger apenas as primeiras três formantes e enfatizando as duas primeiras, não compromete significativamente a taxa de acerto do algoritmo permitindo a redução dos custos computacionais;
- a duplicação da resolução espectral, não obstante o maior detalhe nos modelos de vogal resultantes, não beneficia significativamente as taxas de acerto do algoritmo;
- a utilização de modelos no domínio cepstral permite reduzir o número de coeficientes, sem custos significativos nas taxas de acerto do algoritmo, requerendo apenas a aplicação de uma transformada adicional de cálculo eficiente e expedito (a DCT), com o potencial de serem utilizados também para a análise das características transitórias do sinal.

Os resultados obtidos sugerem também que o algoritmo desenvolvido apresenta taxas de acerto superiores na vogal correspondente à primeira sílaba das palavras dissílabas, o que se ficará a dever à maior estabilidade das características espectrais destes segmentos do sinal e a maior dependência de efeitos de prosódia na segunda e última sílaba.

Tendo sido viável construir modelos de referência apenas para as 9 vogais orais e a realização de testes exaustivos de identificação de vogais sobre 4 das vogais, será necessário ampliar os modelos das vogais de referência por forma a incluir as vogais nasais e validar o algoritmo de identificação das vogais para o novo espaço das vogais, de preferência recorrendo a uma metodologia de construção de modelos de referência mais robusta e consistente do que a que esteve disponível nesta dissertação.

As experiências de vozeamento com segmentos de fala natural obtidos do mesmo orador, apresentadas no capítulo 6 e implementadas de forma automatizada com recurso ao algoritmo protótipo de identificação de vogais sussurradas, permitiu concluir sobre a viabilidade da implantação de segmentos de vogal vozeada, do ponto do vista percetual, preservando-se a informação linguística e a inteligibilidade na generalidade dos cenários testados. A implantação de segmentos de fala vozeada nos segmentos correspondentes à vogal da primeira sílaba de palavras dissílabas conduziu a resultados bastante satisfatórios, com uma ligeira melhoria da projeção do sinal, aproximando-o a um sinal de fala normalmente vozeada, incluindo cenários em que a vogal foi substituída por uma outra vogal próxima no espaço das vogais, em virtude de ter sido incorretamente identificada pelo algoritmo. Estes resultados sugerem também a viabilidade de se utilizarem técnicas de interpolação entre fonemas próximos no espaço das vogais, preservando a informação linguística e melhorando a inteligibilidade.

7.2 Propostas de Trabalho Futuro

Por forma a dar continuidade ao trabalho desenvolvido nesta dissertação, com base nas conclusões obtidas e nas limitações aqui identificadas, propõem-se as seguintes tarefas futuras:

- **melhorar a metodologia de construção de modelos de referência**, por forma a obter modelos mais robustos, sugerindo-se extrair modelos exclusivamente baseados em vogais obtidas de contexto de palavra, com duração razoável, melhorando a representatividade dos modelos e evitando os efeitos devidos à coarticulação com os fonemas contíguos;
- **complementar a base de dados de oradores**, obtendo novos modelos para um conjunto de 14 vogais do Português Europeu padrão, 9 vogais orais e 5 vogais nasais, e validar o algoritmo de identificação de vogais para este novo conjunto de fonemas;
- **desenvolver métodos de interpolação e construção de envolventes espectrais de vogais**, por orador, com vista à síntese e implantação de segmentos de sinal de fala vozeada sintética nos sinais de fala sussurrada originais;
- **conduzir testes formais de carácter percetual**, com base na implantação automatizada de vogais vozeadas sintéticas em sinais de fala sussurrada, com um conjunto alargado de participantes, obedecendo às recomendações previstas para este tipo de testes;
- **avaliar a integração de métodos de natureza estatística no algoritmo** (*e.g.*, HMM-GMM), com vista à melhoria do desempenho na identificação dos fonemas.

Também no âmbito dos trabalhos futuros, será necessário integrar o trabalho desenvolvido com outros módulos cujo desenvolvimento se encontra em curso no projeto DyNaVoiceR, nomeadamente com as tarefas de segmentação fonética automática.

Anexo A

Reprodução das Tarefas de Gravação (DyNaVoiceR)

Procedimento: Fonemas sustentados – dizer cada som 3 vezes em fala normal e 3 vezes em fala sussurrada, 2 a 3 segundos;

S de sapato

X de chapéu

Z de casa

J de janela

I de ilha

E de peso

E de ela

A de água

A de amarelo

O de óculos

O de ovo

U de uva

E de sede

1

Figura A.1: Folha 1 das tarefas de gravação, com instruções

Procedimento: Leitura de palavras – dizer cada palavra 3 vezes em fala normal e 3 vezes em fala sussurrada, pela ordem apresentada;

Col. 1	Col. 2	Col. 3	Col. 4	Col. 5	Col. 6
Nuca	Ripa	Sala	Zaro	Asa	Velho
Lupa	Laje	Juba	Assa	Rita	Hora
Chiba	Face	Chama	Luta	Haja	Buda
Minha	Pica	Vida	Jarra	Ache	Viga
Guga	Vaze	Acha	Fisga		

Procedimento: Leitura de frases –ler cada frase 3 vezes em fala normal e 3 vezes em fala sussurrada;

2

Figura A.2: Folha 2 das tarefas de gravação, com instruções

A Marta e o avô vivem naquele casarão rosa velho.

Sofia saiu cedo da sala.

A asa do avião andava avariada.

Agora é hora de acabar.

A minha mãe mandou-me embora.

O Tiago comeu quatro peras.

Procedimento: Leitura do texto 1 vez em fala normal e 1 vez em fala sussurrada

3

Figura A.3: Folha 3 das tarefas de gravação, com instruções

Tabela A.1: Correspondência entre a identificação na base dados, código IPA e notação utilizada na dissertação para as 9 vogais orais do Português Europeu padrão consideradas na base de dados DyNaVoiceR e incluídas nas tarefas sustentadas.

	Correspondências		
Label (BD)	Vogal	IPA	exemplo
1	/i/	i	'i̲lha'
2	/ê/	ɛ	'pe̲so'
3	/é/	e	'e̲la'
4	/á/	a	'á̲gua'
5	/â/	ɐ	'a̲marelo'
6	/ó/	ɔ	'ó̲culos'
7	/ô/	o	'o̲vo'
8	/u/	u	'u̲va'
9	/e/	ɨ	'sede̲'

Anexo B

Testes de Desempenho do Algoritmo de Identificação de Vogais Sussurradas (Cap. 5)

Tabela B.1: Taxas de acerto e de sucesso para a vogal **/i/ de 'ilha'**, modelo médio sustentadas+palavras, **coeficientes espectrais** (18 a 22 bandas úteis) para as escalas Linear, Mel, Bark e ERB. Resultados consolidados para 20 oradores, 10 femininos e 10 masculinos, num total de 475 ocorrências da vogal.

/i/ - Acertos (%)				
Bandas	Lin.	Mel	Bark	ERB
22	40.8	55.8	64.8	66.5
21	44.4	58.3	65.5	67.6
20	45.9	60.2	66.9	68.0
19	48.8	62.3	68.4	68.8
18	53.3	64.4	70.7	70.5

/i/ - Sucessos (%)				
Bandas	Lin.	Mel	Bark	ERB
22	86.7	98.1	99.8	99.8
21	85.9	98.5	99.8	99.8
20	86.3	99.2	100	100
19	88.6	99.4	100	100
18	94.1	99.6	100	100

Tabela B.2: Taxas de acerto e de sucesso para a vogal **/á/ de 'água'**, modelo médio sustentadas+palavras, **coeficientes espectrais** (18 a 22 bandas úteis) para as escalas Linear, Mel, Bark e ERB. Resultados consolidados para 20 oradores, 10 femininos e 10 masculinos, num total de 647 ocorrências da vogal.

/á/ - Acertos (%)				
Bandas	Lin.	Mel	Bark	ERB
22	73.9	90.9	90.7	91.7
21	79.0	91.3	90.9	91.5
20	80.4	90.7	90.9	91.8
19	80.4	91.0	90.9	91.8
18	87.2	91.5	91.3	92.0

/á/ - Sucessos (%)				
Bandas	Lin.	Mel	Bark	ERB
22	85.2	96.6	96.9	97.1
21	87.8	96.9	96.9	97.1
20	88.4	96.8	97.2	97.2
19	90.9	97.2	97.4	97.2
18	93.2	97.4	97.7	97.7

Tabela B.3: Taxas de acerto e de sucesso para a vogal **/â/ de 'amarelo'**, modelo médio sustentadas+palavras, **coeficientes espectrais** (18 a 22 bandas úteis) para as escalas Linear, Mel, Bark e ERB. Resultados consolidados para 20 oradores, 10 femininos e 10 masculinos, num total de 1365 ocorrências da vogal.

/â/ - Acertos (%)				
Bandas	Lin.	Mel	Bark	ERB
22	30.1	47.5	47.5	47.3
21	30.9	46.7	47.6	47.2
20	32.7	46.2	47.5	46.7
19	35.2	46.9	47.1	46.7
18	37.1	45.7	47.0	46.4

/â/ - Sucessos (%)				
Bandas	Lin.	Mel	Bark	ERB
22	59.5	71.9	75.7	79.1
21	58.6	71.4	75.2	78.8
20	59.3	70.9	75.2	77.8
19	61.2	71.3	75.2	78.1
18	63.2	72.2	74.8	77.9

Tabela B.4: Taxas de acerto e de sucesso para a vogal **/u/ de 'uva'**, modelo médio sustentadas+palavras, **coeficientes espectrais** (18 a 22 bandas úteis) para as escalas Linear, Mel, Bark e ERB. Resultados consolidados para 20 oradores, 10 femininos e 10 masculinos, num total de 459 ocorrências da vogal.

/u/ - Acertos (%)				
Bandas	Lin.	Mel	Bark	ERB
22	75.4	85.8	85.0	84.5
21	78.9	85.6	85.0	84.5
20	82.8	85.6	84.7	84.7
19	84.1	85.4	84.3	84.7
18	83.9	85.8	84.3	84.1

/u/ - Sucessos (%)				
Bandas	Lin.	Mel	Bark	ERB
22	95.4	95.9	95.0	95.6
21	97.6	95.9	95.0	95.6
20	97.6	95.6	95.2	95.4
19	97.8	96.3	95.2	95.0
18	96.7	96.5	95.2	94.8

Tabela B.5: Taxas de acerto e de sucesso para a vogal **/i/ de 'ilha'**, modelo médio sustentadas+palavras, **coeficientes cepstrais** (18 a 22 bandas úteis) para as escalas Linear, Mel, Bark e ERB. Resultados consolidados para 20 oradores, 10 femininos e 10 masculinos, num total de 475 ocorrências da vogal.

/i/ - Acertos (%)				
Bandas	Lin.	Mel	Bark	ERB
22	41.3	48.8	54.5	54.1
21	45.1	52.8	54.1	55.4
20	46.5	53.7	55.6	56.8
19	47.6	56.0	59.8	60.0
18	52.8	60.0	63.6	64.4

/i/ - Sucessos (%)				
Bandas	Lin.	Mel	Bark	ERB
22	85.5	94.9	97.1	96.6
21	84.8	94.7	97.5	97.1
20	84.8	94.5	97.5	97.3
19	87.8	96.8	97.9	98.3
18	93.3	98.5	99.4	99.6

Tabela B.6: Taxas de acerto e de sucesso para a vogal **/á/ de 'água'**, modelo médio sustentadas+palavras, **coeficientes cepstrais** (18 a 22 bandas úteis) para as escalas Linear, Mel, Bark e ERB. Resultados consolidados para 20 oradores, 10 femininos e 10 masculinos, num total de 647 ocorrências da vogal.

/á/ - Acertos (%)				
Bandas	Lin.	Mel	Bark	ERB
22	60.4	85.9	87.8	87.5
21	67.1	86.9	87.6	87.5
20	76.7	86.4	87.6	87.3
19	76.4	86.9	88.4	88.1
18	84.9	89.8	90.1	89.6

/á/ - Sucessos (%)				
Bandas	Lin.	Mel	Bark	ERB
22	85.2	95.4	96.0	94.1
21	87.8	95.2	96.1	94.3
20	88.4	94.4	96.0	94.9
19	90.9	95.4	96.0	95.1
18	93.2	96.0	96.3	95.8

Tabela B.7: Taxas de acerto e de sucesso para a vogal **/â/ de 'amarelo'**, modelo médio sustentadas+palavras, **coeficientes cepstrais** (18 a 22 bandas úteis) para as escalas Linear, Mel, Bark e ERB. Resultados consolidados para 20 oradores, 10 femininos e 10 masculinos, num total de 1365 ocorrências da vogal.

/â/ - Acertos (%)				
Bandas	Lin.	Mel	Bark	ERB
22	26.0	44.6	47.6	44.8
21	28.6	45.0	47.2	45.9
20	29.7	45.5	46.7	45.8
19	32.2	44.8	45.9	44.8
18	35.8	44.8	46.1	44.8

/â/ - Sucessos (%)				
Bandas	Lin.	Mel	Bark	ERB
22	53.0	69.2	72.2	74.1
21	52.6	69.9	72.3	74.1
20	54.7	71.0	74.0	74.2
19	57.2	71.6	73.6	75.6
18	60.8	71.9	73.8	74.9

Tabela B.8: Taxas de acerto e de sucesso para a vogal **/u/ de 'uva'**, modelo médio sustentadas+palavras, **coeficientes cepstrais** (18 a 22 bandas úteis) para as escalas Linear, Mel, Bark e ERB. Resultados consolidados para 20 oradores, 10 femininos e 10 masculinos, num total de 459 ocorrências da vogal.

/u/ - Acertos (%)				
Bandas	Lin.	Mel	Bark	ERB
22	59.3	87.6	85.8	85.4
21	62.5	86.7	85.6	85.4
20	66.4	87.6	85.6	84.5
19	65.6	86.7	85.4	84.7
18	74.9	86.5	85.4	84.5

/u/ - Sucessos (%)				
Bandas	Lin.	Mel	Bark	ERB
22	95.4	98.3	97.2	98.3
21	97.6	97.6	97.8	97.6
20	97.6	97.6	96.9	97.8
19	97.8	96.5	97.2	96.9
18	96.7	97.2	97.4	96.7

Tabela B.9: Taxas de acerto e de sucesso para a vogal **/i/ de 'ilha'**, modelo médio sustentadas+palavras, coeficientes espectrais de **alta resolução** (36 a 44 bandas úteis) para as escalas Linear, Mel, Bark e ERB. Resultados consolidados para 20 oradores, 10 femininos e 10 masculinos, num total de 475 ocorrências da vogal.

/i/ - Acertos (%)				
Bandas	Lin.	Mel	Bark	ERB
44	40.6	55.8	63.4	65.9
42	44.0	57.1	64.8	66.9
40	45.5	58.7	65.9	67.4
38	48.2	61.5	66.9	67.8
36	52.4	63.6	68.8	69.5

/i/ - Sucessos (%)				
Bandas	Lin.	Mel	Bark	ERB
44	87.8	98.1	99.8	99.8
42	87.4	98.9	99.8	100
40	87.2	99.2	100	100
38	88.6	99.4	100	100
36	93.5	99.6	100	100

Tabela B.10: Taxas de acerto e de sucesso para a vogal **/á/ de 'água'**, modelo médio sustentadas+palavras, coeficientes espectrais de **alta resolução** (36 a 44 bandas úteis) para as escalas Linear, Mel, Bark e ERB. Resultados consolidados para 20 oradores, 10 femininos e 10 masculinos, num total de 647 ocorrências da vogal.

/á/ - Acertos (%)				
Bandas	Lin.	Mel	Bark	ERB
44	81.6	92.4	93.8	94.1
42	82.2	92.7	93.7	94.0
40	83.3	92.6	94.0	94.6
38	82.8	93.0	94.0	94.4
36	89.2	92.9	94.4	94.6

/á/ - Sucessos (%)				
Bandas	Lin.	Mel	Bark	ERB
44	89.6	96.9	97.7	97.5
42	90.1	97.2	97.8	97.7
40	90.6	97.1	97.8	97.7
38	92.6	97.4	97.8	97.7
36	94.9	97.7	98.1	97.8

Tabela B.11: Taxas de acerto e de sucesso para a vogal **/â/ de 'amarelo'**, modelo médio sustentadas+palavras, coeficientes espectrais de **alta resolução** (36 a 44 bandas úteis) para as escalas Linear, Mel, Bark e ERB. Resultados consolidados para 20 oradores, 10 femininos e 10 masculinos, num total de 1365 ocorrências da vogal.

/â/ - Acertos (%)				
Bandas	Lin.	Mel	Bark	ERB
44	33.0	46.9	48.2	47.6
42	34.2	46.7	48.1	48.4
40	36.0	46.7	48.0	47.7
38	38.5	46.1	47.7	47.4
36	39.6	46.8	47.5	46.7

/â/ - Sucessos (%)				
Bandas	Lin.	Mel	Bark	ERB
44	60.4	70.8	75.3	79.1
42	62.1	70.9	75.1	78.4
40	61.5	71.4	75.1	78.3
38	63.3	70.9	74.7	77.5
36	64.6	71.6	74.9	77.0

Tabela B.12: Taxas de acerto e de sucesso para a vogal **/u/ de 'uva'**, modelo médio sustentadas+palavras, coeficientes espectrais de **alta resolução** (36 a 44 bandas úteis) para as escalas Linear, Mel, Bark e ERB. Resultados consolidados para 20 oradores, 10 femininos e 10 masculinos, num total de 459 ocorrências da vogal.

/u/ - Acertos (%)				
Bandas	Lin.	Mel	Bark	ERB
44	85.2	86.9	85.4	85.0
42	84.1	86.3	85.4	84.3
40	85.8	86.5	85.0	84.7
38	87.4	86.3	85.0	84.7
36	85.6	85.6	84.7	84.5

/u/ - Sucessos (%)				
Bandas	Lin.	Mel	Bark	ERB
44	98.0	96.5	95.9	95.9
42	98.0	96.7	95.9	95.4
40	99.0	95.6	96.1	95.6
38	98.5	96.1	96.1	95.2
36	97.2	96.3	95.6	95.4

Tabela B.13: Taxas de acerto e de sucesso para a vogal **/i/ de 'ilha'**, **modelo médio de sustentadas**, coeficientes espectrais (18 a 22 bandas úteis) para as escalas Linear, Mel, Bark e ERB. Resultados consolidados para 20 oradores, 10 femininos e 10 masculinos, num total de 475 ocorrências da vogal.

/i/ - Acertos (%)				
Bandas	Lin.	Mel	Bark	ERB
22	27.2	37.1	39.2	42.3
21	32.0	37.1	39.8	43.8
20	31.4	37.5	40.2	44.4
19	32.4	38.7	41.1	45.3
18	34.7	38.1	40.8	46.9

/i/ - Sucessos (%)				
Bandas	Lin.	Mel	Bark	ERB
22	56.6	76.0	86.3	88.6
21	58.5	79.8	87.8	89.7
20	61.9	82.9	88.8	89.9
19	64.2	85.3	89.9	90.8
18	72.2	88.8	90.9	92.2

Tabela B.14: Taxas de acerto e de sucesso para a vogal **/á/ de 'água'**, **modelo médio de sustentadas**, coeficientes espectrais (18 a 22 bandas úteis) para as escalas Linear, Mel, Bark e ERB. Resultados consolidados para 20 oradores, 10 femininos e 10 masculinos, num total de 647 ocorrências da vogal.

/á/ - Acertos (%)				
Bandas	Lin.	Mel	Bark	ERB
22	54.1	79.6	81.6	79.6
21	65.1	79.9	81.8	79.8
20	69.4	80.2	81.9	79.9
19	66.9	81.1	81.3	80.4
18	74.5	82.4	82.1	80.5

/á/ - Sucessos (%)				
Bandas	Lin.	Mel	Bark	ERB
22	86.2	95.5	96.1	95.7
21	92.7	96.0	96.0	95.7
20	95.2	96.0	95.8	95.7
19	95.7	95.8	95.8	96.0
18	96.6	96.0	96.0	96.0

Tabela B.15: Taxas de acerto e de sucesso para a vogal **/â/ de 'amarelo'**, **modelo médio de sustentadas**, coeficientes espectrais (18 a 22 bandas úteis) para as escalas Linear, Mel, Bark e ERB. Resultados consolidados para 20 oradores, 10 femininos e 10 masculinos, num total de 1365 ocorrências da vogal.

/â/ - Acertos (%)				
Bandas	Lin.	Mel	Bark	ERB
22	40.1	53.6	53.2	52.0
21	39.7	51.2	53.1	52.0
20	45.4	52.2	52.9	52.2
19	47.0	53.0	52.5	52.3
18	47.4	51.8	52.2	52.1

/â/ - Sucessos (%)				
Bandas	Lin.	Mel	Bark	ERB
22	82.3	88.0	84.9	83.7
21	84.2	86.0	84.0	82.6
20	86.6	84.8	83.7	81.5
19	85.2	84.6	83.2	81.3
18	83.5	84.0	82.1	80.5

Tabela B.16: Taxas de acerto e de sucesso para a vogal **/u/ de 'uva'**, **modelo médio de sustentadas**, coeficientes espectrais (18 a 22 bandas úteis) para as escalas Linear, Mel, Bark e ERB. Resultados consolidados para 20 oradores, 10 femininos e 10 masculinos, num total de 459 ocorrências da vogal.

/u/ - Acertos (%)				
Bandas	Lin.	Mel	Bark	ERB
22	41.2	42.5	38.8	42.0
21	42.5	44.0	39.4	42.5
20	51.6	41.6	39.4	43.1
19	49.2	40.7	38.6	42.9
18	48.1	40.1	39.0	43.6

/u/ - Sucessos (%)				
Bandas	Lin.	Mel	Bark	ERB
22	78.0	88.5	89.1	90.2
21	81.7	89.3	89.5	90.6
20	87.4	90.0	90.0	90.4
19	89.1	90.2	89.8	90.4
18	89.1	90.6	90.2	90.2

Referências

[1] Izabel Christine Seara, Vanessa Gonzaga Nunes, e Cristiane Lazzarotto-Volcão. *Fonética e fonologia do português.* UFSC, Florianápolis, 2011.

[2] Tom Bäckström e Okko Räsänen. *Introduction to Speech Processing.* Aalto University, recurso em linha, 2019.

[3] Maria Raquel Delgado-Martins. Análise acústica das vogais tónicas em português. *Boletim de Filologia*, 1973.

[4] Ian Vince McLoughlin. *Speech and Audio Processing: A MATLAB®-based Approach.* Cambridge University Press, 2016.

[5] Gunnar Fant. *Speech Acoustics and Phonetics: Selected Writings.* Text, Speech and Language Technology. Springer Netherlands, 2007.

[6] E. Zwicker e H. Fastl. *Psychoacoustics: Facts and Models.* Springer series in information sciences. Springer Berlin Heidelberg, 2007.

[7] Aníbal Ferreira. Implantation of voicing on whispered speech using frequency-domain parametric modelling of source and filter information. Em *2016 International Symposium on Signal, Image, Video and Communications (ISIVC)*, páginas 159–66, Novembro 2016.

[8] Vladimir Senatorov, Shirish Satpute, Katherine Perry, David Kaylie, e John Cole. Aphonia induced by simultaneous bilateral ischemic infarctions of the putamen nuclei: A case report and review of the literature. *Journal of medical case reports*, 7:83, Março 2013.

[9] Eliana Fabron, Viviane Marino, Talyssa Nóbile, Luciana Sebastião, e Suely Motonaga. Medical treatment and speech therapy for spasmodic dysphonia: a literature review. *Revista CEFAC*, 15:713–25, Junho 2013.

[10] Charlotte Jacobs. *Carcinomas of the Head and Neck: Evaluation and Management.* Springer, Boston, USA, 1990.

[11] Vincent Callanan, Paul Gurr, David Baldwin, Morwenna White-Thompson, Jane Beckinsale, e Jane Bennett. Provox^{TM} valve use for post-laryngectomy voice rehabilitation. *The Journal of Laryngology and Otology*, 109(11):1068–71, 1995.

[12] James H. Brandenburg. Vocal Rehabilitation After Laryngectomy. *Archives of Otolaryngology*, 106(11):688–91, Novembro 1980.

[13] Jun Wang, Ashok Samal, Jordan R. Green, e Frank Rudzicz. Sentence recognition from articulatory movements for silent speech interfaces. *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, páginas 4985–8, 2012.

[14] Maria J. Soutelo. *MasterVoicing - A whisper to voiced speech assistant*. Dissertação de Mestrado, Faculdade de Engenharia da Universidade do Porto, Porto, Portugal, 2017.

[15] Philip Lieberman. The evolution of human speech: Its anatomical and neural bases. *Current Anthropology*, 48:39–66, Fevereiro 2007.

[16] Lesley Mathieson. *Greene and Mathieson's The Voice and Its Disorders*. Wiley, 2001.

[17] Jason Fagone. The quest to save stephen hawking's voice. *San Francisco Chronicle*, recurso em linha, 2018.

[18] Xuedong Huang, Alex Acero, e Hsiao-Wuen Hon. *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*. Prentice Hall PTR, 2001.

[19] Maria Helena Mira Mateus, Isabel Falé, e Maria João Freitas. *Fonética e fonologia do português*. Universidade Aberta, Lisboa, 2016.

[20] Kenneth N. Stevens. *Acoustic Phonetics*. Current studies in linguistics series. Cambridge, 1998.

[21] D. B. Fry, S. R. Anderson, J. Bresnan, B. Comrie, W. Dressler, e C. J. Ewen. *The Physics of Speech*. Cambridge Textbooks in Linguistics. Cambridge University Press, 1979.

[22] L. R. Rabiner e B. H. Juang. *Fundamentals of Speech Recognition*. Prentice-Hall Signal Processing Series: Advanced monographs. PTR Prentice Hall, 1993.

[23] J. Sundberg. Formant structure and articulation of spoken and sung vowels. *Folia Phoniat*, página 22:28–48, 1970.

[24] J. Sundberg. *The Science of the Singing Voice*. Northern Illinois University Press, 1987.

[25] Paavo Alku. Glottal inverse filtering analysis of human voice production — a review of estimation and parameterization methods of the glottal excitation and their applications. *Sadhana*, 36(5):623–50, Outubro 2011.

[26] Xing Fan e John Hansen. Acoustic analysis and feature transformation from neutral to whisper for speaker identification within whispered speech audio streams. *Speech Communication*, 55:119–34, Janeiro 2013.

[27] Martin F. Schwartz. Power spectral density measurements of oral and whispered speech. *Journal of Speech and Hearing Research*, 13(2):445–6, 1970.

[28] Taisuke Ito, Kazuya Takeda, e Fumitada Itakura. Analysis and recognition of whispered speech. *Speech Communication*, 45:139–52, Fevereiro 2005.

[29] Ingegerd Eklund e Hartmut Traunmüller. Comparative study of male and female whispered and phonated versions of the long vowels of swedish. *Phonetica*, 54:1–21, Janeiro 1997.

[30] Siobodan T. Jovicic. Formant feature differences between whispered and voiced sustained vowels. *Acta Acustica united with Acustica*, 84(4):739–43, 1998.

[31] Ken J. Kallail e Floyd W. Emanuel. Formant-frequency differences between isolated whispered and phonated vowel samples produced by adult female subjects. *Journal of speech and hearing research*, 27:245–51, Julho 1984.

[32] Hamid Sharifzadeh, Ian Mcloughlin, e Martin Russell. A comprehensive vowel space for whispered speech. *Journal of voice : official journal of the Voice Foundation*, 26:49–56, Maio 2011.

[33] Vivien Tartter. Identifiability of vowels and speakers from whispered syllables. *Perception & psychophysics*, 49:365–72, Maio 1991.

[34] Ken J. Kallail e Floyd W. Emanuel. The identifiability of isolated whispered and phonated vowel samples. *Journal of Phonetics*, 13(1):11–7, 1985.

[35] Robert W. Morris. *Enhancement and Recognition of Whispered Speech.* Tese de doutoramento, Georgia Institute of Technology, Atlanta, GA, USA, 2003.

[36] M. Higashikawa, K. Nakai, A. Sakakura, e H. Takahashi. Perceived pitch of whispered vowels-relationship with formant frequencies: A preliminary study. *Journal of Voice*, 10(2):155–8, 1996.

[37] I. B. Thomas. Perceived pitch of whispered vowels. *The Journal of the Acoustical Society of America*, 46(2B):468–70, 1969.

[38] Vivien Tartter e David Braun. Hearing smiles and frowns in normal and whisper registers. *The Journal of the Acoustical Society of America*, 96:2101–7, Novembro 1994.

[39] Keith Johnson. *Acoustic and Auditory Phonetics.* Chichester : Wiley-Blackwell, 2012.

[40] DeLiang Wang e Guy J. Brown. *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications.* Wiley-IEEE Press, 2006.

[41] Brian C. J. Moore. *An Introduction to the Psychology of Hearing.* Brill, Leiden, The Netherlands, 2013.

[42] Ben Gold, Nelson Morgan, e Dan Ellis. *Speech and Audio Signal Processing: Processing and Perception of Speech and Music.* Wiley-Interscience, New York, USA, 2ª edição, 2011.

[43] J. V. Tobias. *Foundations of modern auditory theory.* Número 1 em Foundations of Modern Auditory Theory. Academic Press, 1970.

[44] D. O'Shaughnessy. *Speech communication: human and machine.* Addison-Wesley series in electrical engineering. Addison-Wesley Pub. Co., 1987.

[45] Andreas Spanias, Ted Painter, e Venkatraman Atti. *Audio Signal Processing and Coding.* John Wiley & Sons, Inc., Dezembro 2005.

[46] B. R. Glasberg e B. C. Moore. Derivation of auditory filter shapes from notched-noise data. *Hearing Research*, 47(1-2):103–38, 1990.

[47] Malcolm J. Crocker. *Handbook of Acoustics*. A Wiley-Interscience Publication. Wiley, 1998.

[48] Juan G. Roederer. *The physics and psychophysics of music: an introduction*. Springer-Verlag, 1995.

[49] Gunnar Fant. *Acoustic theory of speech production*. The Hague, The Netherlands, Mouton, 1960.

[50] A. M. Kondoz. *Digital speech: coding for low bit rate communication systems*. Wiley, 2004.

[51] Jeremy S. Bradbury. *Linear Predictive Coding*. recurso em linha, 2000.

[52] Wai C. Chu. *Speech Coding Algorithms: Foundation and Evolution of Standardized Coders*. John Wiley & Sons, Inc., New York, USA, 1ª edição, 2003.

[53] Md Sahidullah e Goutam Saha. Design, analysis and experimental evaluation of block based transformation in mfcc computation for speaker recognition. *Speech Communication*, 54:543–65, Maio 2012.

[54] M. R. Schroeder, H. Quast, e H. W. Strube. *Computer Speech: Recognition, Compression, Synthesis*. Springer Series in Information Sciences. Springer Berlin Heidelberg, 2004.

[55] Robert W. Morris e Mark A. Clements. Reconstruction of speech from whispers. *Medical engineering and physics*, 24:515–20, Setembro 2002.

[56] Ian Mcloughlin, Hamid Sharifzadeh, Su Lim Tan, Jingjie Li, e Yan Song. Reconstruction of phonated speech from whispers using formant-derived plausible pitch modulation. *ACM Transactions on Accessible Computing*, 6:1–21, Maio 2015.

[57] Ian Vince McLoughlin, Jingjie Li, e Yan Song. Reconstruction of continuous voiced speech from whispers. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, páginas 1022–26, Janeiro 2013.

[58] Hamid R. Sharifzadeh, Ian Vince McLoughlin, e Farzaneh Ahmadi. Reconstruction of normal sounding speech for laryngectomy patients through a modified celp codec. *IEEE Transactions on Biomedical Engineering*, 57:2448–58, 2010.

[59] T. Toda, M. Nakagiri, e K. Shikano. Statistical voice conversion techniques for body-conducted unvoiced speech enhancement. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(9):2505–17, Novembro 2012.

[60] Patricia R. Oliveira. *Artificial voicing of whispered speech*. Dissertação de Mestrado, Faculdade de Engenharia da Universidade do Porto, Porto, Portugal, 2015.

[61] Aníbal Ferreira e Deepen Sinha. Audio communication coder. *120th Convention of the Audio Engineering Society*, páginas 1–21, 2006.

[62] Y. Dun e G. Liu. A fine-resolution frequency estimator in the Odd-DFT domain. *IEEE Signal Processing Letters*, 22(12):2489–93, Dezembro 2015.

[63] Aníbal Ferreira. Accurate estimation in the ODFT domain of the frequency phase and magnitude, of stationary sinusoids. *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, páginas 47–50, Fevereiro 2001.

[64] Aníbal Ferreira e Deepen Sinha. Accurate and robust frequency estimation in the odft domain. Em *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, 2005*, páginas 203–6, Novembro 2005.

[65] R. Rowlands. The odd discrete fourier transform. *ICASSP '76. IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1:130–133, Abril 1976.

[66] Alan V. Oppenheim e Ronald W. Schafer. *Discrete-Time Signal Processing*. Prentice Hall Press, Upper Saddle River, NJ, USA, 3ª edição, 2009.

[67] Monson H. Hayes. *Statistical Digital Signal Processing and Modeling*. John Wiley & Sons, Inc., New York, NY, USA, 1ª edição, 1996.

[68] João P. Silva, Marco A. Oliveira, e Aníbal J. Ferreira. Fundamental frequency contour and micro-variation manipulations using a fully parametric harmonic speech model. *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Maio 2020 (submetido).

[69] Yoni Swerdlin, John Smith, e Joe Wolfe. The effect of whisper and creak vocal mechanisms on vocal tract resonances. *The Journal of the Acoustical Society of America*, 127:2590–8, Abril 2010.

[70] Joe Rodgers e Alan Nicewander. Thirteen ways to look at the correlation coefficient. *American Statistician - AMER STATIST*, 42:59–66, Fevereiro 1988.

[71] H. Hermansky. Perceptual linear predictive (plp) analysis of speech. *The Journal of the Acoustical Society of America*, 87 4:1738–52, 1990.

[72] C. Kim e R. M. Stern. Power-normalized cepstral coefficients (pncc) for robust speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(7):1315–1329, Julho 2016.

[73] Marisa Lousada, Luis Jesus, e Andreia Hall. Temporal acoustic correlates of the voicing contrast in european portuguese stops. *Journal of the International Phonetic Association*, 40:261 – 275, Dezembro 2010.