FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Vozeamento artificial de fala não vozeada

Paulo Jorge Proença de Azevedo

Mestrado Integrado em Engenharia Electrotécnica e de Computadores

Orientador: Aníbal Ferreira (Prof. Dr. Eng.)

Co-orientador: Ricardo Sousa (Dr. Eng.)

04 de Outubro de 2012

U.PORTO

FEUP FACULDADE DE ENGENHARIA UNIVERSIDADE DO PORTO

MIEEC - MESTRADO INTEGRADO EM ENGENHARIA ELETROTÉCNICA E DE COMPUTADORES 2011/2012

A Dissertação intitulada

"Vozeamento Artificial de Fala não Vozeada"

foi aprovada em provas realizadas em 04-10-2012

o júri

Presidente Professor Doutor Diamantino Rui da Silva Freitas
Professor Associado do Departamento de Engenharia Eletrotécnica e de Computadores da Faculdade de Engenharia da Universidade do Porto

Professor Doutor José Manuel Neto Vieira
Professor Auxiliar do Departamento de Electrónica e Telecomunicações do Instituto de Engenharia Electrónica e Telemática de Aveiro

Professor Doutor Aníbal João de Sousa Ferreira
Professor Associado do Departamento de Engenharia Eletrotécnica e de Computadores da Faculdade de Engenharia da Universidade do Porto

Doutor Ricardo Teixeira de Sousa
Bolseiro POS - DOC da FCUP

O autor declara que a presente dissertação (ou relatório de projeto) é da sua exclusiva autoria e foi escrita sem qualquer apoio externo não explicitamente autorizado. Os resultados, ideias, parágrafos, ou outros extratos tomados de ou inspirados em trabalhos de outros autores, e demais referências bibliográficas usadas, são corretamente citados.

Autor - Paulo Jorge Proença de Azevedo

Resumo

A voz tem um papel vital na comunicação entre seres humanos. A capacidade de vozear a fala é uma habilidade concedida pelas cordas vocais, e com a qual, a esmagadora maioria da população pode contar. No entanto, um pequeno conjunto de pessoas, por ter sido alvo de operações cirúrgicas na zona laríngea, apenas retém a capacidade de sussurrar.

Atualmente, existem vários métodos de terapia de fala destinados a estes pacientes. Contudo, a insuficiência da qualidade da voz resultante e a natureza intrusiva de alguns métodos são dois dos principais catalisadores da procura de uma solução mais apropriada.

Com esta dissertação, pretende-se avaliar a possibilidade de reconstrução digital de fala vozeada a partir de fala não-vozeada; ou seja, criar voz a partir de sussurros. Pretende-se ainda perceber até que ponto a extração de certas características de um sinal de voz original podem auxiliar o vozeamento artificial do sinal sussurrado.

Para este efeito foram gravadas quatro amostras sonoras em ambiente anecoico: uma amostra de voz e uma amostra de sussurro, ditadas por um orador saudável de cada género.

Criaram-se, com o auxílio de técnicas de processamento de sinal, duas versões de fala vozeada artificialmente para cada um dos oradores. Uma versão construída sem o auxílio da respetiva referência vozeada, e uma versão com esse auxílio.

A avaliação dos resultados foi subjetiva, contando-se com a presença de 35 voluntários. Cada voluntário pontuou os quatro ficheiros finais nas categorias de: degradação, inteligibilidade, naturalidade e identidade do orador.

De seguida, foi realizada uma análise descritiva e inferencial das pontuações. Esta análise permitiu tirar as necessárias conclusões sobre a viabilidade do vozeamento artificial de fala sussurrada.

Por fim, investigaram-se métodos de classificação automática de pequenos segmentos dos sinais sussurrados. Uma componente vital na eventual integração deste algoritmo num aparelho utilizável pelos pacientes laringectomizados.

Conclui-se com uma ponderação dos potenciais futuros desenvolvimentos na área aqui em estudo.

Abstract

Voice plays a vital role in the communication among human beings. The capability of voicing speech is an ability granted by the vocal cords, and on which, the majority of the population can rely. However, a small number of people, having been subject to laryngeal surgery, are only able to whisper.

Nowadays, there are several speech therapy methods aimed to help these patients. Yet, the lackluster resulting voice quality and the intrusive nature of some of these methods are two of the main catalyzers of the search for a better suited solution.

In this dissertation the possibility of digital voicing unvoiced speech, or in other words, reconstructing voice from whispers, is measured. Furthermore, an assessment of how valuable the extraction of some of the characteristics of an original voiced signal, in the process of voicing unvoiced speech, is also performed.

To this end, four sounds samples were recorded in an anechoic environment: a sample of voiced and a sample of unvoiced speech, dictated by a narrator of each gender.

With the help of signal processing technics, for each of the unvoiced samples two artificially voiced samples were obtained. One version created with the help of the original voiced sample, and one without referred help.

The four signals obtained were evaluated subjectively by a group of 35 volunteers. Each volunteer assigned four different scores to each of the signals; one score for each of the following fields: degradation, intelligibility, inartificiality and speaker identity.

Afterwards, a descriptive and inferential analysis of the scores was performed. This analysis allowed the drawing of conclusions concerning the viability of artificial voicing of whispered speech.

Finally, methods for the automatic classification of small segments of the whispered signals were investigated. This is a vital component in the potential integration of this algorithm on a device usable by laryngectomees.

In the final parts of this dissertation, suggestions are made as to what future work could be done following this study.

Agradecimentos

Deixo aqui expressa a minha gratidão ao meu orientador, Professor Doutor Engenheiro Aníbal Ferreira, pelo incentivo e oportunidade de trabalhar neste tema, e também pelo seu apoio nas diversas questões que foram surgindo.

Quero agradecer igualmente ao meu co-orientador, Doutor Engenheiro Ricardo Sousa, pelos seus ensinamentos e pela sua incessante paciência e prontidão no auxílio do desenvolvimento desta dissertação, sem os quais teria sido impossível alcançar os resultados aqui presentes.

Presto também um agradecimento à minha namorada, Mariana Almeida, pela sua fantástica pré-disposição em ajudar em tudo o que foi necessário; com destaque para a gravação de amostras de fala e no recrutamento de voluntários para o teste subjetivo dos resultados.

Uma outra palavra de agradecimento vai para a minha tia, Cláudia Proença, pela correção da versão escrita da dissertação aqui exposta.

Agradeço também a todos os voluntários que participaram na avaliação dos resultados, cuja contribuição não foi menos que crucial.

Mais ainda, vejo reconhecida a valiosa ajuda do meu amigo Ricardo Dias, uma vez que me guiou nos primeiros passos de ambientação ao programa de computador usado na escrita da dissertação. Aproveito também para felicitar o resto das pessoas que trabalharam sob o mesmo orientador, pelo ambiente de entreajuda criado.

Uma menção ao Professor Doutor Engenheiro Diamantino pela sua disponibilidade expressada em partilhar o seu conhecimento de processamento de sinal comigo.

Por último, quero manifestar um profundo agradecimento aos meus pais e amigos pelo seu precioso apoio; não só na elaboração desta dissertação, mas duma forma geral no sucesso académico que me ajudaram a alcançar.

Paulo Azevedo

"If you can't explain it simply, you don't understand it well enough."

Albert Einstein

Conteúdo

Lista de Figuras

Lista de Tabelas

Abreviaturas e Símbolos

ACP	Análise de Componentes Principais
ACR	*Absolute Category Rating*
AFI	Alfabeto Fonético Internacional
AUC	*Area Under the Curve*
CELP	*Code Excited Linear Prediction*
CP	Componentes Principais
DCR	*Degradation Category Rating*
F_0	Frequência fundamental
F_n	Frequência formante de índice n, para n diferente de 0.
F_s	Frequência de amostragem(=*Sampling*)
LF	Liljencrants-Fant
Matlab	MATrix LABoratory
MFCC	*Mel-Frequency Cepstral Coefficients*
MVS	Máquina de Vetores de Suporte
PCM	*Pulse Code Modulation*
PE	Português Europeu
RBF	*Radial Basis Function*
RNA	Rede Neuronal Artificial
ROC	*Receiver Operating Characteristic*
RoR	*Rate of Rise*
sp	*Synthesised Power*
SPSS	Statistic Package for the Social Sciences
VAD	*Voice Activity Detector*
VCV	Vogal-Consoante-Vogal
vp	*Voiced Power*
WAD	*Whisper Activity Detector*
WAV	*WAVeform audio format*
WEKA	Waikato Environment for Knowledge Analysis
wp	*Whispered Power*
WPC	*Whispered Phoneme Classifier*

Capítulo 1

Introdução

Desde o dia em que nascemos, fazemos por tentar comunicar com os outros seres humanos! Daí, são apenas cerca de 2 anos até aprendermos as primeiras palavras, e pouco mais até articularmos as primeiras frases. Esta capacidade de aprender e utilizar uma linguagem na sua forma verbal remonta a, pelo menos, 5000 A.C. [18]. Nos dias de hoje, a esmagadora maioria da população adulta, felizmente tem como garantido, há anos, o correto funcionamento dos seus órgãos produtores de fala.

1.1 Âmbito

Esta dissertação destina-se à minoria dos habitantes do mundo que não possuem a capacidade de falar vozeadamente, mas que retêm a capacidade de falar não vozeadamente. Por outras palavras, pessoas que apesar não conseguirem falar normalmente, conseguem sussurrar. Em termos fisiológicos, refiro-me a indivíduos que mantêm a capacidade de exalar ar dos pulmões e a capacidade de modular o fluxo de ar através do trato vocal, mas que não possuem a capacidade de utilizar as pregas vocais. Este é o caso, por exemplo, dos pacientes que tenham paralisia das pregas vocais, ou dos pacientes que tenham sido sujeitos a uma laringectomia parcial.

Em termos de áreas de conhecimento, esta dissertação está contida, essencialmente, na área do processamento digital de sinal. No entanto, são também abrangidas as áreas da medicina e da linguística.

1.2 Motivação

A motivação desta dissertação surge da necessidade de colmatar a dificuldade de comunicação dos pacientes incapazes de vozear a fala. Apesar do sussurro ser uma forma verbal de expressão disponível a ser utilizada pelo ser humano na sua maioria, não é nada conveniente para estes pacientes a impossibilidade crónica de vozear a sua fala. Tendo isto em mente, existem atualmente diversas tentativas à ultrapassagem do problema. No entanto, nenhuma é, ao mesmo tempo, simples de pôr em prática e detentora de resultados satisfatórios. Assim, desponta a necessidade

de trabalhar no sentido de criar uma alternativa mais próxima do ideal, a que todos os pacientes, nomeadamente os laringectomizados parciais, possam recorrer.

Paralelamente, a eliminação deste problema pela via de processamento digital de sinal, só por si detém interesse, uma vez que a solução obtida consistiria numa ferramenta de processamento poderosa e inovadora.

1.3 Objetivo

O objetivo desta dissertação não é, no entanto, tão ambicioso quanto produzir a solução quasi-ideal mencionada. Em concreto, o que se pretende obter como resultado deste estudo, é a prova de conceito da possibilidade ou impossibilidade de transformar digitalmente um sinal de sussurro num sinal de voz. Mais ainda, deseja-se perceber até que ponto, o uso de certas de características de um sinal vozeado original, contribui para um sinal artificialmente vozeado com maior qualidade do que um sinal artificialmente vozeado que não faça uso de uma referência original.

Assim, neste estudo, recolhem-se pares de amostras de fala sussurrada e vozeada que ostentem o mesmo conjunto ordenado de sons, todos eles de oradores perfeitamente saudáveis.

Uma vez obtido um conjunto de sinais artificialmente vozeados com diferentes graus de dependência da referência vozeada, serão comparados com esta referência e tecer-se-ão as devidas conclusões.

1.4 Organização da Dissertação

Para além da introdução, esta dissertação contém mais 7 capítulos. No capítulo 2, aborda-se o sistema de produção de fala humano, bem como certos conceitos de codificação digital de sinais de fala aplicados neste estudo. No capítulo 3, faz-se uma alusão ao estado da arte em terapia de fala para pacientes laringectomizados. No capítulo 4, enumeram-se as ferramentas utilizadas neste trabalho e de que modo prestaram utilidade. No capítulo 5, explica-se o procedimento de recolha dos sinais mencionados. No capítulo 6, descreve-se a metodologia empregue no vozeamento artificial do sinal não vozeado e faz-se uma análise da qualidade do sinal obtido. No capítulo 7, faz-se um trabalho complementar na área da deteção das zonas do sinal não vozeado que devem ser vozeadas. No capítulo 8, reflete-se sobre a viabilidade do processo de vozeamento artificial e prestam-se sugestões de trabalho futuro.

Capítulo 2

Revisão Bibliográfica

Neste capítulo, abordam-se conceitos essenciais à compreensão dos restantes conteúdos desta dissertação. Primeiro, começa-se com a descrição do mecanismo humano de produção de fala segundo o modelo matemático simplificado "fonte-filtro". Seguidamente, enumeram-se os diferentes tipos de sons existentes na língua portuguesa europeia e a forma como estes se organizam. Por fim, apresentam-se as principais diferenças entre os tipos de fala sussurrada e vozeada aqui em observação.

2.1 Modelo fonte-filtro

A maioria dos sistemas físicos capazes de produzir som são constituídos por três elementos: um elemento excitador, um elemento vibrador e um elemento ressoador. O primeiro é o responsável por fornecer energia ao sistema que, no caso da fala, é o ar expelido pelos pulmões. Um elemento vibrador que, mediante a excitação do primeiro elemento, produz um padrão acústico vibratório cuja frequência determina a tonalidade do som produzido; no caso da fala, trata-se das pregas vocais. E um elemento ressoador, que tem por função converter as oscilações do elemento vibrador em vibrações sonoras do ar circundante, impondo timbre ao som; equivalente, no caso da fala, ao trato vocal [19]. É importante notar que o volume de ar à saída dos lábios não é igual ao volume de ar à frente do ouvinte ou dispositivo gravador, embora seja comum incorporar-se este efeito de radiação labial no final do trato vocal. Na designação "modelo fonte-filtro", a fonte corresponde ao fluxo de ar após ação das pregas vocais, e o filtro trata-se do trato vocal.

A Figura 2.1 revela o comportamento em frequência do módulo de cada um dos blocos do modelo fonte-filtro (neste caso o filtro labial encontra-se isolado do trato vocal). Como se pode observar, o bloco correspondente à fonte apresenta várias riscas, a primeira correspondendo à frequência da oscilação das pregas vocais, denominada de frequência fundamental (ou F_0), e as restantes aos seus harmónicos. A magnitude dos harmónicos segue um decaimento de cerca de -12dB/oit [20]. O bloco correspondente ao filtro do trato vocal apresenta uma amplitude plana, exceto pelas elevações nas suas frequências de ressonância conhecidas como frequências formantes

(ou F_n, onde n=1,2,3,...). Por fim, o bloco correspondente ao filtro de radiação labial comporta-se como um diferenciador [20], de onde se deduz que o seu declive é de +6dB/oit.

Uma das maiores limitações deste modelo está em assumir a completa separação dos dois sistemas "fonte"e "filtro", algo que em termos fisiológicos não acontece. No entanto, este é atualmente um dos modelos que melhor representa a produção de fala e, por conseguinte, é no pressuposto do seu correto funcionamento que os procedimentos desta dissertação se baseiam.

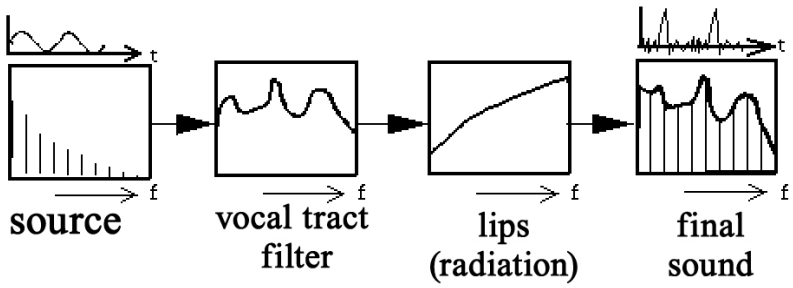

Figura 2.1: Diagrama de blocos da produção de fala de acordo com o modelo fonte-filtro [1].

2.2 Fonética do português europeu

O Português Europeu (PE) é constituído por um conjunto de sons distintos que resultam do comportamento dos articuladores do trato vocal e do estado das pregas vocais. A Figura 2.2 ilustra o modelo fisiológico do agrupamento trato vocal e pregas (=cordas) vocais. De acordo com a categorização contemplada no Alfabeto Fonético Internacional (AFI), a língua portuguesa é composta por 35 sons diferentes. Estes 35 sons são agrupados em grupos consoante as semelhanças ou dissemelhanças no seu processo de produção.

A mais conhecida divisão é entre vogais e consoantes; de facto, existem 14 vogais, 19 consoantes e ainda 2 semivogais (semelhantes a vogais mas com menos energia e duração). Esta distinção é feita com base na existência ou não da constrição do trato vocal num ponto: no caso das vogais, não existe qualquer constrição. Por oposição, no caso das consoantes, existe. Mais ainda, nas vogais existe sempre fonação, isto é, as pregas vocais vibram periodicamente na produção de qualquer vogal. A estes sons chama-mos vozeados. No caso das consoantes, pode ou não existir fonação, ou seja, podem ser ou não vozeadas. Uma vogal pode ser subsequentemente categorizada como nasal ou oral, caso exista ou não passagem de ar através da cavidade nasal, respetivamente. Tanto as vogais orais como nasais são posteriormente classificadas de acordo com a posição do dorso e da raiz da língua (alta, média, baixa) e com a posição dos lábios (anterior, central, posterior ou velar). O anexo A contém tabelas com todas as vogais da língua portuguesa e suas divisões.

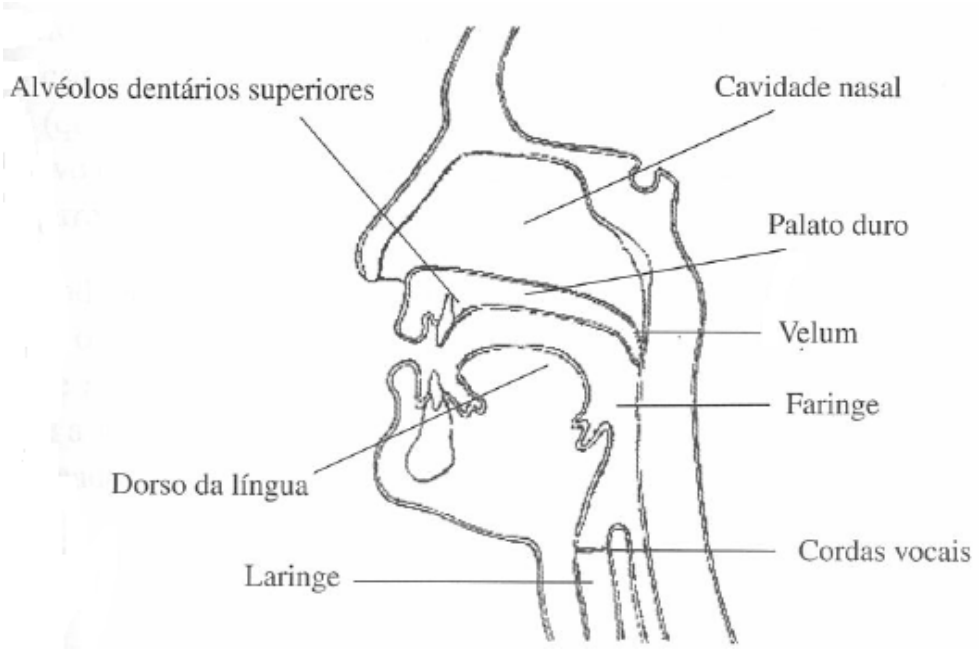

Figura 2.2: Trato vocal, seus articuladores e pregas (=cordas) vocais [2].

Em contrapartida, uma consoante é classificada quanto ao seu ponto de articulação, o seu modo de articulação, a posição do palato mole e o estado das pregas vocais. A Figura 2.3 apresenta o conjunto das consoantes com as suas subdivisões. Na coluna da esquerda dividem-se as consoantes de acordo com o ponto de articulação: bilabial, labiodental, dental, alveolar, palatal, velar, uvular; e estado das pregas vocais: vozeada ou não-vozeada; e na linha do topo estão separados os 4 modos de articulação: oclusiva (ou plosiva), fricativa, lateral, vibrante; e as duas posições do palato mole: a fechar o acesso à cavidade nasal no caso de uma consoante oral, e a permitir o acesso no caso de uma consoante nasal. Quanto a esta última divisão, quando não mencionadas, as consoantes são orais.

No interesse desta dissertação, existe uma divisão mais importante do que todas as outras: a existência ou não de fonação (ou vozeamento). Como foi mencionado, todas as vogais são vozeadas e, do total de 19 consoantes, apenas 3 oclusivas e 3 fricativas são não vozeadas. O porquê da importância desta divisão é explicado na secção 2.5, após se discutirem as diferenças entre os dois tipos de fala em questão.

Ponto e voz. \ Modo		Oclusiva		Fricativa	Lateral	Vibrante
		Oral	Nasal			
Bilabial	Vozeada	b	m			
	Não-vozeada	p				
Labiodental	Vozeada			v		
	Não-vozeada			f		
Dental	Vozeada	d		z		
	Não-vozeada	t		s		
Alveolar	Vozeada		n		l	ɾ
	Não-vozeada					
Palatal	Vozeada		ɲ	ʒ	ʎ	
	Não-vozeada			ʃ		
Velar	Vozeada	g				
	Não-vozeada	k				
Uvular	Vozeada					ʀ
	Não-vozeada					

Figura 2.3: Classificação articulatória tradicional das consoantes do português europeu padrão [2].

2.3 Características da voz e do sussurro

Fisiologicamente falando, a principal diferença na produção destes dois tipos encontra-se ao nível da glote. Segundo [21], as pregas vocais formam uma abertura estreita na fala sussurrada e uma vibração quasi-periódica para fala vozeada. No entanto, segundo [3], existem outros estudos que descrevem a abertura das pregas vocais como sendo maior na fala sussurrada.

Considere-se o modelo fonte-filtro. Se já não existe vibração das pregas vocais, então a fonte já não possui uma frequência fundamental nem harmónicos; ao invés, a fonte pode ser vista agora como ruído de turbulência causado pela passagem de ar pelas pregas vocais sem repetida obstrução. Uma outra consequência da abertura da glote é um acoplamento acústico da parte superior do trato vocal com a cavidade subglotal.

Importa então perceber as diferenças a nível espectral da fala sussurrada. Já foi visto que o sinal de excitação é agora ruído em vez de um sinal periódico, o que causa o desaparecimento da frequência fundamental e dos harmónicos. Este facto é visível na Figura 2.4. Assim, a estrutura harmónica visível no espectro da vogal vozeada não se encontra no espectro da vogal sussurrada.

A segunda grande diferença é ao nível da amplitude das frequências formantes. A envolvente espectral da vogal sussurrada apresenta uma forma mais planar, onde as magnitudes das frequências formantes encontram-se menos proeminentes do que na envolvente da contraparte vozeada.

Por fim, os próprios valores das frequências formantes também se alteram. Nas vogais, a primeira formante na fala sussurrada apresenta sempre uma frequência mais alta do que na fala vozeada: cerca de 39% mais alta nos homens e 32% nas mulheres [3], ou cerca de 22% de acordo com [21]. A frequência da segunda formante apresenta uma deslocação média muito inferior, na

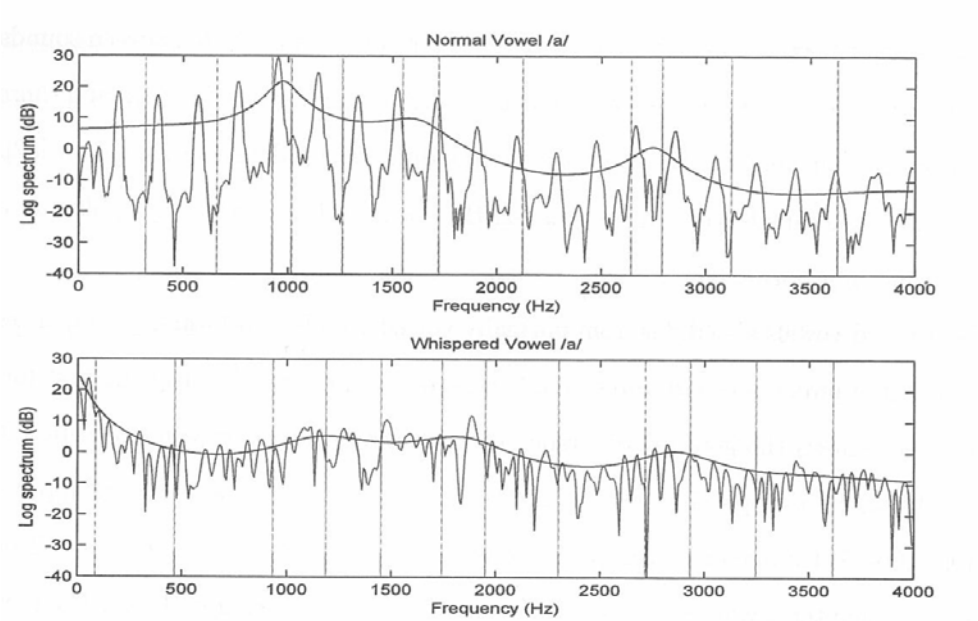

Figura 2.4: Espectros da vogal [a] na sua versão vozeada e sussurrada [3].

maioria dos casos positiva, embora em [21] 2 das vogais apresentassem deslocações negativas em 2 das vogais no caso feminino. A terceira frequência apresenta uma deslocação ainda menor e sempre positiva, exceto numa única vogal. Infelizmente, estes dados baseiam-se no estudo das vogais da língua inglesa, não tendo sido encontrada informação relativa às vogais da língua portuguesa.

Relativamente à deslocação das frequências formantes nas consoantes vozeadas, confirma-se uma subida das frequências das primeiras formantes semelhante à das vogais.

Relativamente às consoantes não vozeadas, o seu espectro é semelhante na fala vozeada e sussurrada [22].

No geral, a fala não vozeada possui menos 20dB de potência espectral do que a sua contraparte vozeada [3].

2.4 Análise de fala de curta duração

Para as secções que se seguem, faz sentido perceber em primeiro lugar, como obter os espectros dos diferentes sons presentes numa amostra de fala.

Na análise da fala, não se obteria informação útil se se obtivesse a transformada de Fourier do sinal todo, uma vez que daí resultaria uma média dos espectros de todos os sons proferidos nessa amostra. É, então preciso segmentar a amostra em partes de curta duração (ordem das dezenas de milissegundos) denominadas de tramas e, posteriormente, realizar a transformada de Fourier de cada uma das tramas. Esta segmentação pode ser vista como a multiplicação do sinal total, $x[n]$

por uma janela, $w[n]$, que é nula para valores $|n| > N/2$. Assim, a transformada de Fourier numa determinada trama é dada por [19]:

$$X_m(e^{jw}) = \sum_{n=-\infty}^{\infty} x[n]w[m-n]e^{jwn}. \quad (2.1)$$

Geralmente, estas tramas não são partes disjuntas do sinal, mas sim parcialmente sobrepostas.

Existe uma ferramenta muito útil de representação no domínio das frequências de todas as tramas do sinal: os espectrogramas. Os espectrogramas são representações de duas dimensões onde o eixo das abcissas corresponde ao tempo e o eixo das ordenadas às frequências. O valor do espectrograma em dB num determinado ponto é dado por:

$$S_m(e^{jw})_{dB} = 10 \times log_{10}(|X_m(e^{jw})|^2). \quad (2.2)$$

Para se obter a realização visual converte-se o valor de $S_m(e^{jw})_{dB}$ para uma escala de cor (geralmente cinzentos). A Figura 2.5 ilustra um exemplo de um espectrograma da sequência vozeada "A Sofia".

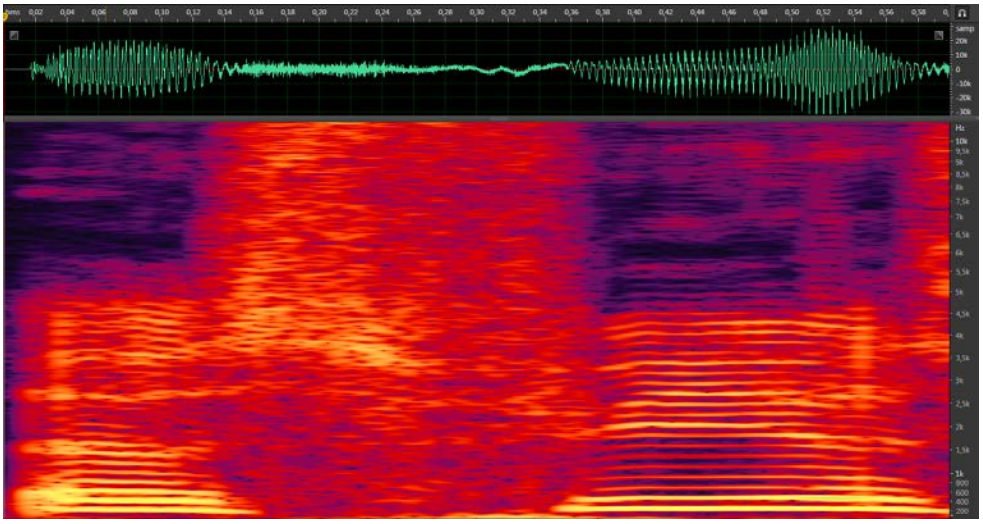

Figura 2.5: Sequência de fala vozeada "A Sofia"e seu espectrograma obtido no Adobe Audition [4].

2.5 Conclusão

Dada a divisão entre os tipos de fala vozeada e sussurrada e entre os sons vozeados e não vozeados, percebe-se que só faz sentido vozear artificialmente o sinal de fala sussurrada nas partes que correspondem a sons vozeados. Como vimos no capítulo anterior, os sons não vozeados

mantêm-se semelhantes nos dois tipos de fala exceto na sua potência. Já os sons vozeados, necessitam de sofrer uma série de alterações para soarem a vozeados, a saber: introdução de uma frequência fundamental e harmónicos; deslocação das frequências formantes; estreitamento da largura de banda das frequências formantes e aumento da sua magnitude; aumento da potência total; e, possivelmente, o ajuste do decaimento da envolvente espectral, uma vez que os -12dB/oit presentes na fonte glotal podem não se manter quando esta fonte é ruído de turbulência.

Capítulo 3

Estado da arte

Neste capítulo, serão descritos os três métodos de terapia da fala a que os pacientes em reabilitação pós-laringectomia recorrem com maior frequência. No final, e porque interessa também perceber quais os avanços em termos de algoritmia, descrever-se-á um algoritmo cuja base é a de um codec CELP e cujo objetivo é o de reconstruir fala vozeada a partir de um sussurro, semelhante a esta dissertação.

3.1 Fala esofágica

Este método de fala consiste no uso do conjunto estômago e esófago para gerar e transportar o fluxo de ar necessário, ao invés dos pulmões e da traqueia. A língua deve permanecer encostada ao teto da boca para que a passagem do ar se dê para o esófago ao invés da traqueia. A Figura 3.1 ilustra o fluxo na fala esofágica.

A favor deste método, está o facto da fala gerada ser surpreendentemente inteligível. No entanto, é um método bastante complicado de dominar, sendo que apenas 30% dos pacientes laringectomizados o experimentam, e apenas um 1/5 dessas pessoas (6% do total), consegue usá-lo com sucesso [3].

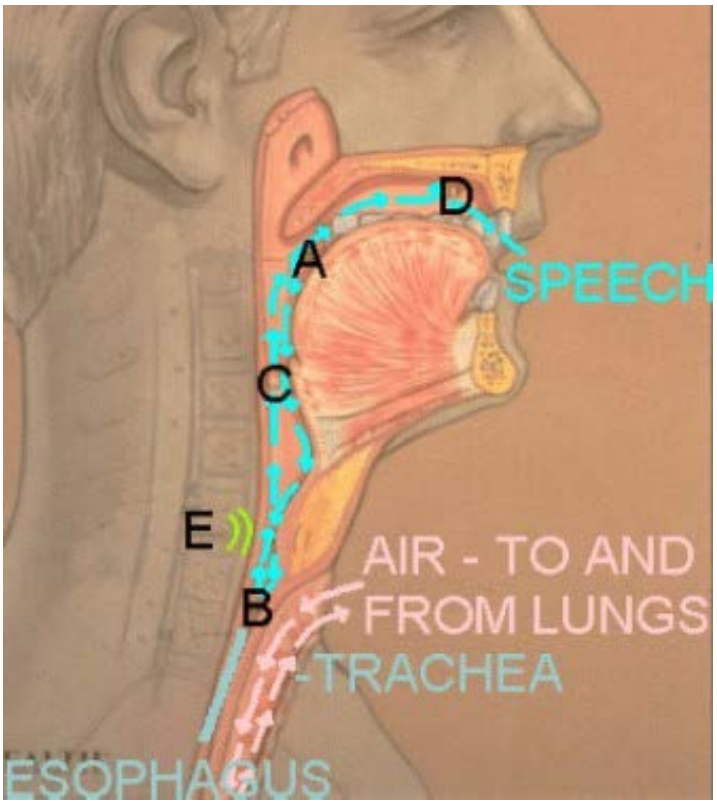

Figura 3.1: Fluxo de ar na fala esofágica [5].

3.2 Fala traqueoesofágica

Este método de reabilitação de fala é essencialmente usado em pacientes que sofreram uma laringectomia total e que respiram através de um *stoma* (traduz para estoma). É também o método preferido pela classe médica. Neste método, é criado um buraco entre a traqueia e o esófago. Este buraco é seguidamente preenchido com uma válvula unidirecional (prótese) que permite a passagem de ar da traqueia para o esófago, para que possa chegar ao trato vocal. Para que o ar não escape através do estoma, é necessário que a pessoa tape o estoma com o dedo durante o período de fala. A Figura 3.2 ilustra o fluxo na fala traqueoesofágica.

A favor deste método, estão a relativa boa qualidade da fala gerada e a relativa fácil usabilidade. No entanto, a prótese:

- Necessita de tratamento diário pela pessoa;
- Causa derrames com o tempo, necessitando de ser substituída;
- Aumenta o risco de infeções na área circundante.

Estima-se que 30% dos pacientes laringectomizados recorra a este método de terapia de fala [3].

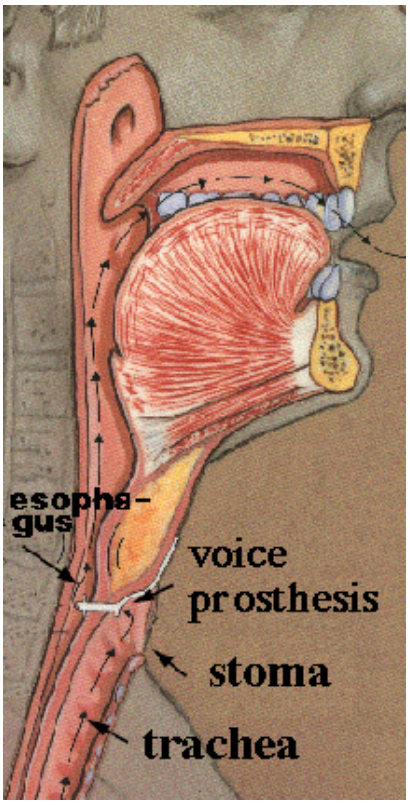

Figura 3.2: Fluxo de ar na fala traqueoesofágica [6].

3.3 Eletrolaringe

A eletrolaringe é um dispositivo que serve para introduzir vibração na fala produzida, simulando, grosseiramente, uma frequência fundamental no sinal de fala. Esta ressonância é modelada ao nível do trato vocal para produzir os diferentes fonemas. A Figura 3.3 ilustra o fluxo no uso de uma eletrolaringe.

A favor deste método, está a superior facilidade de uso relativamente aos anteriores. No entanto, o sinal de fala é caracterizado pejorativamente como "robótico". Estima-se que mais de 55% dos pacientes laringectomizados recorram a este engenho como forma de reabilitação de fala normal, fazendo deste a principal forma de reabilitação usada hoje em dia [3].

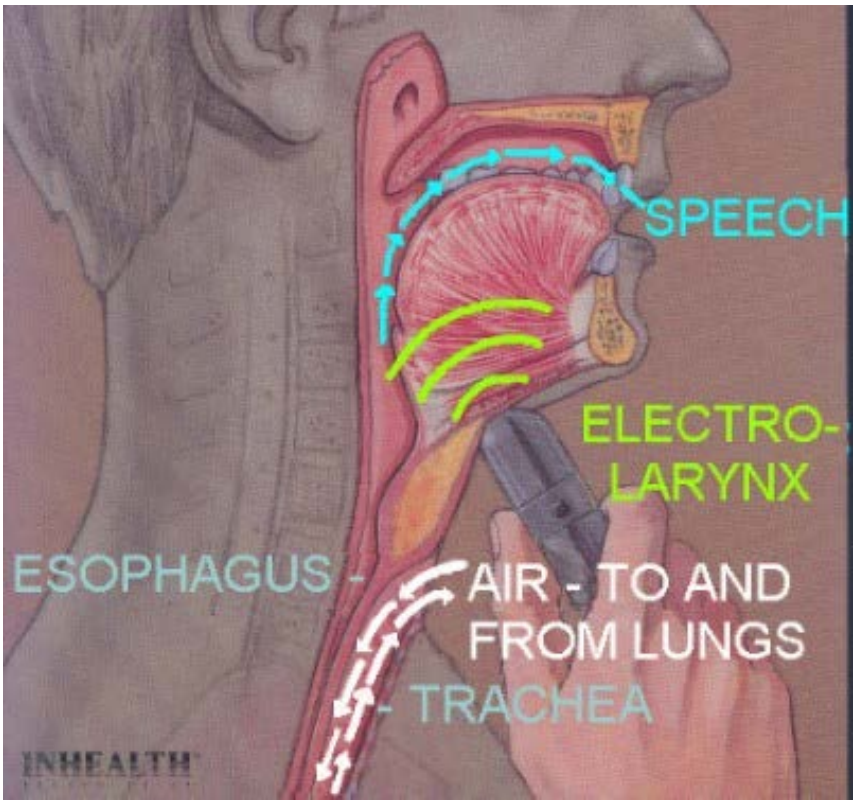

Figura 3.3: Fluxo de ar no uso de uma eletrolaringe [7].

3.4 Codec CELP modificado

Em Janeiro de 2011, a Universidade Tecnológica de Nanyang em Singapura desenvolveu um algoritmo cujo objetivo era o semelhante ao desta dissertação [3], exceto em três principais pontos:

- a língua alvo é o inglês;
- não são usadas amostras vozeadas como apoio ao vozeamento artificial da amostra sussurrada;
- existe um requisito de funcionamento do algoritmo em tempo real.

A abordagem consistiu em produzir um algoritmo modificado de um codec *Code Excited Linear Prediction* (CELP) para codificação do sinal de fala. Resumidamente, um codec CELP é um codificador usado nas comunicações móveis que, como outros codificadores baseados em predição linear, assume que a amostra $x[n]$ é calculável através de uma combinação linear de p amostras anteriores:

$$x[n] \quad = \quad \sum_{k=1}^{p} a_k x[n-k]. \tag{3.1}$$

Ainda com vista a reduzir o débito, este codec possui um módulo de deteção de voz denominado comummente por *Voice Activity Detector* (VAD) de forma a não gastar bits a transmitir momentos de silêncio.

Para que o codec funcionasse como vozeador artificial, tiveram de se realizar várias alterações ao seu funcionamento base:

- Foi necessário detetar fala sussurrada e não fala normal;
- Foi necessário classificar sons na fala sussurrada;
- Foi necessário transformar fala sussurrada em fala normal.

Os resultados obtidos desta experiência revelaram uma qualidade de fala normal produzida, de uma forma geral, superior à da qualidade da fala produzida pela eletrolaringe (a forma mais comum de reabilitação, como visto anteriormente). No entanto, foram encontrados dois grandes obstáculos não completamente ultrapassados:

- Sons problemáticos: pela forma como foi desenhado o algoritmo, alguns sons na fala sussurrada enfrentam uma transformação ambígua, havendo uma probabilidade se obter um som na fala vozeado que não corresponde ao desejado;
- Individualidade do orador: chegou-se à conclusão que a fidelidade à voz da pessoa não é completamente garantida com este método. Em suma, pensa-se que as características individuais introduzidas pelos pulmões e pelo trato vocal são mantidas; já aquelas que são introduzidas a nível da laringe, não se conseguem sintetizar.

3.5 Conclusão

Enquanto que a Universidade Tecnológica de Nanyang, em Singapura, colocou grande ênfase em vozear um sinal sussurrado sem qualquer referência, com este projeto deseja-se perceber, até que ponto, o uso de certas características desta referência aumentaria a qualidade do sinal artificialmente vozeado. Para tal, serão concebidas, não só uma versão do sinal artificialmente vozeado independente de qualquer referência, mas também uma ou mais versões que tirem partido do sinal vozeado original.

Capítulo 4

Ferramentas utilizadas

Neste capítulo, é feita uma enumeração das ferramentas de *software* utilizadas no desenvolvimento desta dissertação. A ordem pela qual são apresentadas é a mesma pela qual foram necessárias.

4.1 Adobe Audition CS6

O Adobe Audition CS6 [4] é uma poderosa ferramenta de gravação, análise e edição de som. Esta ferramenta foi muito útil em várias atividades ao longo da dissertação:

- Na gravação de amostras em formato *WAVeform audio format* (WAV);
- Na subamostragem destas amostras e na sua conversão para formato *Pulse Code Modulation* (PCM);
- Na representação visual de espectrogramas, tanto das amostras gravadas como das sintetizadas;
- Na reprodução das amostras sintetizadas de fala vozeada durante a realização do teste subjetivo descrito na secção 6.5.2.

A Figura 4.1 exibe a interface desta ferramenta.

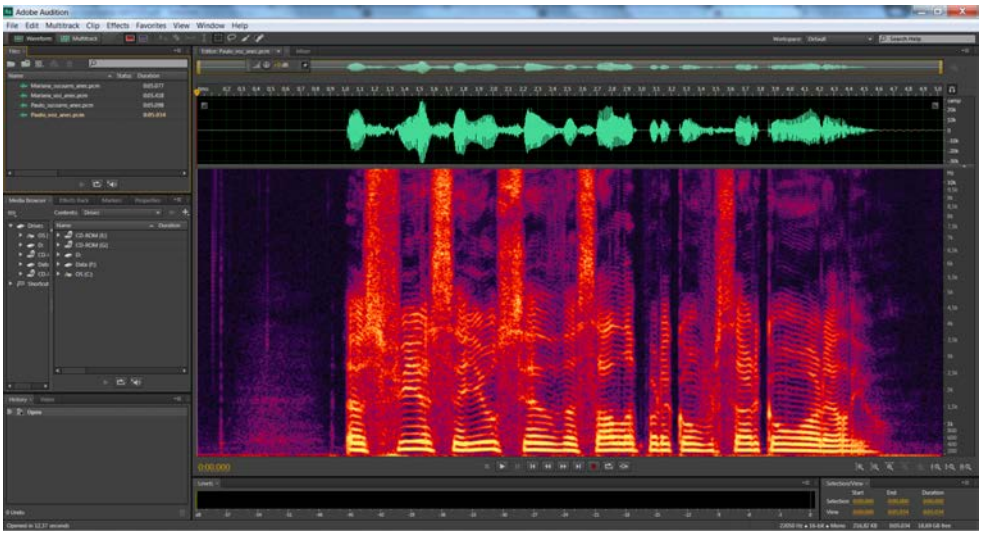

Figura 4.1: Interface do Adobe Audition CS6 [4].

4.2 Praat

O Praat [8] é uma ferramenta de gravação, análise e síntese de som.
As funcionalidades deste *software* foram pertinentes das seguintes formas:

- Novamente na visualização de espectrogramas das amostras gravadas em WAV;
- Na estimação automática de F_0 e $F_{1,2,3}$, bem como da energia do sinal;
- Na visualização de uma "fatia"vertical do espectrograma;
- Na segmentação manual das regiões a implantar vozeamento, através tanto da própria audição como da visualização do espectrograma dessas amostras.

A Figura 4.2 mostra a interface desta ferramenta quando se abre um ficheiro de som acompanhado de um *textgrid*. O *textgrid* é uma funcionalidade do Praat que permite a fácil segmentação manual do sinal sonoro. Neste caso, foi usado para a segmentação do sinal, como referido acima.

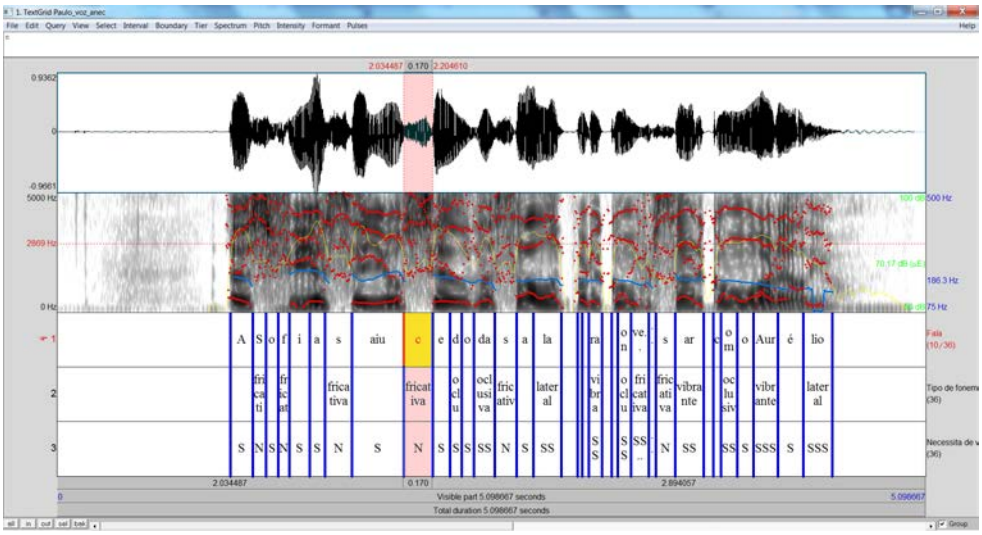

Figura 4.2: Interface do Praat [8] com um ficheiro de som e um *textgrid*. O "c"de "cedo"encontra-se selecionado sendo possível ouvir somente essa região do sinal.

4.3 MATrix LABoratory (Matlab)

O Matlab é uma poderosa ferramenta de programação orientada à resolução de problemas relacionados com cálculo numérico. Foi neste ambiente que a esmagadora maioria do processamento de sinal foi feito. De facto, todos os conteúdos expostos no capítulo 6, bem como a análise feita na secção 7.2.4, foram realizáveis graças a esta ferramenta. A Figura 4.3 expõe a interface de trabalho do Matlab. Na parte de trás, o *workspace* e, no centro, a codificação de uma função, a título de exemplo.

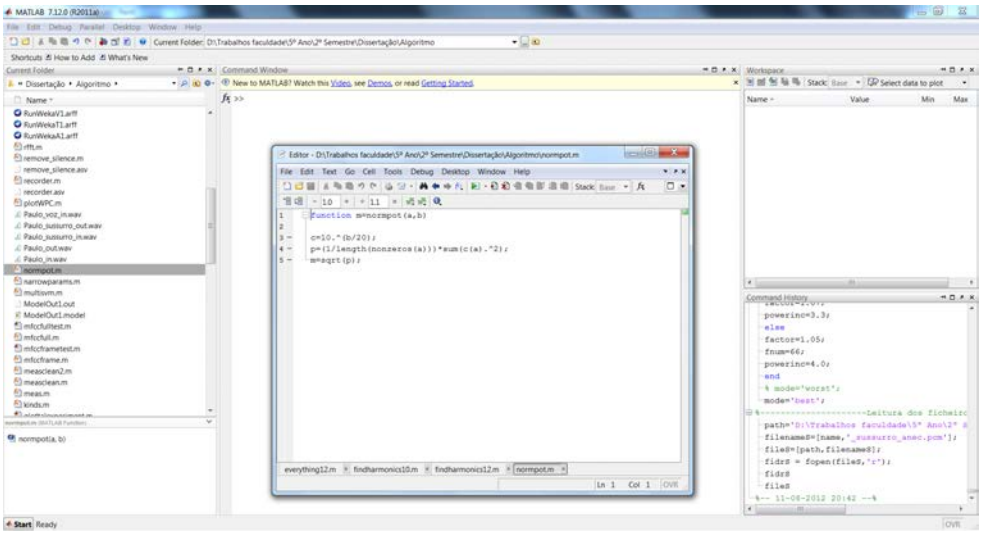

Figura 4.3: Interface do Matlab [9] (a grande) e um exemplo de uma função codificada nesta linguagem.

4.4 Waikato Environment for Knowledge Analysis (WEKA)

O WEKA [10] é um ambiente de *data mining* que proporciona a exploração dos dados. Incluiu algoritmos de redução de características, classificação e avaliação das propriedades discriminativas. Permite, ainda, receber dados extraídos de outros ambientes.

Esta ferramenta foi usada para simular dois algoritmos de aprendizagem de máquina supervisionados: Máquinas de Vetores de Suporte (MVS) e Redes Neuronais Artificiais (RNA). Ambos os algoritmos foram usados no sentido de apurar até que ponto é possível, no sinal de sussurro, a classificação automática de uma trama como fazendo parte de uma região que necessita de vozeamento ou não. Este procedimento é tratado na secção 7.2.

A Figura 4.4 ilustra uma das interfaces disponíveis desta ferramenta - a *explorer*, com um ficheiro de dados aberto. Neste caso, esta interface foi a usada para executar as simulações dos algoritmos mencionados.

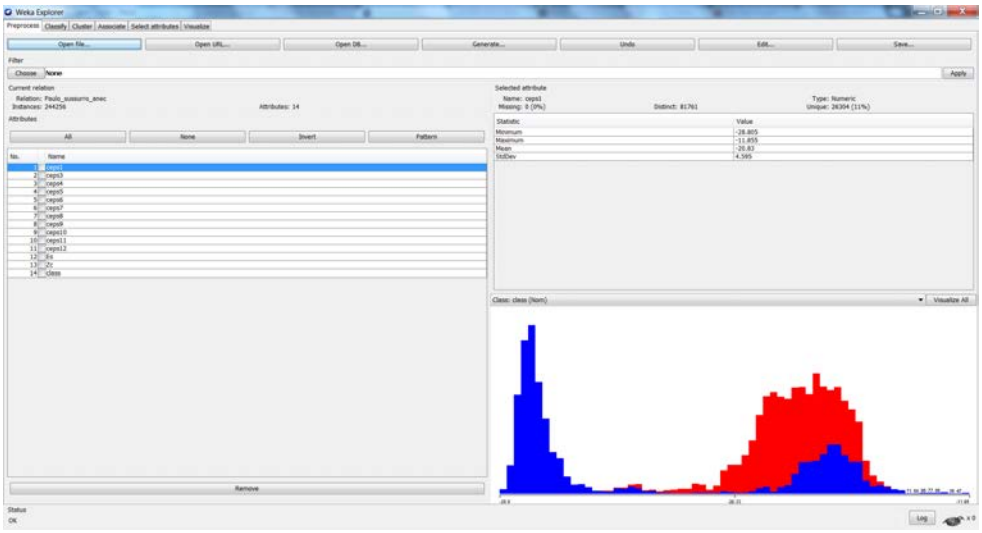

Figura 4.4: Interface *explorer* do WEKA [10].

4.5 Statistic Package for the Social Sciences (SPSS)

O SPSS [11] é uma ferramenta que contém funcionalidades de análise descritiva (caracterização da amostra) e inferencial (testes estatísticos) essencial para tratar os dados relacionados com a avaliação percetiva.

Esta ferramenta permitiu perceber até que ponto os parâmetros escolhidos para a classificação automática foram bem eleitos. Para este efeito, foi usada a análise Kruskal Wallis descrita na secção 7.2.3. A Figura 4.5 exibe a interface deste *software*.

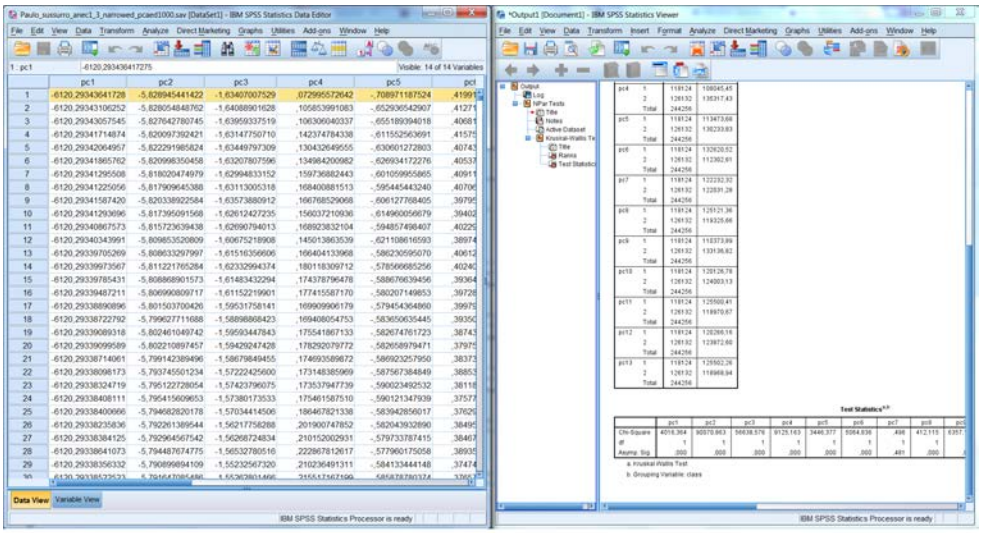

Figura 4.5: Interface do SPSS [11] dividida em 2 partes. Do lado esquerdo são visíveis os dados e do lado direito o resultado do teste Kruskal Wallis.

Capítulo 5

Recolha de amostras

Neste capítulo, aborda-se a primeira fase do desenrolar dos procedimentos desta dissertação: a recolha das amostras vozeadas e não vozeadas.

Este procedimento requer a escolha de um conjunto de atributos que foram pensados para melhor servir o propósito desta dissertação. O conjunto de palavras escolhido, o conjunto de oradores, o local e o equipamento usado são os pontos desta experiência que mais importam referir e cuja especificação se encontra disposta separadamente neste capítulo.

5.1 Frase

De acordo com o orientador do projeto, Eng. Aníbal Ferreira, o conjunto de sons a usar na realização desta prova de conceito formam uma única frase: "A Sofia saiu cedo da sala para conversar com o Aurélio". Esta frase é uma versão estendida da frase protocolo em terapia de fala: "Sofia saiu cedo da sala" [23]. Esta versão alongada foi concebida por forma a conter um bom alcance dos tipos de sons existentes na língua portuguesa, tornando os resultados desta dissertação mais válidos. Nesta frase encontram-se:

- pelo menos uma consoante oclusiva vozeada: "d"em "cedo";
- pelo menos uma consoante oclusiva não vozeada: "p"em "para";
- pelo menos uma consoante fricativa vozeada: "v"em "conversar";
- pelo menos uma consoante fricativa não vozeada: "S"em "Sofia";
- pelo menos uma consoante lateral: "l"em "Aurélio";
- pelo menos uma consoante vibrante: "r"em "conversar";
- vários tipos de vogais.

No entanto, é preciso estar alerta, uma vez que nem todos os sons são enunciados tal e qual como se se encontrassem fora do contexto de uma frase, o que altera a segmentação da frase

em regiões vozeadas e não vozeadas. Mais informação sobre a segmentação desta frase pode ser encontrada nas secções 6.2.1 e 6.3.1.

5.2 Oradores

Uma vez que o objetivo desta dissertação é a prova de conceito da possibilidade de vozear artificialmente um sinal de sussurro, não existe a necessidade de usar um numeroso conjunto de oradores. No entanto, teve-se em atenção, dois aspetos neste processo de escolha: a variedade de género e a saúde dos órgãos produtores de fala. Assim, foram selecionados dois oradores jovens, um do sexo masculino e um do sexo feminino, com idades respetivas de 22 e 21 anos.

5.3 Local

Inicialmente, fizeram-se gravações numa sala cujas propriedades acústicas não eram as melhores. A sala apresentava uma forma retangular comum e continha muitos objetos que contribuíam para comportamentos acústicos não favoráveis.

Após analisar o nível de ruído obtido nessas gravações percebeu-se que tinha de ser feita uma alteração ao nível da qualidade acústica do local. Usou-se então uma câmara anecoica, isto é, uma sala que, pela sua construção, não produz eco. A Figura 5.1 é uma fotografia tirada da câmara anecoica na direção que os oradores enfrentavam aquando a gravação das amostras.

Figura 5.1: Vista da câmara anecoica na direção para onde os oradores falavam.

5.4 Equipamento

Foram usadas 2 peças de equipamento na gravação das amostras, para além do próprio computador onde ficaram registadas: um microfone de alta qualidade e um pré amplificador inserido entre a ligação microfone-computador.

O microfone é o Ear Set 1 da Sennheiser [12]. Este microfone é extremamente sensível e apresenta uma inteligibilidade de discurso ótima, o que funciona bem com o ambiente de baixo ruido em questão. Mais ainda, o Ear Set 1 é adaptável à estrutura craniana do utilizador. A Figura 5.2 corresponde à imagem deste aparelho.

Figura 5.2: Ilustração do microfone Ear Set 1 da Sennheiser [12] utilizado na gravação de amostras de fala.

O pré amplificador é o UA-25EX da Roland [13]. Este amplificador é compatível com o microfone referido e possuí a capacidade de gravar com 24 bits de resolução e a frequências de amostragem de 96kHz. Para além disso, é compacto e portátil. A Figura 5.3 corresponde à imagem deste aparelho.

Figura 5.3: Ilustração do pré amplificador UA-25EX da Roland [13] utilizado na gravação de amostras de fala.

5.5 Especificações adicionais do procedimento

O conjunto de especificações relativas ao procedimento que ficou por definir nas secções anteriores encontra-se aqui listado:

- O microfone foi ajustado para se situar lateralmente aos lábios dos oradores a cerca de 2cm destes, no sentido de evitar a deteção de sopros de ar;
- A potência do pré amplificador foi iterativamente ajustada para que as amostras gravadas tivessem grande parte da sua amplitude acima dos 50% do máximo, mas sem nunca incorrer em *clipping*;
- Houve um grande cuidado em que, para cada orador, a duração das versões vozeada e não vozeada não diferisse muito, de forma a que, posteriormente, fosse mais fácil qualquer processo de sincronização entre as duas;
- O *software* utilizado na gravação das amostras foi o Adobe Audition CS6 [4], tal como está mencionado na secção 4.1;
- O formato das amostras foi WAV, com Frequência de amostragem (=*Sampling*) (F_s) de 48kHz para permitir captar todo o espectro audível e com 32 bits de resolução, no sentido de minimizar o erro de resolução;
- No entanto, por questões de compatibilidade com o código, as amostras tiveram posteriormente de sofrer uma conversão para 22050Hz de F_s e 16 bits de resolução. As amostras foram ainda convertidas para PCM. Todas estas conversões foram realizadas no Adobe Audition CS6 [4].
- A duração das amostras, sem contar com os períodos de silêncio, antes e depois da enunciação, são as dadas pela tabela 5.1.

Tabela 5.1: Duração de cada amostra recolhida, descontando os períodos de silêncio iniciais e finais.

Tipo de fala	**Orador**	
	Masculino (s)	**Feminino (s)**
Vozeada	3,600	3,519
Não Vozeada	3,646	3,541

O anexo B contém o conjunto completo de ilustrações dos sinais recolhidos pós conversão. Também se encontram presentes os espectrogramas respetivos.

Capítulo 6

Transformação do sinal sussurrado em vozeado

Neste capítulo, explicar-se-ão sucintamente os procedimentos que permitiram transformar artificialmente a amostra sussurrada numa amostra vozeada. No final, tecem-se as conclusões que a análise dos resultados permitiu elaborar.

Em concordância com o objetivo desta dissertação, foi desenhado um algoritmo cuja saída são duas versões distintas do sinal artificialmente vozeado. A primeira versão faz uso das seguintes características do sinal vozeado original: em cada trama, as frequências a que os seus harmónicos ocorrem e a energia total da trama. A segunda versão pode ser considerada completamente independente do sinal vozeado original: a única informação retirada é o valor médio da frequência fundamental ao longo de todas as tramas, algo que podia ser facilmente uma informação exterior sobre o orador ou até um parâmetro ajustável em função da sonoridade resultante. Por motivos de simplicidade, atribuem-se, a estas versões, os nomes "dependente"e "independente", respetivamente.

Pretende-se, com esta dualidade de resultados, perceber até que ponto é necessária, ou útil, a informação presente na glote para vozear o sinal sussurrado, bem como, perceber até que ponto a informação da glote é a principal detentora das características que concedem identidade ao orador. Mais ainda, procura-se avaliar outras características, tais como a inteligibilidade da voz e a sua naturalidade. Onde inteligibilidade significa: até que ponto se percebeu o que foi dito; e naturalidade significa: até que ponto a voz soa natural, ou seja, sem características de artificialidade.

As duas versões aqui em causa, em termos de dependência do sinal vozeado original para sua construção, são dois extremos do escopo: a versão dependente é muito dependente, e a versão independente é muito independente. Assim, podiam ter sido contempladas outras versões com graus de dependência entre estes limites, cujos benefícios seriam outras conclusões ao nível da importância da informação original da glote na produção de fala. Por exemplo, podia existir uma versão semelhante à versão aqui chamada independente que, em vez de se basear num F_0 médio, usasse o valor de F_0 de cada trama. Apesar das suas vantagens, tal não foi feito uma vez que as amostras produzidas são objeto de avaliação por um conjunto de voluntários, e o facto

de cada pessoa ter de avaliar mais amostras poderia facilmente tornar o processo de avaliação cansativo, algo que se quis evitar. Mais informação sobre a análise subjetiva de resultados pode ser encontrada na secção 6.5.2.

Na próxima secção deste capítulo, dá-se uma visão dos vários módulos do algoritmo. Nas secções posteriores, exclusive a última, explica-se cada um dos módulos mais pormenorizadamente. Sempre que for necessário, explicitar-se-ão as diferenças no algoritmo que permitem obter cada uma das versões referidas. Por fim, na última secção, criticam-se os resultados obtidos recorrendo a uma avaliação subjetiva dos sinais produzidos.

6.1 Abordagem

No sentido de dar ao leitor uma visão geral das várias partes do algoritmo, foi composto um diagrama de blocos presente na Figura 6.1.

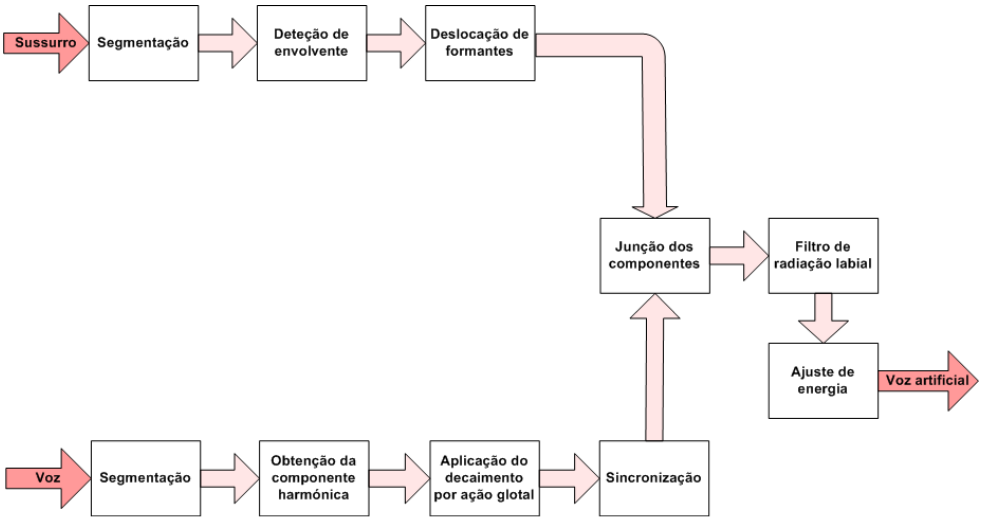

Figura 6.1: Diagrama de blocos do algoritmo de vozeamento artificial desenvolvido.

Neste sistema, as duas entradas correspondem às duas amostras recolhidas de um dos oradores. Cada um dos ficheiros possui uma frequência de amostragem de 22050Hz, 16 bits de resolução e uma duração aproximada de 5 segundos, como referido na secção 5.5. A leitura é feita em tramas de 1024 amostras com uma sobreposição de 50%.

O sinal de sussurro é segmentado nas partes que devem ou não ser vozeadas. De seguida, nas tramas que necessitam de vozeamento, é feita a extração da envolvente espectral, no sentido de tentar modular o filtro do trato vocal. Por fim, modifica-se este filtro de forma a compensar o deslocamento de formantes que ocorre de um tipo de fala para o outro.

Por sua vez, o sinal de voz é também segmentado, desta vez nas partes vozeada e não vozeada. Seguidamente, das tramas consideradas vozeadas extraem-se os harmónicos e faz-se uma normalização de amplitude dos mesmos para remover a informação de magnitude do trato vocal.

Posteriormente, aplica-se um decaimento aos harmónicos normalizados, no sentido de simular o decaimento glótico. Por fim, como a frase de teste não apresenta o mesmo ritmo e duração nos diferentes ficheiros, é necessário sincronizar as tramas vozeadas no sinal vozeado com as tramas que necessitam de vozeamento no sinal sussurrado.

De acordo com o modelo fonte-filtro, um sinal de fala vozeado é a multiplicação da componente harmónica pelo filtro do trato vocal e pelo filtro de radiação labial. Assim, multiplicam-se as duas componentes já obtidas: harmónicos e filtro do trato vocal. O terceiro passo é a multiplicação deste resultado por um filtro de radiação labial sintetizado, obtendo-se um sinal de fala artificial. Finalmente, há que ajustar a amplitude das tramas sintetizadas de forma a que obtenham um valor de energia correspondente à energia de uma trama vozeada.

Encontra-se terminado o algoritmo. A escrita no ficheiro é feita em *overlap-add* e segue uma lógica simples: se a trama a ser escrita for uma trama que diga respeito a uma região que necessite de vozeamento artificial, escreve-se a trama sintetizada correspondente; de outra forma, escreve-se a trama presente no sinal de sussurro original correspondente.

6.2 Processamento do sinal sussurrado

Nesta secção, detalhar-se-ão os procedimentos referentes à amostra sussurrada recolhida, estendendo assim a sumarização presente na secção anterior.

Pretende-se, para cada trama do sinal de sussurro, obter o filtro do trato vocal presente no modelo fonte-filtro.

6.2.1 Segmentação

A segmentação do sinal de sussurro consiste na distinção de duas áreas da amostra. Uma área correspondente ao sons vozeados e outra aos sons não vozeados, como descrito na secção 2.2. Embora, em rigor, num sussurro não existam zonas vozeadas, devido à completa inação das pregas vocais em termos de repetição periódica, por motivos de simplicidade, denominam-se estes sons "vozeados"e "não vozeados", pela nomenclatura que herdam por serem enunciados em fala vozeada.

Portanto, primeiro há que distinguir quais as áreas da amostra que necessitam de vozeamento e quais as áreas que não precisam. Pelo que se sabe dos ensinamentos da fonética, de todos os sons do PE, apenas as consoantes oclusivas não vozeadas e as consoantes fricativas não vozeadas são sons não vozeados - o que significa que não devem ser alvo de vozeamento artificial na conversão do sussurro. Assim, segmentando estas consoantes no sinal sussurrado, fica-se automaticamente com as duas categorias desejadas: uma porção do sinal que não deve ser alvo de vozeamento artificial e uma porção do sinal que deve.

A segmentação do sinal sussurrado foi feita manualmente recorrendo ao Praat [8]. A funcionalidade *textgrid* do Praat possui uma ferramenta de camadas, bastante útil na tarefa de segmentação - é possível ouvir partes do sinal e dar-lhes uma designação, por exemplo: "necessita de vozeamento"ou "não necessita de vozeamento", isto enquanto se observa o espectrograma do sinal.

Note-se que a segmentação teve de sofrer uma alteração ao longo da dissertação, visto que, apesar do "o"de "Sofia"representar um som vozeado, as suas tramas não apresentam estrutura harmónica definida. De facto, na enunciação de "Sofia"é muito difícil de distinguir o "o", ouvindo-se "Sfia". Tal alteração leva ao desaparecimento da região vozeada correspondente ao "o", sendo que se considera agora "Sof"como sendo um segmento não vozeado. Tal diferença entre escrita e enunciação é normal quando se trata da enunciação de uma série de sons agregados (=palavras) [2].

Feita esta segmentação no Praat, apenas foi necessário passar os tempos em que começam e terminam as regiões que necessitam de vozeamento para o algoritmo em Matlab. Com base nestes tempos, sinalizam-se as tramas que necessitam de vozeamento.

Importa perceber que, numa aplicação para uso quotidiano deste algoritmo à fala sussurrada dos pacientes laringectomizados haveria, por razões óbvias, uma obrigatoriedade da segmentação ser feita automaticamente. Considerando este facto, foi desenvolvido um método de segmentação automática discutido separadamente no capítulo 8. A razão pela qual este método de classificação automática não foi integrado no algoritmo aqui descrito, deve-se ao facto deste método apresentar resultados tão bons que a sua integração não alteraria praticamente o resultado final.

6.2.2 Deteção de envolvente

O objetivo deste módulo é o de estimar o filtro do trato vocal implícito no espectro de cada trama a vozear.

Sabe-se que num sinal de fala sussurrada a fonte é ruído, o trato vocal é aproximadamente plano com picos nas frequências formantes, e o filtro de radiação labial é um diferenciador. Pensou-se então que, eliminando a ação do ruído e do filtro de radiação labial, restaria apenas o filtro do trato vocal.

Dada a natureza suave do filtro do trato vocal, procurou-se uma estratégia para suavizar o espectro de cada trama na tentativa de eliminar o mais possível a presença do ruído da fonte. Portanto, se se conseguisse aplicar um filtro passa-baixo não à trama, mas ao espectro da trama, obter-se-ia um espectro suavizado livre de grande parte do ruído. Esta operação não é mais do que calcular a envolvente cepstral de uma trama: calcular o espectro do espectro (na escala logarítmica) e colocar a zero os p últimos coeficientes obtidos, onde p é a ordem da envolvente cepstral.

O valor de p foi escolhido de forma a conservar o mais possível a informação das formantes e, ao mesmo tempo eliminar, a informação do ruído de fonte. Conseguiu-se um resultado apelativo com $p = 100$. A Figura 6.2 ilustra a envolvente cepstral da primeira trama do "A"de "A Sofia...". Neste, são perfeitamente visíveis as 4 primeiras formantes nas frequências: 600, 1780, 3080 e 4110 Hertz.

Havendo retirado grande parte da informação do ruído da fonte, é ainda visível um decaimento nos picos das formantes que não faz parte do filtro do trato vocal. Este decaimento pode ser visto como a soma do decaimento negativo do ruído da fonte com o aumento introduzido pelo filtro de radiação labial. O que se faz neste algoritmo é usar esta envolvente cepstral como aproximação do trato vocal e no filtro de radiação labial sintetizado na secção 6.4.2 corrige-se o decaimento que houver a corrigir.

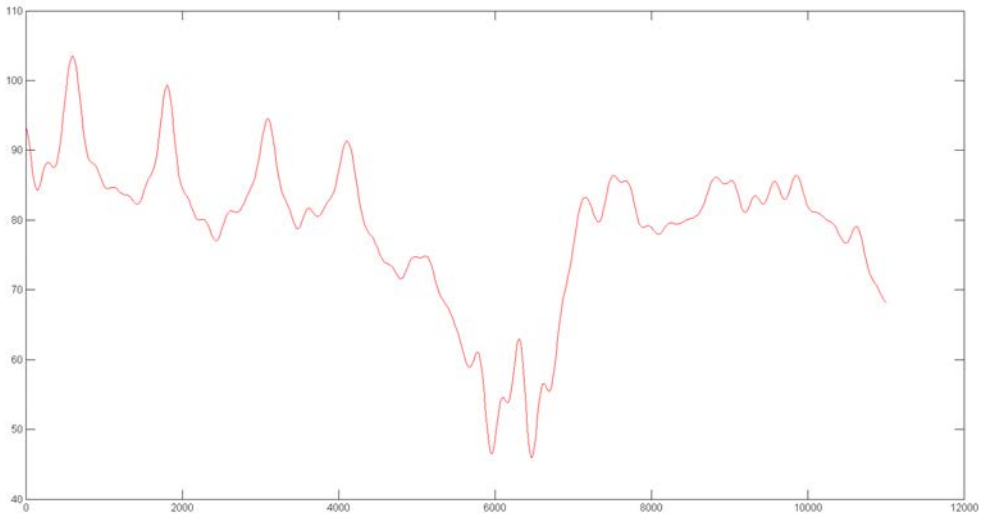

Figura 6.2: Envolvente cepstral de ordem 100 de uma trama correspondente ao início do "A"de "A Sofia...".

6.2.3 Deslocação de formantes

A envolvente cepstral, ao contrário da envolvente LPC, não possui uma forma fácil de descobrir as frequências formantes e de as alterar. Assim, teve de ser utilizada uma estratégia alternativa por forma a conseguir este deslocamento de frequências. Pensou-se, então, que seria uma alternativa razoável realizar um escalonamento do espectro.

Se se representar o espectro da envolvente cepstral de uma trama m por $|S_m(w)|_{dB}$, e a for o fator de escalonamento, então faz-se: $|S_m(aw)|_{dB}$. Se $a > 1$ o espectro é comprimido. Os valores de a que demonstraram melhores resultados foram 1.07, no caso do orador masculino e 1.05, no caso do orador feminino.

A desvantagem desta técnica é que, em vez de deslocar apenas as formantes, todo o espectro é deslocado. Para além disso, quando mais alta a frequência, mais para trás é deslocado o valor de magnitude do gráfico correspondente, o que significa que frequências formantes mais altas vão ser deslocadas de um valor maior. Este segundo ponto é especialmente incómodo, uma vez que vimos na secção 2.3 que, em média, quanto mais alto o valor da frequência formante, menos ela se deve deslocar.

Contudo, os resultados obtidos após inserção deste módulo no algoritmo revelaram-se superiores em qualidade face aos obtidos até à data.

6.3 Processamento do sinal vozeado

Nesta secção, pormenorizar-se-ão os procedimentos realizados sobre a amostra vozeada recolhida, também estes resumidos na secção 6.1.

Pretende-se, para cada trama do sinal de voz, obter a componente harmónica de acordo com o modelo fonte-filtro.

6.3.1 Segmentação

A segmentação do sinal vozeado foi, em tudo, semelhante à segmentação do sinal sussurrado. Identificaram-se as zonas vozeadas e não vozeadas no Praat [8], passaram-se esses tempos para o Matlab e sinalizaram-se as tramas correspondentes às regiões vozeadas.

Não foi pensado qualquer método de segmentação automática do sinal vozeado, uma vez que, numa aplicação prática deste algoritmo, não se faria uso de uma amostra vozeada para vozear a sussurrada.

6.3.2 Obtenção da componente harmónica

Este módulo pretende obter a componente harmónica de cada trama vozeada do sinal de voz.

É também o módulo onde se encontra a principal diferença entre as duas versões geradas. Como havia sido referido, este algoritmo produz uma versão dependente que usa informação de cada trama vozeada do sinal de voz, e uma versão independente que apenas usa uma frequência fundamental média extraída do sinal de voz. Assim, esta secção divide-se em duas partes correspondentes às duas versões mencionadas. Em cada uma explicita-se a lógica por trás dos procedimentos efetuados, seguido da explicação dos procedimentos em si.

6.3.2.1 Versão dependente

Sabe-se que o sinal vozeado em análise é composto por uma componente harmónica (fonte), pela ação do filtro do trato vocal (filtro) e pela ação do filtro de radiação labial. O objetivo deste módulo é o de extrair apenas a componente harmónica, visto que esta é a informação que não se encontra no sinal de sussurro. Tal procedimento encontra-se descrito nos próximos parágrafos.

Primeiramente, o espectro de cada trama é analisado no sentido de se descobrir a frequência fundamental. Posteriormente, calculam-se uns harmónicos "teóricos"como sendo os múltiplos inteiros da frequência fundamental encontrada. De seguida, analisa-se novamente o espectro da trama sinalizando-se, para cada harmónico teórico, o máximo local mais próximo em frequência; consideram-se que estes novos picos são os verdadeiros harmónicos.

Uma vez conhecidas as frequências dos harmónicos, descarta-se a trama vozeada original e sintetizam-se novos harmónicos nas frequências encontradas. A amplitude destes harmónicos é toda aproximadamente igual, de forma a eliminar qualquer influência do trato vocal. O valor da amplitude em si não importa, visto que no final do algoritmo se multiplica o espectro todo por um fator, por forma a ajustar a energia de cada trama. Para além da informação do trato vocal, perde-se também a informação do decaimento por ação glotal (-12dB/oit) e do filtro de radiação labial (+6dB/oit). O primeiro decaimento é imposto na secção 6.3.3 e o segundo é imposto, logicamente, apenas no sinal de fala vozeada artificial final.

A Figura 6.3 ilustra a componente harmónica normalizada da primeira trama do "A"de "A Sofia...". Acima dos 2500Hz não foi detetada mais componente harmónica no sinal vozeado original. Portanto a síntese do harmónicos acaba por aí.

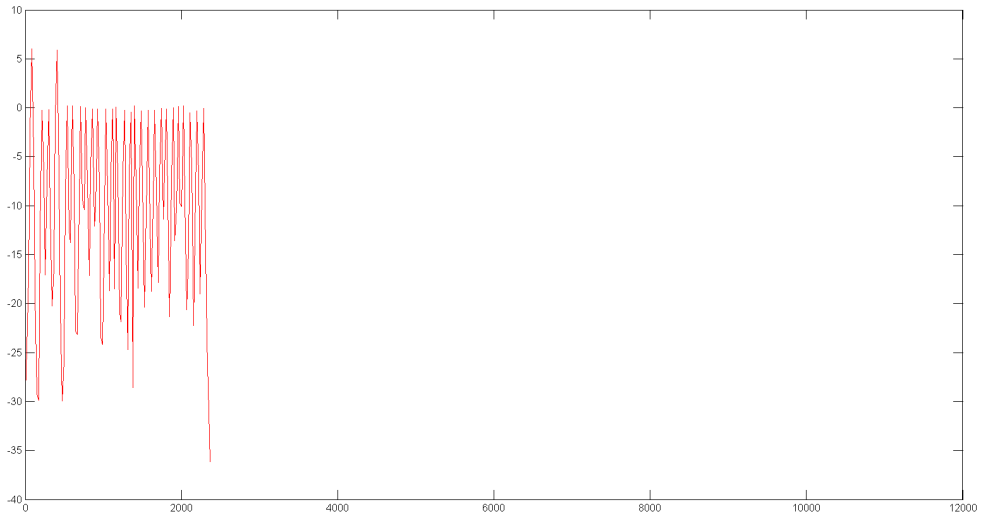

Figura 6.3: Componente harmónica normalizada da primeira trama do "A"de "A Sofia...".

6.3.2.2 Versão independente

Como foi mencionado anteriormente, esta versão não depende praticamente do sinal vozeado recolhido. A única informação disponível é a frequência fundamental média do orador. Assim, há primeiro que calcular a frequência fundamental média do orador e, depois, sintetizar uns harmónicos iguais para todas as tramas com base nessa frequência fundamental.

Usa-se a mesma rotina usada na versão dependente para calcular F_0 para cada trama. No final, obtém-se a média destes valores. Mais nenhuma vez se recorre ao sinal vozeado original.

Os harmónicos teóricos para todas as tramas são calculados como sendo os múltiplos inteiros da frequência fundamental média. Não havendo mais informação, a síntese dos harmónicos usa estes valores teóricos e a amplitude é também normalizada da mesma forma e pela mesma razão que na secção anterior.

6.3.3 Aplicação do decaimento por ação glotal

Sintetizada a componente harmónica normalizada, há que introduzir o decaimento por ação glotal de -12dB/oit. Inicialmente, foi feita a multiplicação nas frequências da componente harmónica por uma rampa com o declive mencionado. Posteriormente, melhorou-se o resultado obtido ao usar a síntese de um único impulso glotal como rampa.

Para cada trama, usa-se o valor de F_0 em causa (sempre o mesmo valor na versão independente) para sintetizar um impulso glotal nos tempos. Este impulso glotal tem sido alvo de estudo há anos. O modelo usado aqui é o popular modelo Liljencrants-Fant (LF). A Figura 6.4 ilustra a forma

de onda do impulso glotal de acordo com o modelo LF e sua derivada incluindo os diferentes parâmetros que o caraterizam.

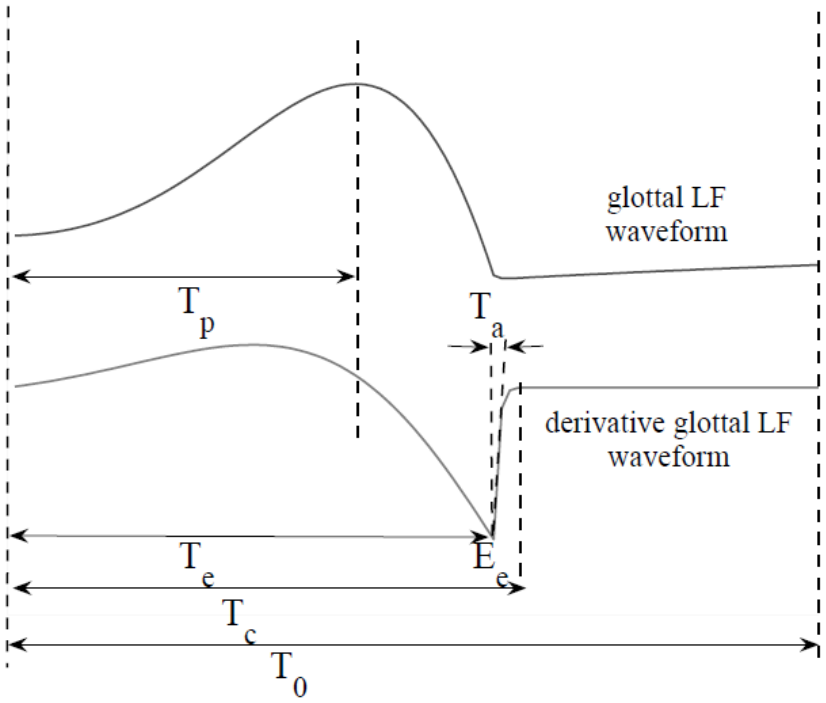

Figura 6.4: Ilustração da forma de onda do impulso glotal LF e da sua derivada [14].

Os diferentes parâmetros de afinação Tp, Te, Ta, Tc e Ee foram retirados de [24], de modo a se obter um modelo glotal representante de uma voz saudável. No sentido de se simular um impulso glótico mais parecido com o impulso glótico humano, acrescentou-se ruído branco gaussiano à forma de onda do impulso glotal sintetizado. De acordo com o co-orientador desta dissertação, o ruído deve ser acrescentado na zona aberta da glote. Assim, realizou-se este acréscimo da seguinte maneira:

$$igr[n] \quad = \quad ig[n] + ig[n] \times r[n]. \tag{6.1}$$

Onde $igr[n]$ e $ig[n]$ são as formas de onda do impulso glotal após e antes da inserção do ruído, respetivamente, e $r[n]$ é o ruído branco gaussiano. Uma vez que, durante a fase em que a glote se encontra quase fechada o nível de pressão do ar é quase zero, a multiplicação entre o ruído e a forma de onda leva à forte atenuação do sinal de ruído original nas zonas onde a glote se encontra quase fechada. Somando este novo sinal com a forma de onda do impulso glótico original, obtém-se um impulso glótico com uma variação de pressão, ligeira mas significativa, apenas durante o período em que a glote se encontra aberta. Um exemplo do novo impulso glótico pode ser visualizado na Figura 6.5.

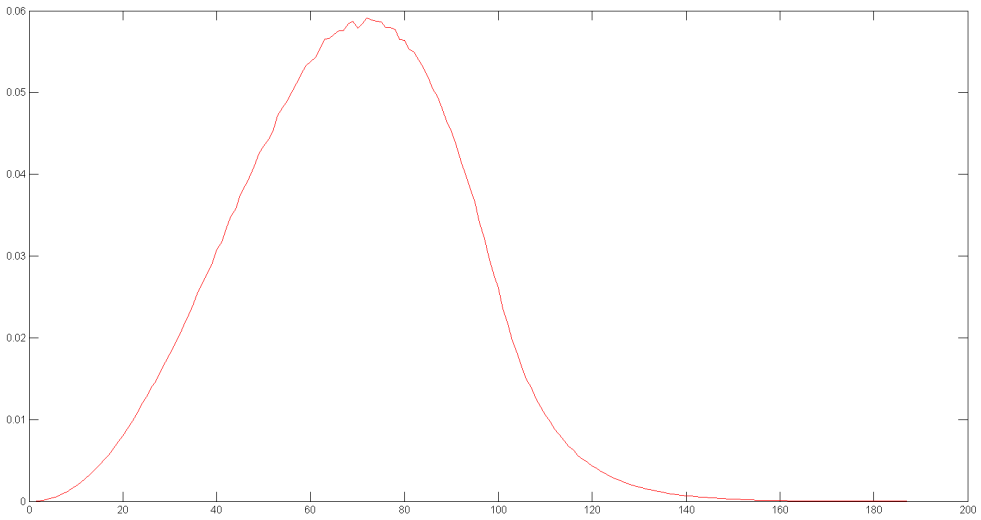

Figura 6.5: Ilustração da forma de onda do impulso glotal LF, após acréscimo de ruído branco gaussiano, na primeira trama de "A"em "A Sofia...".

Depois de gerado no domínio dos tempos, o impulso glotal foi alvo da transformada de Fourier. O seu módulo, na primeira trama de "A"em "A Sofia...", encontra-se na Figura 6.6.

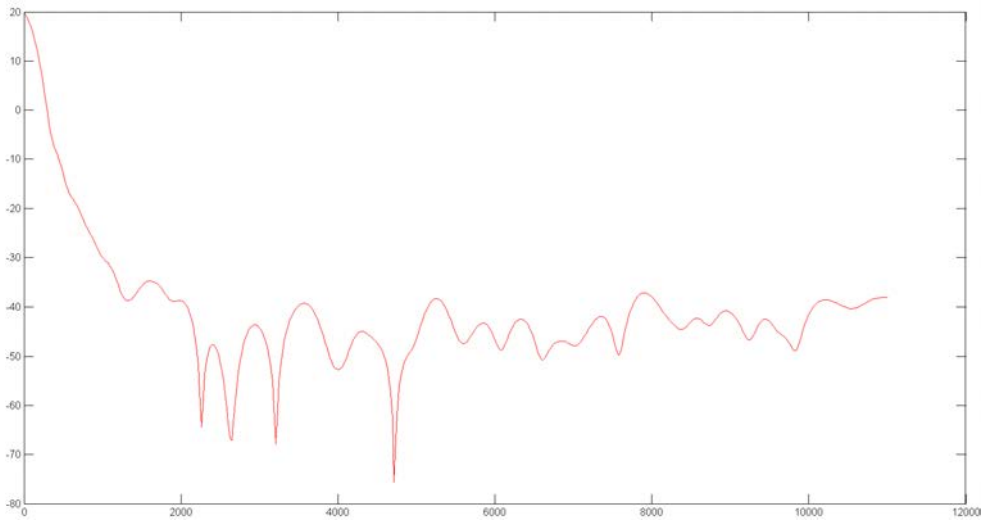

Figura 6.6: Módulo da transformada do impulso glotal gerado para a primeira trama de "A"em "A Sofia...".

6.3.4 Sincronização

Como havia sido mencionado no capítulo 5, houve um cuidado intra orador no sentido de enunciar a frase com a mesma cadência e duração total em ambos os tipos de fala. Logicamente,

não foi obtida tal perfeição, por conseguinte, existe a necessidade de desenvolver um módulo que faça corresponder as tramas de uma região vozeada no sinal vozeado às tramas que necessitam de vozeamento no sinal sussurrado, para que se consigam realizar operações entre os dois sinais.

Existem dois problemas a colmatar neste módulo:

- o facto das regiões vozeadas e a vozear análogas não começarem nos mesmos instantes;
- o facto das regiões vozeadas e a vozear análogas não terem a mesma duração.

Para resolver o primeiro problema, o que se faz é transladar as regiões vozeadas do sinal vozeado até a primeira trama de cada região vozeada se encontrar alinhada com a primeira trama da região a vozear correspondente. Para resolver o segundo problema, faz-se uma decimação ou interpolação linear a cada região vozeada do sinal vozeado, de forma que cada uma destas regiões passe a ter o mesmo tamanho que a região a vozear correspondente.

Feito isto, estão reunidas as condições para a realização de operações entre o filtro do trato vocal estimado e a componente harmónica sintetizada.

6.4 Reconstrução do sinal

Nas secções integrantes desta secção, explicitam-se os procedimentos que permitiram trabalhar as duas componentes obtidas até aqui, num único sinal de fala vozeada.

6.4.1 Junção das componentes

No sentido de produzir um sinal vozeado artificial, é necessária a fonte glótica, o filtro do trato vocal e o filtro de radiação labial. Neste momento possui-se, para a versão dependente, uma fonte glótica obtida à custa da amostra vozeada original, e, para a versão independente, uma fonte glótica criada a partir de um valor F_0 constante. Estas fontes glóticas encontram-se em sincronismo com um filtro do trato vocal, filtro este que é o mesmo, não obstante a versão final que se pretenda gerar.

A operação realizada entre estes dois elementos foi uma simples multiplicação nas frequências. Há, no entanto, um ponto muito interessante nesta multiplicação que se prende com o domínio da fase; algo que, até agora não havia sido mencionado. Na geração do filtro do trato vocal, a fase que se considerou foi a fase original do sinal sussurrado, trama a trama. Já na obtenção da componente harmónica, mais concretamente, na sintetização do impulso glotal, não foi guardada a fase da transformada deste impulso, uma vez que, não houve a preocupação de deslocar nos tempos o impulso glótico de modo a manter a continuidade da forma de onda ao longo das várias tramas; de facto, o impulso gerado em cada trama vozeada começava sempre no início da fase de abertura da glote e terminava no final do ciclo glótico. Contudo, existe uma razão para esta escolha. Na secção 6.3.3, foi mencionado que inicialmente se havia usado uma rampa -12dB/oit ideal como forma de modular o decaimento por ação glotal. Nessa mesma altura, e na geração da versão dependente, guardou-se a fase de cada trama vozeada do sinal vozeado e considerou-se que a

componente harmónica gerada herdaria essa fase inalterada. Ora, ao somar essa mesma fase com a fase do filtro do trato vocal o resultado final era, percetualmente, idêntico a um sinal final onde se considerava inexistente a fase da componente harmónica. Tal realização levou ao relaxamento da obtenção da fase da componente harmónica, aquando do uso do impulso glótico na modelização do decaimento por ação glotal.

Assim, a fase do sinal vozeado artificialmente é, unicamente, a fase do sinal sussurrado original.

A Figura 6.7 ilustra o módulo do sinal reconstruido na primeira trama de "A"em "A Sofia...".

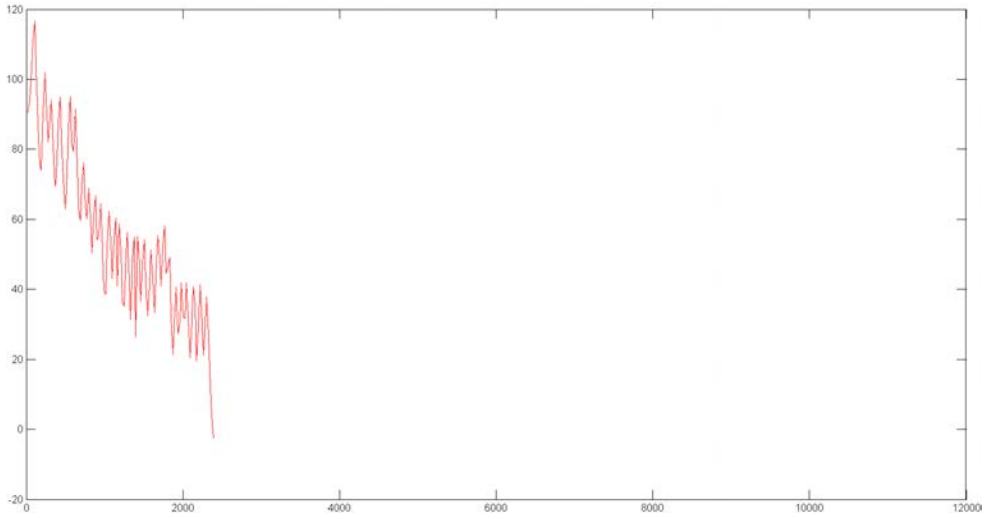

Figura 6.7: Módulo da junção da componente harmónica com o filtro do trato vocal na primeira trama de "A"em "A Sofia...".

6.4.2 Filtro radiação labial

O objetivo deste módulo é o de simular o efeito que os lábios provocam na produção de fala. Tal como havia sido referido, este efeito pode ser aproximado como a passagem do sinal obtido até este ponto por um diferenciador, ou seja, a multiplicação (ou soma em dB) do sinal por uma reta de declive +6dB/oit. No entanto, neste caso em particular, poderia não ser bem este o declive, uma vez que o filtro do trato vocal estimado apresentava um decaimento indesejado. Contudo, após várias tentativas com valores superiores a +6dB/oit, o resultado final não se revelou qualitativamente superior, pelo que foi decidido manter-se os +6dB/oit.

Foi então multiplicado este efeito diferenciador ao módulo do sinal obtido até agora, tendo a fase permanecida inalterada.

A Figura 6.8 ilustra o módulo do sinal após atuação do módulo do filtro de radiação labial na primeira trama de "A"em "A Sofia...".

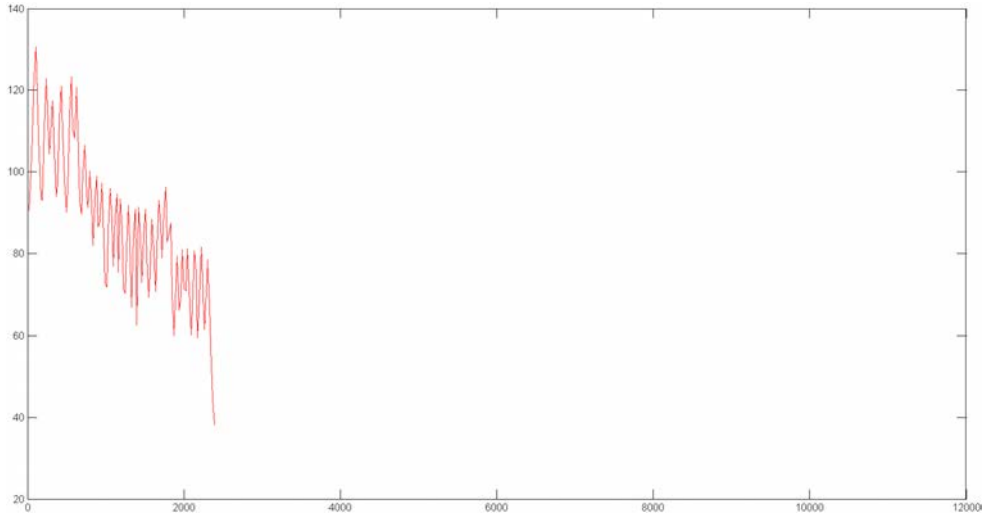

Figura 6.8: Módulo da junção da componente harmónica, com o filtro do trato vocal e com o filtro de radiação labial na primeira trama de "A"em "A Sofia...".

6.4.3 Ajuste de energia

Pode-se dizer que a forma do sinal vozeado artificial final estava, neste momento, completamente determinada tanto em módulo como em fase. O único aspeto ainda por ajustar era a energia do sinal resultante.

Na síntese da componente harmónica normalizada, houve a preocupação de se obter a correta forma dos harmónicos, no entanto os valores das suas magnitudes, embora todos semelhantes, foram arbitrados sem qualquer critério. Isto, porque, caso se tivesse conservado, por exemplo, a energia das tramas vozeadas (na versão dependente), do processo de junção das componentes fonte glótica e filtro do trato vocal, sendo este uma multiplicação no domínio das frequências, resultariam tramas com magnitude média muito superior a qualquer trama vozeada normal. Tal facto levaria, de igual forma, a um ajuste da energia final das tramas.

Assim, concentrou-se a necessidade da correção da energia no último módulo deste algoritmo. O que se faz é multiplicar o módulo do espectro por um valor constante, de modo a tornar a energia resultante da trama, semelhante à energia da trama vozeada original correspondente. Existe, neste ponto, uma pequena diferença entre as versões dependente e independente. A razão sendo que, na versão independente, não se tem acesso à energia das tramas do sinal vozeado original. Os próximos parágrafos explicam como se ajustou a energia para cada um dos casos.

Na versão dependente, calculou-se a potência média da trama sintetizada e a potência da trama vozeada original correspondente. Se se apelidar a primeira de *Synthesised Power* (sp) e a segunda de *Voiced Power* (vp), tem-se:

$$S_{m2}(w) \quad = \quad S_{m1}(w) \times \sqrt{vp/sp}. \tag{6.2}$$

Onde $S_{m1}(w)$ era a trama sintetizada antes do ajuste, e $S_{m2}(w)$ é uma nova trama cuja energia é igual à da trama do sinal vozeado original correspondente.

A Figura 6.9 ilustra o módulo do sinal reconstruido após ajuste de energia na primeira trama de "A"em "A Sofia...".

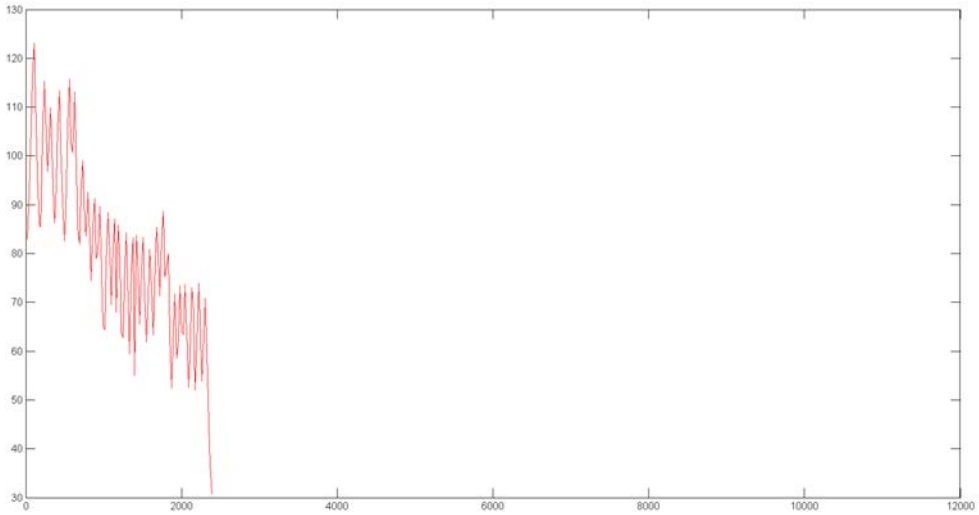

Figura 6.9: Módulo do sinal reconstruido após ajuste de energia na primeira trama de "A"em "A Sofia...".

Na versão independente, não se conhecendo o sinal vozeado original, teve de se usar uma abordagem diferente. A única declaração que a literatura relata, tanto quanto a pesquisa pôde apurar, é que as tramas correspondentes a vogais possuem, em média, mais 20dB de potência na fala vozeada do que na sua contraparte não vozeada. No entanto, não tendo sido feita uma distinção entre vogais e consoantes aquando da segmentação da frase, todas as tramas vozeadas artificialmente terão de ser sujeitas ao mesmo aumento de potência. Denomine-se a potência média de uma trama sussurrada original de *Whispered Power* (wp) e o incremento de potência em dB face à potência da trama sussurrada original de *power increment* (pi). Faz-se:

$$S_{m2}(w) \quad = \quad S_{m1}(w) \times \sqrt{wp/sp} \times \sqrt{10^{pi/10}}. \tag{6.3}$$

Começou-se por experimentar com $pi = 20$ mas tal valor provou ser demasiado alto, uma vez que o sinal gravado no ficheiro sofreu *clipping* em grande parte da sua duração. O valor de *pi* foi ajustado iterativamente até o máximo absoluto do sinal ser inferior ao nível máximo permitido na gravação do sinal no ficheiro. Os valores de *pi* finais foram 10,6dB no caso do orador e 12dB no caso da oradora. Note-se, no entanto, que esta diferença de valores não deve ser interpretada com significância, visto que o sinal de sussurro original da oradora não se encontra tão perto dos limites de amplitude como o do orador, por uma questão de insuficiência do amplificador. Tal diferença leva, naturalmente, à necessidade de um maior incremento de energia no caso da oradora, para que

o sinal resultante se situe na mesma gama de valores que o sinal resultante do orador.

6.4.4 Gravação

Serve esta secção para explicitar o modo como foi gravado o sinal final num ficheiro.

O formato de saída foi igualmente PCM e tanto a F_s como o número de bits de resolução permaneceram, naturalmente, idênticos: 22050Hz e 16 bits. O método de reconstrução do sinal foi baseado no procedimento *overlap-add*. Em detalhe, por cada trama a ser gravada, se esta fizesse parte de uma região a vozear, então escrevia-se a trama vozeada artificialmente correspondente; de outra forma, escrevia-se a trama correspondente do sinal de sussurro original.

O anexo C contém o conjunto total dos sinais sintetizados, e seus espectrogramas: as versões dependente e independente, do orador e da oradora.

6.5 Resultados

Uma vez gravadas as duas versões, falta esclarecer que conclusões se podem tirar desta experiência. Faz sentido, neste ponto, dividir as conclusões em dois grupos: um grupo de conclusões complementares aos objetivos da dissertação, que surgem de observações realizadas durante o desenvolvimento do algoritmo; e um outro grupo de conclusões referente aos objetivos da dissertação propriamente ditos.

6.5.1 Resultados complementares

Durante o período de desenvolvimento do algoritmo descrito, a experimentação levou à observação de curiosos comportamentos dos sinais sintetizados. Enumeram-se seguidamente as principais observações registadas.

O escalonamento do espectro, realizado com o objetivo de alterar o valor das frequências formantes, não pôde ser igual nas amostras sussurradas de ambos os oradores. De facto, começou por se ajustar o fator de escalonamento para 1.07, uma vez que este providenciava o melhor resultado. Seguidamente, testou-se o mesmo valor para a amostra da oradora e verificou-se um resultado medíocre em comparação ao caso anterior. Usou-se então o valor 1.05 para se obter um resultado melhor. Tal realização vai de encontro ao que está descrito na secção 2.3, quando se refere que o deslocamento das frequências formantes ocorre em maior escala nos oradores do sexo masculino.

Se se assumir que, tal como foi dito na secção 2.3, existe uma discrepância de 20dB entre vogais na fala vozeada e na fala sussurrada, e sabendo que os valores usados neste algoritmo rondaram os 11dB, pode-se afirmar que as vogais precisam de um ajuste de energia maior do que as consoantes vozeadas. O que por sua vez significa que as vogais perdem muito mais energia na ausência de vibração das pregas vocais, do que as consoantes vozeadas. Esta conclusão vai de encontro ao que é dito em [19], quando se afirma que, na fala vozeada, as vogais são produzidas quando existe fonação, sem turbulência. Já as consoantes (vozeadas) são produzidas quando existe fonação e turbulência. Assim sendo, parece lógico que, se se retirar a fonação nos dois grupos de

sons, o grupo das consoantes vozeadas retenha uma percentagem maior da sua energia original que o grupos das vogais.

A terceira, e última, observação registada já foi referida com algum detalhe anteriormente. No entanto, reforça-se aqui a ideia de que a fase do impulso glótico se manifestou irrelevante na síntese do sinal. Pelo menos, tanto quanto se pôde apurar através da audição desta frase protocolar.

6.5.2 Resultados principais

O objetivo desta dissertação era investigar, até que ponto, é possível obter um sinal vozeado artificialmente de boa qualidade. Mais, investigar em que medida este sinal artificial possui uma maior qualidade quando gerado com o auxílio de um sinal vozeado original.

No sentido de avaliar a qualidade dos sinais obtidos, optou-se por usar um método subjetivo. Isto é, ao invés de se usar uma fórmula matemática que tenta avaliar a qualidade do resultado, explorou-se a oportunidade de contar com um conjunto de ouvintes, cuja opinião média serve como indicador da qualidade dos sinais produzidos.

O procedimento relativo à avaliação subjetiva assegurou que os resultados obtidos seriam, por um lado, válidos pelas condições a que os ouvintes foram sujeitos durante a experiência, e por outro, interessantes no sentido de avaliar corretamente a qualidade dos sinais produzidos.

Contou-se com a participação de 35 pessoas, 19 do sexo masculino e 16 do sexo feminino com uma idade média de 23,5 anos. A audição foi feita numa sala silenciosa com um *headset* de alta qualidade, Steelseries Siberia V2 [15] visível na Figura 6.10.

Figura 6.10: Ilustração do headset Steelseries Siberia V2 [15] usado no teste subjetivo.

Foi requisitado aos ouvintes que ouvissem os ficheiros originais e os sintetizados numa certa ordem, e que, para cada ficheiro sintetizado, atribuíssem uma pontuação a uma série de categorias idealizadas no sentido de avaliar os sinais produzidos em diferentes vertentes. A pontuação seguiu uma escala de Likert de 5 níveis (5 para a melhor situação e 1 para a pior); as diferentes categorias são:

- Degradação - a quantidade de degradação da qualidade geral percebida da amostra vozeada original para a amostra sintetizada;
- Inteligibilidade - quão bem se percebeu o que foi dito;
- Naturalidade - quão natural (=humana) a voz soou;
- Identidade - quão bem é que se tornava possível identificar quem estava a falar.

Facilmente se repara que, tanto a inteligibilidade como a identidade, são categorias que só fazem sentido ser pontuadas havendo uma comparação entre a amostra original e a sintetizada. Já a inteligibilidade e a naturalidade apenas requerem a audição do sinal sintetizado para serem classificadas. Abaixo encontra-se o conjunto de passos seguido pelos ouvintes no processo de audição, o qual respeita as obrigatoriedades mencionadas:

1. Ouvir a amostra vozeada original do orador masculino.
2. Ouvir a amostra vozeada artificial dependente do orador masculino.
3. Pontuar a degradação percebida.
4. Ouvir a amostra vozeada artificial dependente do orador masculino.
5. Pontuar a inteligibilidade percebida.
6. Ouvir a amostra vozeada artificial dependente do orador masculino.
7. Pontuar a naturalidade percebida.
8. Ouvir a amostra vozeada original do orador masculino.
9. Ouvir a amostra vozeada artificial dependente do orador masculino.
10. Pontuar a identidade percebida.
11. Repetir para os outros três ficheiros.

Existe um outro fator envolvido na elaboração do guião acima: a norma de telecomunicações P.900 [25], a qual especifica como devem ser feitos testes subjetivos de codecs som. Esta norma descreve, entre outros, dois tipos de testes válidos: o "*Absolute Category Rating*"(ACR) e o "*Degradation Category Rating*"(DCR). O teste ACR consiste na audição da amostra sintetizada uma única vez, sem anterior audição da amostra original, e na pontuação segundo uma escala de Likert de 5 níveis da qualidade geral percebida; níveis estes que possuem um significado bem definido na norma. O teste DCR é em tudo semelhante ao anterior, exceto o facto de se ouvir, anteriormente à amostra sintetizada, a amostra original.

No âmbito desta dissertação, adotaram-se os procedimentos e adaptaram-se as escalas para servirem as categorias escolhidas. Os três primeiros passos do guião representam um teste DCR,

tal e qual como se encontra na norma, os passos quatro a cinco e seis a sete representam 2 testes ACR e os passos oito a dez representam um teste DCR, todos eles com escalas adaptadas da original. Podem ser encontrados no anexo D, para cada uma destas categorias, os significados de cada nível de Likert.

Uma vez obtidas as pontuações dos ouvintes, dividem-se os resultados nas quatro categorias definidas. A primeira secção diz respeito à degradação - nesta é explicado e justificado um tratamento estatístico realizado sobre os dados registados. É também aqui que se tece um conjunto de conclusões relativas à degradação, derivadas da análise estatística. O conteúdo das secções posteriores é semelhante, embora dizendo respeito a cada uma das categorias restantes.

6.5.2.1 Degradação

Existem quatro grupos de dados objetos de análise no que diz respeito à degradação, a saber:

- as opiniões dos ouvintes relativas ao ficheiro pertencente ao orador do sexo Masculino, Dependente da referência vozeada e pontuado na categoria Degradação - abreviado MDDeg;
- as opiniões dos ouvintes relativas ao ficheiro pertencente ao orador do sexo Masculino, Independente da referência vozeada e pontuado na categoria Degradação - abreviado MIDeg;
- as opiniões dos ouvintes relativas ao ficheiro pertencente ao orador do sexo Feminino, Dependente da referência vozeada e pontuado na categoria Degradação - abreviado FDDeg;
- as opiniões dos ouvintes relativas ao ficheiro pertencente ao orador do sexo Feminino, Independente da referência vozeada e pontuado na categoria Degradação - abreviado FIDeg.

O tratamento estatístico dos dados realizou-se em duas partes. Inicialmente, fez-se uma análise descritiva de cada um dos grupos mencionados, no sentido de se obter uma visão geral das distribuições. Posteriormente, combinaram-se os grupos de dados dois a dois, das quatro formas possíveis. Cada par foi sujeito a uma análise inferencial compreendida por um teste de Wilcoxon e um teste de Friedman, ambos com um nível de significância de 5%. O teste de Wilcoxon serve para comprovar estatisticamente se as medianas dos dois grupos são iguais, isto é, se os dois grupos podiam ser duas amostras de um só grupo (com uma só mediana). Já o teste de Friedman tenta comprovar estatisticamente que as distribuições dos dois grupos são iguais, ou seja, que os dois grupos podiam ser duas amostras de um só grupo (com uma só distribuição). Considera-se que dois grupos de dados não apresentam diferenças estatisticamente relevantes se se comprovar que as suas medianas e as suas distribuições são simultaneamente iguais. Ou em termos práticos, se os resultados de ambos os testes forem superiores a 0,05.

A tabela 6.1 apresenta a análise descritiva e inferencial para o par: MDDeg e MIDeg.

Da observação da tabela, constata-se que a média de MDDeg é visivelmente superior à de MIDeg, e que os valores de mediana e dos quartis sugerem que, de uma forma geral, a distribuição de MDDeg compreende valores superiores à de MIDeg. O resultado da análise inferencial confirma essa disparidade. De facto, os resultados de 0,004, em ambos os testes, sendo inferiores

Tabela 6.1: Análise descritiva e inferencial dos grupos de dados referentes às versões dependente e independente do orador do sexo masculino no estudo da degradação.

Parâmetro	**Análise descritiva**					**Teste Wilcoxon**	**Teste Friedman**
	Média	**Moda**	**Q25**	**Mediana**	**Q75**		
MDDeg	2,77	3	2	3	3	0,004	0,004
MIDeg	2,34	3	2	2	3		

a 0,05, não são suficientemente altos para que se possa afirmar que os dois grupos de dados podiam ser originários do mesmo grupo; o que significa que a versão dependente do orador do sexo masculino e a versão independente do mesmo orador apresentam, face à amostra original, níveis de degradação de qualidade estatisticamente diferentes, com a versão dependente a aproximar-se mais da referência vozeada do que a sua contraparte. Tal resultado era expectável dada a ausência de informação da referência vozeada na síntese da versão independente e face à presença dessa mesma informação na geração da versão dependente.

Comparem-se agora, na tabela 6.2, as versões dependente e independente referentes à oradora.

Tabela 6.2: Análise descritiva e inferencial dos grupos de dados referentes às versões dependente e independente do orador do sexo feminino no estudo da degradação.

Parâmetro	**Análise descritiva**					**Teste Wilcoxon**	**Teste Friedman**
	Média	**Moda**	**Q25**	**Mediana**	**Q75**		
FDDeg	3,20	3	3	3	4	0,000	0,000
FIDeg	2,23	2	2	2	3		

A análise descritiva permite observar uma discrepância de degradação entre versões, superior no caso feminino em relação ao caso masculino. O resultado dos testes de Wilcoxon e Friedman, agora ainda mais que anteriormente, revelam uma diferença de valores entre os dois grupos de dados, reprovando, igualmente, a hipótese de ambas as versões apresentarem uma fidelidade à referência vozeada semelhante.

As duas próximas tabelas, 6.3 e 6.4, não pretendem comparar diferenças entre versões dependente e independente, mas sim entre versões com o mesmo grau de dependência para os dois oradores. Embora esta comparação não faça, estritamente falando, parte do objetivo da dissertação, a sua realização ajuda a caracterizar a qualidade dos sinais sintetizados.

A tabela 6.3 apresenta as diferenças existentes entre oradores na avaliação dos sinais dependentes.

Tabela 6.3: Análise descritiva e inferencial dos grupos de dados referentes à versão dependente de ambos os oradores no estudo da degradação.

Parâmetro	**Análise descritiva**					**Teste Wilcoxon**	**Teste Friedman**
	Média	**Moda**	**Q25**	**Mediana**	**Q75**		
MDDeg	2,77	3	2	3	3	0,007	0,012
FDDeg	3,20	3	3	3	4		

É visível, tanto da análise descritiva como da análise inferencial, que, apesar de apresentarem valores próximos, estes dois grupos são ainda estatisticamente diferentes, o que significa que, a versão dependente relativa à fala da oradora, se encontra mais próxima da respetiva referência vozeada, do que a versão dependente se encontra próxima da referência análoga. Ou seja, recorrendo à amostra vozeada original, conseguem-se resultados superiores na fala de uma oradora em relação à fala de um orador.

Por seu turno, a tabela 6.4 apresenta as diferenças existentes entre oradores na avaliação dos sinais independentes.

Tabela 6.4: Análise descritiva e inferencial dos grupos de dados referentes à versão independente de ambos os oradores no estudo da degradação.

Parâmetro	**Análise descritiva**					**Teste Wilcoxon**	**Teste Friedman**
	Média	**Moda**	**Q25**	**Mediana**	**Q75**		
MIDeg	2,34	3	2	2	3	0,467	0,593
FIDeg	2,23	2	2	2	3		

Contrariamente ao caso anterior, na comparação de géneros dos sinais independentes, os dois grupos de dados são suficientemente semelhantes para passarem em ambos os testes da análise inferencial. O que significa que, quando não se extrai informação da amostra vozeada, a qualidade do sinal sintetizado é independente do género do orador.

Dos três elementos que se extraem da amostra vozeada: curva de F_0, valores exatos dos harmónicos, energia da trama, é seguro afirmar-se que a curva de F_0 é o elemento que mais contribui para a amostra sintetizada soar semelhante à amostra vozeada original. Assim sendo, é provável que a dissemelhança entre as duas últimas tabelas signifique que a curva de F_0 é um elemento mais importante na fala vozeada da oradora do que na do orador.

Estas considerações disseram respeito à avaliação da degradação de qualidade entre os sinais originais e sintetizados. De uma forma geral, os ouvintes quantificaram a degradação das amostras dependentes em cerca de 3; as amostras independentes, por sua vez, foram quantificadas próximas de 2,3. Qualitativamente falando, 2 significa que a degradação percebida foi "incomodativa"e 3 significa que esta foi "um pouco incomodativa". Em suma, a versão independente não apresenta a qualidade necessária para competir com a dependente.

A próxima secção retrata os resultados do ponto de vista da inteligibilidade.

6.5.2.2 Inteligibilidade

Inteligibilidade define-se como "quão bem se percebe o que é dito".

De forma semelhante à secção anterior, esta secção exprime as opiniões dos ouvintes em quatro tabelas, cuja observação serve de base para a elaboração de um conjunto de conclusões. A tabela 6.5 apresenta a comparação entre as versões dependente e independente, no que diz respeito ao orador do sexo masculino.

Tabela 6.5: Análise descritiva e inferencial dos grupos de dados referentes às versões dependente e independente do orador do sexo masculino no estudo da inteligibilidade.

Parâmetro	**Análise descritiva**					**Teste Wilcoxon**	**Teste Friedman**
	Média	**Moda**	**Q25**	**Mediana**	**Q75**		
MDInt	4,74	5	5	5	5	0,527	0,257
MIInt	4,69	5	4	5	5		

A análise inferencial confirma o que é evidente na análise descritiva. A inteligibilidade apresenta valores altíssimos e muitíssimo semelhantes em ambas versões. Aparentemente, a dificuldade de perceção do conteúdo sonoro foi quase inexistente, no caso do orador.

A tabela 6.6 retrata os resultados no caso da oradora.

Tabela 6.6: Análise descritiva e inferencial dos grupos de dados referentes às versões dependente e independente do orador do sexo feminino no estudo da inteligibilidade.

Parâmetro	**Análise descritiva**					**Teste Wilcoxon**	**Teste Friedman**
	Média	**Moda**	**Q25**	**Mediana**	**Q75**		
FDInt	4,94	5	5	5	5	0,011	0,008
FIInt	4,71	5	4	5	5		

Ainda com uma ligeira semelhança, os dois conjuntos são agora estatisticamente diferentes. Por outras palavras, a extração de parâmetros da amostra vozeada original contribui significativamente, no caso da oradora feminina, para o aumento da inteligibilidade. Uma vez que a inteligibilidade (tal como a naturalidade e a identidade) pode ser vista como uma área particular da qualidade geral do sinal, a dissemelhança aqui observada é coerente com a dissemelhança observada no estudo da degradação.

A tabela 6.7 avalia a diferença entre géneros na pontuação da versão dependente.

Tabela 6.7: Análise descritiva e inferencial dos grupos de dados referentes à versão dependente de ambos os oradores no estudo da inteligibilidade.

Parâmetro	**Análise descritiva**					**Teste Wilcoxon**	**Teste Friedman**
	Média	**Moda**	**Q25**	**Mediana**	**Q75**		
MDInt	4,74	5	5	5	5	0,020	0,014
FDInt	4,94	5	5	5	5		

Constata-se que a versão dependente da oradora apresenta resultados superiores de inteligibilidade, face à do orador. Ou seja, é mais impactante a extração de informação da amostra vozeada no caso da oradora do que no caso do orador.

Veja-se, na tabela 6.8 o que acontece na versão independente.

Tabela 6.8: Análise descritiva e inferencial dos grupos de dados referentes à versão independente de ambos os oradores no estudo da inteligibilidade.

Parâmetro	Análise descritiva					Teste Wilcoxon	Teste Friedman
	Média	Moda	Q25	Mediana	Q75		
MIInt	4,69	5	4	5	5	0,782	0,527
FIInt	4,71	5	4	5	5		

Quando não se extraem parâmetros das amostras vozeadas originais, a inteligibilidade dos sinais artificialmente vozeados dos dois oradores é muito semelhante.

Este estudo disse respeito à avaliação da inteligibilidade dos sinais sintetizados. De uma forma geral, os ouvintes pontuaram a inteligibilidade das amostras dependentes com 4,8; as amostras independentes, por sua vez, foram pontuadas com 4,7. Qualitativamente falando, 4 significa que o ouvinte "ou não percebeu uma palavra ou percebeu tudo mas com esforço"e 5, que o ouvinte "percebeu o que foi dito sem problema". Provou-se que, para um orador masculino, se percebe tão bem o que foi dito na versão dependente como na independente; com um orador feminino, percebe-se significativamente melhor o que foi dito na versão dependente; no entanto, esta diferença é para benefício da versão dependente e não para malefício da independente.

Na próxima secção apresentam-se e discutem-se os resultados da naturalidade.

6.5.2.3 Naturalidade

A naturalidade pode ser entendida como a proximidade da voz a uma voz humana e saudável.

Observem-se os resultados das pontuações nas quatro tabelas habituais.

Tabela 6.9: Análise descritiva e inferencial dos grupos de dados referentes às versões dependente e independente do orador do sexo masculino no estudo da naturalidade.

Parâmetro	Análise descritiva					Teste Wilcoxon	Teste Friedman
	Média	Moda	Q25	Mediana	Q75		
MDNat	2,91	3	2	3	3	0,000	0,000
MINat	1,77	2	1	2	2		

A tabela 6.9 prova que a extração de parâmetros da referência vozeada é muito importante para a naturalidade, quando se trata de um orador.

Tabela 6.10: Análise descritiva e inferencial dos grupos de dados referentes às versões dependente e independente do orador do sexo feminino no estudo da naturalidade.

Parâmetro	Análise descritiva					Teste Wilcoxon	Teste Friedman
	Média	Moda	Q25	Mediana	Q75		
FDNat	3,17	3	3	3	4	0,000	0,000
FINat	1,57	1	2	2	2		

Por sua vez, a tabela 6.10 descreve a disparidade de pontuações no caso feminino como sendo ainda maior que a do caso anterior.

Em ambas as tabelas observa-se que o uso de elementos da referência vozeada contribui, largamente, para a melhoria da naturalidade da voz sintetizada. Neste caso, a curva de F_0 é indubitavelmente um dos principais fatores de deterioramento, mas não é o único. Existe, na literatura [19], informação de que a vibração das pregas vocais não é exatamente periódica, mas sim quasi-periódica; tal facto provoca o aparecimento de *jitter* no valor dos harmónicos até um 1Hz e de *shimmer* até 1dB. O facto de a componente harmónica da versão independente ser composta por harmónicos sem este tipo de oscilação é, sem dúvida, um ponto responsável pela disparidade de naturalidade entre versões.

Comparem-se agora os dois géneros.

Tabela 6.11: Análise descritiva e inferencial dos grupos de dados referentes à versão dependente de ambos os oradores no estudo da naturalidade.

Parâmetro	**Análise descritiva**					**Teste Wilcoxon**	**Teste Friedman**
	Média	**Moda**	**Q25**	**Mediana**	**Q75**		
MDNat	2,91	3	2	3	3	0,061	0,074
FDNat	3,17	3	3	3	4		

A tabela 6.11 mostra que não existe diferença estatística entre géneros na versão dependente, ainda que estes grupos de dados passem diminutamente nos testes da análise inferencial.

Tabela 6.12: Análise descritiva e inferencial dos grupos de dados referentes à versão independente de ambos os oradores no estudo da naturalidade.

Parâmetro	**Análise descritiva**					**Teste Wilcoxon**	**Teste Friedman**
	Média	**Moda**	**Q25**	**Mediana**	**Q75**		
MINat	1,77	2	1	2	2	0,251	0,317
FINat	1,57	1	2	2	2		

A tabela 6.12 afirma que, com uma maior certeza no caso da versão independente, não existe uma diferença entre géneros no que respeita a inteligibilidade.

Ainda que ambas as versões não apresentem diferença significativa entre géneros, é compreensível que a diferença existente seja maior quando se constrói um sinal com uma componente harmónica variável no tempo, como é o caso da versão dependente.

Este estudo disse respeito à avaliação da naturalidade dos sinais sintetizados. De uma forma geral, os ouvintes atribuíram, à naturalidade das amostras dependentes, o nível 3; já às amostras independentes atribuíram, em média, 1,7. Qualitativamente falando, 1 significa que a voz escutada soou "completamente robótica", 2 significa que "deu para perceber que era um humano a falar, mas por pouco"e 3 significa que a voz ficou "a meio caminho entre humano e robô". Concluiu-se, como se esperava, que a naturalidade é um ponto de grande dificuldade. Sabe-se que a informação de

altas frequências é o principal contribuinte para o aumento da naturalidade; neste caso, ambas as versões não possuem informação harmónica para além dos 7kHz, algo que faz com que os valores de naturalidade sejam globalmente baixos. Adicionalmente, a versão independente sofre um decréscimo acentuado, uma vez que possui uma curva de F_0 plana e harmónicos sem qualquer tipo de oscilação.

Na próxima secção, estudam-se os resultados relativos à última categoria: a identidade.

6.5.2.4 Identidade

A identidade avalia o quanto se mantiveram as características acústicas de um determinado orador, após modificação de um sinal. Neste caso, o quanto a voz do sinal vozeado artificialmente se assemelha à voz da referência vozeada.

Uma vez mais, as tabelas seguintes expõem as opiniões dos ouvintes.

Tabela 6.13: Análise descritiva e inferencial dos grupos de dados referentes às versões dependente e independente do orador do sexo masculino no estudo da identidade.

Parâmetro	**Análise descritiva**					**Teste Wilcoxon**	**Teste Friedman**
	Média	**Moda**	**Q25**	**Mediana**	**Q75**		
MDIde	3,57	4	3	4	4	0,000	0,000
MIIde	2,51	3	2	3	3		

A tabela 6.13 comprova uma perda de identidade entre versões, quando se trata do orador.

Tabela 6.14: Análise descritiva e inferencial dos grupos de dados referentes às versões dependente e independente do orador do sexo feminino no estudo da identidade.

Parâmetro	**Análise descritiva**					**Teste Wilcoxon**	**Teste Friedman**
	Média	**Moda**	**Q25**	**Mediana**	**Q75**		
FDIde	3,57	4	3	4	4	0,000	0,000
FIIde	2,37	2	2	2	3		

A tabela 6.14, de forma semelhante à anterior, confirma a perda de identidade nas versões que concernem a oradora.

Em conjunto, as duas tabelas levam a crer que, também na perceção da identidade, os parâmetros extraídos da referência vozeada contêm uma boa parte da informação do orador.

Estude-se a semelhança entre os dois géneros no que diz respeito a esta categoria.

Tabela 6.15: Análise descritiva e inferencial dos grupos de dados referentes à versão dependente de ambos os oradores no estudo da identidade.

Parâmetro	**Análise descritiva**					**Teste Wilcoxon**	**Teste Friedman**
	Média	**Moda**	**Q25**	**Mediana**	**Q75**		
MDIde	3,57	4	3	4	4	0,861	0,655
FDIde	3,57	4	3	4	4		

A tabela 6.15 revela que a versão dependente possui a mesma informação de identidade independentemente do género do orador.

Tabela 6.16: Análise descritiva e inferencial dos grupos de dados referentes à versão independente de ambos os oradores no estudo da identidade.

Parâmetro	**Análise descritiva**					**Teste Wilcoxon**	**Teste Friedman**
	Média	**Moda**	**Q25**	**Mediana**	**Q75**		
MIIde	2,51	3	2	3	3	0,302	0,225
FIIde	2,37	2	2	2	3		

O mesmo se verifica na tabela 6.16 quando se avalia a versão independente para os dois géneros.

Com isto, é possível afirmar que os valores de identidade apenas são influenciados pela extração, ou não, de parâmetros da amostra vozeada original.

As avaliações aqui redigidas exploraram a identidade dos sinais sintetizados. De um modo geral, os ouvintes concederam à identidade das amostras dependentes um nível médio de 3,6 e, às amostras independentes, concederam 2,4. Qualitativamente falando, 2 significa "podia ser qualquer pessoa do mesmo género do orador da amostra original", 3 significa que o ouvinte "não conseguiu perceber quem era mas, se lhe dissessem quem era a falar, ele concordaria"e 4 significa que o ouvinte crê que "era provavelmente o orador em questão a falar". Pode-se afirmar que a informação da identidade do orador se encontra presente, tanto no trato vocal, como nas pregas vocais. A diferença de um nível de Likert entre versões pode ser atribuído, com pesos a determinar, ao conjunto de componentes extraído da referência vozeada. Mais ainda, a diferença de géneros em nada afeta a perceção de identidade no processo de vozeamento artificial.

6.5.2.5 Conclusão

Da análise dos resultados feita acima, pode-se sumarizar um conjunto de conclusões.

Como se demonstrou, a categoria degradação apresentou valores intermédios da escala de Likert. A disparidade entre amostras foi, em média, de 0,7 níveis, sendo que a degradação percebida na audição de amostras dependentes foi, em média, um pouco incomodativa; já a das amostras independentes foi incomodativa.

A inteligibilidade foi a categoria que apresentou os melhores resultados em ambas as versões do sinal. Foi também a categoria cuja disparidade (de 0,1 níveis) foi menor, significando que, a grande maioria das pessoas percebeu o que foi dito sem qualquer esforço.

A naturalidade deteve as pontuações mais baixas do conjunto de categorias. Foi também a categoria onde se observou a maior disparidade de pontuação entre amostras, chegando aos 1,3 níveis. Os ouvintes consideraram que a voz da amostra dependente se encontrava a meio caminho entre uma voz humana e uma voz robotizada, e que a amostra independente soava completamente robótica.

A identidade foi um domínio que, em semelhança à degradação, exibiu resultados intermédios; no entanto, a disparidade de valores entre as duas versões foi um pouco maior neste domínio, alcançando os 1,2 níveis. Da audição da amostra dependente, cerca de metade dos voluntários respondeu que não percebiam quem estava a falar, mas que concordariam se lhes dissessem quem era; a outra metade respondeu que achavam provavelmente ser o orador em questão a falar. Da audição da amostra independente, novamente cerca de metade dos ouvintes opinou que poderia ter sido qualquer pessoa a falar, desde que esta fosse do mesmo género do orador original, e a outra metade respondeu, mais uma vez, que não conseguiam perceber quem estava a falar, mas que concordariam se lhes fosse dito quem era.

Provou-se, através da análise inferencial, que só no domínio da inteligibilidade se podem confundir as versões dependente e independente. Em todos os outros domínios, serão precisos esforços adicionais no desenvolvimento da versão independente, para que se consigam alcançar resultados tão favoráveis como os da versão dependente. Crê-se que, se se conseguir fazer variar a curva de F_0 no sentido de imitar a curva de F_0 presente na amostra vozeada, aumentar-se-iam significativamente as pontuações das três categorias que reprovaram a hipótese nula da análise inferencial. Adicionalmente, pensa-se que a introdução de *jitter* e *shimmer* na componente harmónica, contribuiria também, embora em menor escala, para o aumento das mesmas, embora a maior diferença se situasse provavelmente na categoria da naturalidade. Com o intuito de se aumentarem os resultados de ambas as versões, poder-se-ia, por um lado, empregar uma técnica de deslocamento de formantes mais precisa, e, por outro, construir uma componente harmónica com um número mais elevado de harmónicos; prevê-se que ambas as técnicas contribuíssem para o aumento das três categorias medíocres, com a primeira a influenciar principalmente a identidade e a segunda a naturalidade.

Capítulo 7

Segmentação automática do sinal sussurrado

Neste capítulo, vai-se um pouco mais longe do que os objetivos desta dissertação investigando-se a possibilidade de se detectar automaticamente, nos sinais sussurrados, quais as zonas que necessitam de vozeamento, e quais as que não necessitam. Tal deteção é, não só útil, como vital, quando se pretende tornar a geração da versão independente um processo completamente automático, o que, por sua vez permitiria o uso deste algoritmo como *software* num aparelho a usar pelos pacientes laringectomizados.

De início, investigou-se a informação que havia na literatura, no sentido de se perceber o que já havia sido explorado em termos desta deteção automática. Foram descobertas e implementadas duas estratégias de deteção automática de regiões do sinal sussurrado. Na secção seguinte, explicam-se brevemente estas abordagens e conclui-se sobre até que ponto foram bem sucedidas.

Na secção posterior, descreve-se uma metodologia desenvolvida de raiz, no sentido de criar uma estratégia de deteção automática mais poderosa e de desempenho razoavelmente rápido. Esta metodologia tem como núcleo o treino e teste de algoritmos de aprendizagem de máquina supervisionada.

Em ambas as secções, utilizam-se as versões originais dos sinais gravados de características: 48kHz de frequência de amostragem e 32 bits de resolução. Mais ainda, a dimensão das tramas passou a ser de 480 amostras (10 milissegundos) e a sobreposição de 479 amostras; duas alterações realizadas com o intuito de aumentar o número de tramas, e assim, aumentar a eficácia da classificação.

7.1 Segmentação sem um algoritmo de aprendizagem de máquina

Foram encontradas duas metodologias de classificação automática de tramas na literatura que se enquadram, até certo ponto, no âmbito desta dissertação. Essas metodologias foram implementadas em Matlab e o seu desempenho foi avaliado. No teste destes dois algoritmos apenas se usou a

amostra sussurrada correspondente ao orador do sexo masculino, uma vez que, nada se alcançaria em testar as duas amostras, se para uma delas o resultado não fosse aceitável.

Cada uma das secções integrantes desta secção retrata sucintamente uma destas metodologias.

7.1.1 Deteção de silêncios e classificação de sons

Tal como o nome desta secção sugere, a classificação de tramas, segundo este método [3], é feita em duas etapas. Na primeira etapa, diferenciam-se tramas correspondentes a períodos de silêncio e tramas correspondentes a fala sussurrada, descartando-se as primeiras. Na segunda etapa, tenta-se classificar as tramas restantes como "fricativa", "plosiva não vozeada"ou "vogal". Não é feita uma distinção entre plosivas vozeadas e não vozeadas; adicionalmente, não há menção dos tipos de consoantes vibrante e lateral.

7.1.1.1 Deteção de silêncios

Na primeira etapa, denominada de *Whisper Activity Detector* (WAD), é calculada a energia e a taxa de passagens por zero de cada trama, sendo que a última se define como o número de passagens do sinal por zero num determinado período de tempo.

Seguidamente, uma trama é classificada como sussurro ou silêncio, de acordo com a seguinte função:

$$WAD_m = \begin{cases} 1 & E_m \geq \xi_2 \\ -1 & E_m < \xi_1 \\ sign(Z_m - \xi_z) & \xi_1 \leq E_m < \xi_2 \end{cases}$$

Onde E_m e Z_m são, respetivamente, a energia e a taxa de passagens por zero da trama m; $WAD_m = 1$ significa que a trama m pertence a um período de sussurro; $WAD_m = -1$ significa que a trama faz parte de um período de silêncio; ξ_1, ξ_2 e ξ_z são constantes (diferentes para cada orador) desconhecidas à partida; e sign(.) é uma função que vale 1 ou -1, se o seu argumento for positivo ou negativo, respetivamente.

No âmbito desta dissertação, calcularam-se estas constantes de acordo com uma heurística. Neste cálculo, faz-se uso de uma classificação manual das tramas no auxílio da classificação automática, no sentido de se tentar alcançar o que seria a solução ótima do valor destas constantes. Os passos são os seguintes:

- realiza-se uma segmentação manual das tramas em sussurro e silêncio;
- ξ_1 assume o valor da energia da trama sussurro com menor energia;
- ξ_2 assume o valor da energia da trama silêncio com maior energia;
- ξ_z assume o valor de taxa de passagens por zero que produzir os melhores resultados de classificação.

Testado o algoritmo, o sucesso da classificação automática (quando comparada com a manual), foi muito boa: quase todos os silêncios detetados e poucas tramas de sussurro identificadas como silêncio. No entanto, é de relembrar que o valor das constantes teria de ser ajustado sem recurso à classificação manual, se este método viesse a ser implementado.

7.1.1.2 Classificação de sons

A segunda etapa, dada pelo nome de "*Whispered Phoneme Classifier*"(WPC), apenas tem em conta as tramas classificadas como sussurro da etapa anterior. O objetivo nesta fase é classificar as tramas sussurro em "fricativa", "plosiva não vozeada"ou "vogal". Para se avaliarem os resultados, foram também classificadas manualmente as tramas em "plosivas", "fricativas"(ambas não vozeadas) ou "a vozear". A classificação funciona de acordo com a seguinte árvore de decisão binária:

1. Primeiro, testa-se se a trama corresponde a uma fricativa, comparando a potência nas bandas de frequência acima e abaixo dos 3kHz;

2. Se não se tratar de uma fricativa, compara-se a energia presente na banda 1-3kHz com a banda 6-7.5kHz para se decidir entre vogal e plosiva.

Implementou-se, respetivamente ao primeiro passo, um cálculo de potência abaixo e acima dos 3kHz, semelhante ao cálculo da energia. Concretizado o cálculo, se acima dos 3kHz existisse mais potência, então, estar-se-ia na presença de uma fricativa. A taxa de sucesso na classificação de fricativas foi surpreendentemente alta, com escassas tramas classificadas incorretamente. Relativamente ao segundo passo, a comparação da energia nas bandas referidas não surtiu os resultados desejados. De facto, nenhuma trama plosiva, manualmente identificada, foi corretamente classificada.

Dada esta dificuldade em se classificar plosivas, houve a necessidade de se experimentar outras abordagens.

7.1.2 Deteção de inícios de consoantes não vozeadas

Este método, descrito em [26], tem um objetivo um pouco diferente do que se estava à procura. O que este algoritmo pretende conseguir é encontrar, em fala vozeada e em segmentos sonoros isolados Vogal-Consoante-Vogal (VCV), quais os segmentos correspondentes a consoantes plosivas não vozeadas e quais os correspondentes a fricativas não vozeadas. Apesar das suas diferenças, este algoritmo foi aplicado ao sinal sussurro completo e o seu desempenho foi avaliado.

O princípio é que, com a taxa de passagens por zero, o valor eficaz da energia e a derivada do valor eficaz da energia em ordem ao tempo, denominada de *Rate of Rise* (RoR), se consegue identificar qual a consoante não vozeada presente num determinado segmento VCV. As operações que se fazem sobre estes três são demasiado complexas para, no âmbito desta dissertação, serem expostas e explicadas. Assim sendo, passar-se-á à avaliação do desempenho do método.

Das 6 fricativas não vozeadas, 4 foram corretamente identificadas; e das 3 plosivas não vozeadas, 2 foram corretamente identificadas, não se tendo verificado nenhum falso positivo na identificação. Assim, o resultado da deteção das consoantes fricativas foi inferior ao método anterior; já o das plosivas foi muito superior. Existem, no entanto, 2 graves problemas com o método aqui discutido, intrínseco ao seu desenho: o de apenas detetar os inícios das consoantes, não fornecendo qualquer informação sobre a sua duração e o de não detetar períodos de silêncio.

Conseguir-se-ia solucionar o problema da deteção de silêncios e aumentar o sucesso na identificação de fricativas se se conseguisse, de alguma forma, combinar este método com o anterior. No entanto a duração das plosivas seria um problema por resolver. Tamanha complexidade suscitou a procura de uma solução mais prática e elegante: aprendizagem de máquina.

7.2 Segmentação com um algoritmo de aprendizagem de máquina

Aprendizagem de máquina é uma área científica que pretende responder à pergunta: "como desenvolver programas de computador que se melhorem a eles mesmos através de experiência?" [27]. O resultado é um conjunto de algoritmos que conseguem aumentar a sua capacidade de realizar uma tarefa pelo número de vezes que a realizam.

No âmbito desta dissertação, procura-se um algoritmo capaz de classificar uma trama como "necessita de vozeamento"ou "não necessita de vozeamento", após absorver experiência através de um conjunto de tramas classificadas manualmente. Ao processo de fornecimento de experiência ao algoritmo chama-se "treino", e, ao processo de classificação automática, dá-se o nome de "teste". A função classificadora gerada com o processo de treino apelida-se de "classificador".

Assim, é necessário distinguir um conjunto de tramas de treino e um conjunto de tramas de teste. No entanto, existe um número finito de tramas. E, se por um lado, no sentido de obter a classificação automática mais exata, pretende-se tornar o conjunto de tramas de treino o maior possível em prejuízo do conjunto de tramas de teste; por outro, deseja-se testar o classificador para o maior número de tramas possível, com o objetivo de se obter uma ideia mais correta do seu desempenho na presença de diferentes sons.

Uma técnica usada frequentemente para ultrapassar este impasse chama-se "validação cruzada estratificada", significando, para os objetivos deste capítulo, a divisão do conjunto total de tramas em conjuntos menores, onde as proporções de tramas que necessitam de vozeamento e as que não necessitam, permanecem semelhantes às proporções do conjunto total. A forma como se dividem estes conjuntos e como deles se obtêm conjuntos de treino e teste, depende da versão de validação cruzada estratificada usada. Nesta dissertação foi usada a comum versão de validação cruzada estratificada de 10 partições. Nesta versão, o conjunto total de tramas é dividido em 10 subconjuntos, onde 9 são o conjunto de treino e o restante o conjunto de teste. Terminado o primeiro processo de aprendizagem e teste, repete-se o processo 9 vezes até se esgotarem todos os agrupamentos possíveis de 9 subconjuntos como conjuntos de treino. No final, os 10 classificadores gerados são combinados num só (geralmente obtendo-se uma média dos seus parâmetros), o qual é considerado o classificador mais apto a ser usado para testar um futuro conjunto de tramas de

propriedades desconhecidas. A eficácia do classificador resultante é também uma combinação das eficácias dos 10.

Existe um importante aspeto a saber sobre o funcionamento geral de algoritmos de aprendizagem de máquina. Quando se fala em dotar um algoritmo de tramas de treino e de teste, isto não significa que o algoritmo receba todas as amostras de determinada trama; tal feito tornaria as tarefa de treino e de teste computacionalmente impossíveis. Ao invés, passa-se um conjunto de parâmetros calculados à custa do processamento das tramas. A escolha deste conjunto de parâmetros foi cuidadosamente pensada, e encontra-se descrita na próxima secção.

Tal como a escolha dos parâmetros a usar, a escolha do algoritmo é igualmente importante. Diferentes algoritmos apresentam diferentes vantagens e desvantagens, sendo necessário escolher o que melhor se adapta ao problema em causa. Na secção subsequente à da escolha dos parâmetros, descrevem-se e testam-se dois algoritmos e, em função dos resultados, escolhe-se o mais apropriado.

Uma vez obtida uma eficácia elevada, parte-se para outro objetivo: agilizar o procedimento. Só com um processo de classificação suficientemente rápido é que o mesmo se torna útil na aplicação prática discutida. Assim, faz-se um conjunto de três análises para se reduzirem o número de parâmetros usados na classificação, sacrificando o menos possível a eficácia. Estas análises encontram-se encadeadas na seguinte ordem:

1. Uma análise Kruskal Wallis para se reduzir o número de parâmetros calculados;
2. Uma análise de componente principais para se eliminar a redundância entre parâmetros;
3. Uma segunda análise Kruskal Wallis para se reduzir o número final de parâmetros que o classificador processa.

Nas penúltimas secções deste capítulo, explica-se com um pouco mais de detalhe o funcionamento destas análises, bem como os resultados que delas surtiram.

Termina-se o capítulo com uma conclusão sumária das observações assimiladas durante o desenvolvimento deste classificador automático.

7.2.1 Escolha de parâmetros

A escolha dos parâmetros foi feita com base nos seus poderes discriminativos na classificação de uma trama.

Uma vez que foram utilizados, nos métodos anteriores, a energia, a taxa de passagens por zero e o RoR, estes integraram automaticamente o conjunto de parâmetros. No entanto, estes três parâmetros, por si só, não satisfizeram uma taxa de deteção correta de tramas suficientemente elevada. Assim, acrescentaram-se treze *Mel-Frequency Cepstral Coefficients* (MFCC). Os MFCC são valores de potência medidos numa espécie de espectrograma cepstral numa escala de frequências não linear; entenda-se, uma série de espectros de logaritmos de espectros todos encaixados em sequência temporal, com a particularidade adicional de usarem uma escala de frequência cuja relação com a escala de frequência tradicional é dada por [28]:

$$m = 2595 \times log_{10}(1 + f/700). \quad (7.1)$$

Onde f é um valor da escala de frequência tradicional e m o valor correspondente na escala mel. Esta nova escala foi concebida para adaptar as variações de frequência ao sistema auditivo humano. Treze coeficientes MFCC significa que o "espectrograma cepstral"tem uma resolução vertical de treze pontos: um valor escolhido no compromisso entre preservação de informação e rapidez de computação, tanto no processo de obtenção dos MFCC, como no do tratamento dos mesmos pelo algoritmo de aprendizagem de máquina.

Possui-se agora toda a informação necessária para se projetar um algoritmo de aprendizagem de máquina. Definiu-se o conjunto de tramas a utilizar em treino e teste, as categorias possíveis "necessita de vozeamento"e "não necessita de vozeamento", e, por fim, o conjunto de parâmetros inicial fornecido ao algoritmo: 13 coeficientes mel cepstrais, energia, taxa de passagens por zero e RoR. Falta apenas escolher o algoritmo em si.

7.2.2 Escolha do algoritmo

Existem diferentes algoritmos de aprendizagem de máquina desenvolvidos para se adaptarem a uma larga gama de tipos de problemas. Dos existentes, testaram-se duas opções que, pelos seus princípios teóricos, se adequam ao problema em questão: máquinas de vetores de suporte e redes neuronais artificiais. Nas próximas secções descrevem-se os seus funcionamentos e classificam-se os seus desempenhos.

7.2.2.1 Máquinas de vetores de suporte

Uma Máquina de Vetores de Suporte, ou MVS, é um algoritmo de aprendizagem de máquina relativamente simples e de complexidade computacional baixa, principalmente no processo de teste. A explicação feita em seguida, no sentido de se tornar mais simples, diz respeito a uma máquina de vetores de suporte com 2 categorias de classificação possíveis, com 2 parâmetros de entrada e com uma divisão linear.

A algoritmia do processo de treino de uma máquina de vetores de suporte é, de uma forma resumida, a seguinte:

1. Considere-se um plano ortogonal com dois eixos. Cada um dos eixos contém os valores possíveis para um parâmetro de entrada do algoritmo;

2. Mapeie-se o conjunto de dados de treino segundo esses eixos;

3. Atribuam-se cores diferentes aos pontos, de acordo com as suas categorias;

4. Calcule-se a reta que separa, com o maior hiato possível, os pontos pertencentes a cada uma das categorias.

A Figura 7.1 ilustra este processo; note-se que a divisão feita pela reta H2 é a única correta, uma vez que garante o maior hiato entre categorias.

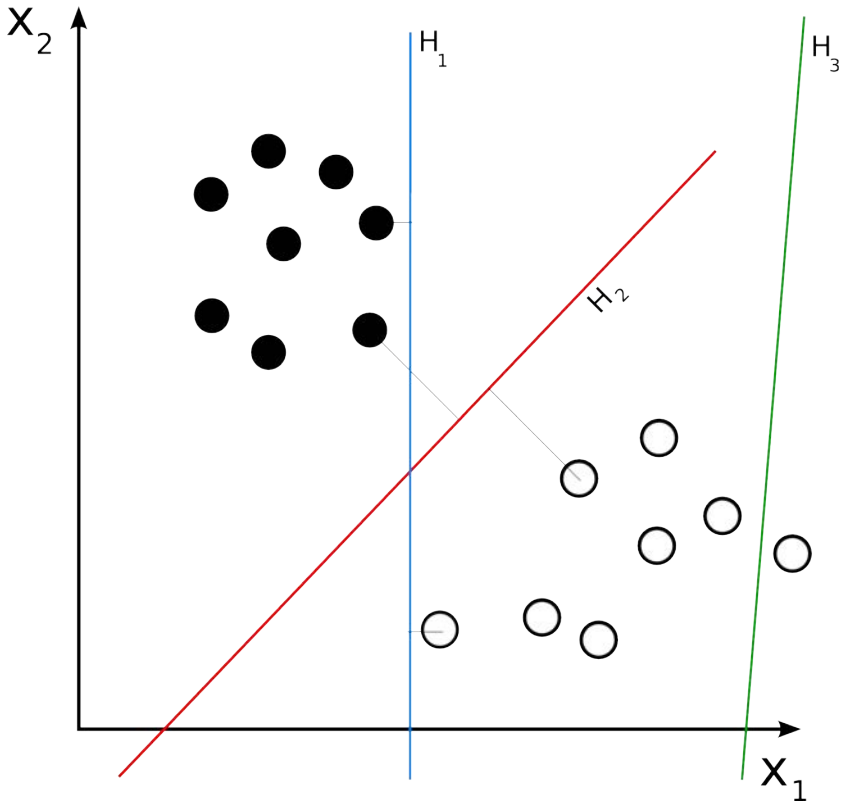

Figura 7.1: Ilustração gráfica do modo de funcionamento de uma máquina de vetores de suporte de 2 categorias e com 2 parâmetros [16].

No processo de teste, a equação da reta é a base da classificação: a categoria de um novo ponto, pertencente ao conjunto de teste, é ditada pelo lado da reta em que este se encontrar, uma vez mapeado.

No aumento do número de parâmetros, o espaço necessário para mapear os pontos é de dimensão superior (igual ao número de parâmetros). A divisão dos pontos é então feita por um hiperplano, ao invés de uma reta.

No aumento do número de categorias, existem duas estratégias: *one-against-one* e *one-against-all*. A primeira tenta encontrar as separações entre categorias criando hiperplanos para cada par de categorias; já a segunda, tenta encontrar as separações entre cada categoria e uma super categoria, resultante do agrupamento das restantes.

Pode-se ainda trocar o que é a forma linear de divisão (reta) por uma forma curvilínea, no sentido de melhorar a precisão da classificação.

As modificações feitas ao modelo simplista descrito, no desenvolvimento desta dissertação, são: o aumento de parâmetros de 2 para 16; e uma curva radial (*Radial Basis Function* (RBF)), ao invés da reta.

Este algoritmo foi então testado no WEKA [10], e, de todos os critérios avaliadores do seu desempenho contempla-se o valor da "área debaixo da curva *Receiver Operating Characteristic* (ROC)"ou "*Area Under the Curve* (AUC)", uma vez sendo o mais aceite como descritor do desempenho de um classificador [29]. Declara-se que AUC=1 significa que o classificador é infalível; à medida que o valor desce, desce também a qualidade do classificador sendo 0 o valor mínimo.

O valor de AUC obtido na simulação deste algoritmo foi 0,924. Tal valor é insuficientemente alto para que se considere que a classificação foi bem sucedida. Posto isto, surge a necessidade de se recorrer a um algoritmo mais robusto: uma rede neuronal artificial.

7.2.2.2 Redes neuronais artificiais

Uma rede neuronal artificial, RNA, ou só rede neuronal, é um algoritmo de aprendizagem de máquina que segue um modelo inspirado no funcionamento das redes neuronais humanas. Este algoritmo usa um conjunto de células de processamento (neurónios), as quais formam ligações (sinapses) entre si para produzir um resultado final [30]. A alta precisão deste algoritmo é o seu elemento mais forte, a custo de um processo de treino extremamente demorado. A lógica matemática por de trás do processo de treino do algoritmo é a seguinte:

1. Para um dado item do conjunto de dados de treino, atribua-se o valor de cada parâmetro a um neurónio de entrada da rede;

2. Criem-se uma ou mais camadas de neurónios intermédios, onde o valor de cada neurónio intermédio é uma combinação pesada dos valores dos neurónios da camada anterior;

3. Ajustem-se iterativamente os valores dos pesos de todas as ligações de neurónios até o valor do neurónio de saída ser igual ao valor atribuído à classe a que o item pertence;

4. Reajustem-se os pesos das ligações de forma a que o valor de determinada função objetivo, por exemplo, o erro quadrático médio, entre as saídas verdadeiras e as saídas calculadas, seja minimizado para todos os itens.

A Figura 7.2 ilustra o formato de uma rede neuronal com três camadas intermédias de neurónios.

O processo de teste consiste no cálculo do valor do neurónio final, utilizando os pesos estabelecidos pelo processo de treino.

Uma vez testado, o valor de AUC obtido para a rede neuronal foi 0,997. Havendo sido obtido um resultado ótimo de qualidade do classificador, há que tentar aumentar a sua eficiência, diminuindo o número de parâmetros que este recebe. As próximas secções relatam os procedimentos estatísticos que levaram a tal possibilidade.

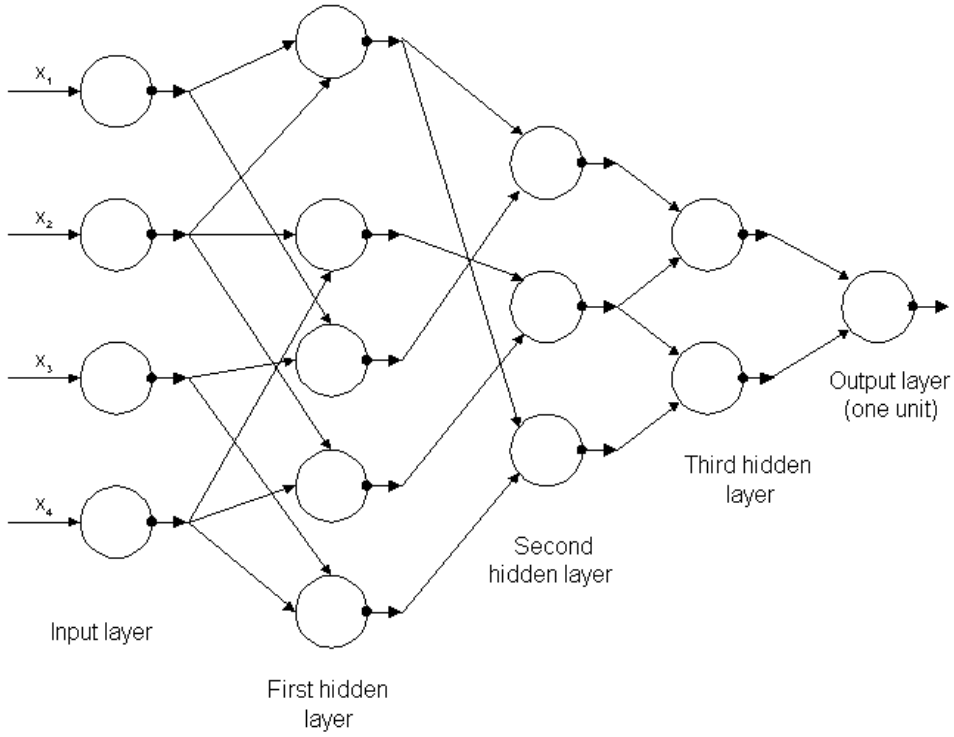

Figura 7.2: Ilustração de uma rede neuronal com três camadas escondidas [17].

7.2.3 Primeira análise Kruskal Wallis

No sentido de eliminar, dos 16 parâmetros, aqueles com fraco poder discriminativo, usou-se a análise Kruskal Wallis. A análise Kruskal Wallis é um teste estatístico que, semelhante ao teste de Friedman, testa se duas populações possuem a mesma distribuição; a grande diferença é que, onde o teste Kruskal Wallis é usado para amostras independentes, o teste de Friedman é usado para amostras dependentes. Na análise de resultados, exposta na secção 6.5.2, usou-se o teste de Friedman, visto um ouvinte produzir uma pontuação para cada orador e cada versão; neste caso, os valores de cada parâmetro de uma trama e a classificação atribuída a essa mesma trama, são valores independentes. Mais ainda, o teste Kruskal Wallis analisa a semelhança de distribuições através de uma pontuação, denominada chi-quadrado, e não uma significância; quanto mais alta esta pontuação mais correlação existe entre distribuições. Por fim, o teste Kruskal Wallis permite comparar mais de duas populações, simultaneamente.

A aplicação deste teste ao objetivo deste capítulo é o de perceber, até que ponto, determinado parâmetro é importante na classificação de uma trama. Infere-se que um parâmetro é significativo, se a sua distribuição, ao longo de todas as tramas, estiver fortemente correlacionada com a distribuição das categorias nas mesmas tramas. Isto é, se a pontuação do teste de Kruskal Wallis, entre determinado parâmetro e a categoria das tramas, for alta.

Uma vez removidos os parâmetros menos discriminativos, aumentar-se-á a eficiência na classificação de uma trama, tanto pelo facto de ser preciso calcular menos coeficientes, como pelo

facto de o classificador trabalhar menos parâmetros; no entanto, a eficácia da classificação é presumivelmente pior, dada a exclusão de informação. A operação de remoção dos parâmetros deve ser tal, que a nova classificação exiba um valor de AUC muito próximo do anterior à remoção.

A tabela 7.1 contém os resultados do teste Kruskal Wallis realizado.

Tabela 7.1: Teste Kruskal Wallis entre o conjunto original de parâmetros e a categoria das tramas.

Parâmetro	**MFCC1**	**MFCC2**	**MFCC3**	**MFCC4**	**MFCC5**	**MFCC6**
Chi-quadrado	37778,280	550,955	119039,202	5096,306	105509,028	4131,201
Parâmetro	**MFCC7**	**MFCC8**	**MFCC9**	**MFCC10**	**MFCC11**	**MFCC12**
Chi-quadrado	2930,477	50200,121	1768,709	4326,197	24886,437	1883,134
Parâmetro	**MFCC13**	**E**	**Z**	**RoR**		
Chi-quadrado	23,419	25536,466	3797,897	24,838		

Definiu-se um limiar de 10.000 pontos, no sentido de discernir quais os principais parâmetros a ter em conta. Como se pode observar na tabela, os parâmetros com pontuações acima do limiar são: MFCC1, MFCC3, MFCC5, MFCC8, MFCC11 e E. Simulou-se novamente a rede neuronal, agora só com estes parâmetros, e reavaliou-se o seu desempenho. O novo valor de AUC obtido foi de 0,970. revelando que o novo conjunto de parâmetros é suficientemente discriminativo na classificação das tramas.

Reduzir mais o número de parâmetros a calcular por trama podia prejudicar demasiado o desempenho do classificador. No entanto ainda se pode tentar reduzir o número de parâmetros com que o algoritmo de aprendizagem de máquina lida. Nas próximas duas secções realiza-se um procedimento cujo resultado comprova esta possibilidade.

7.2.4 Análise de componentes principais e segunda análise Kruskal Wallis

Considere-se a distribuição dos diversos parâmetros. Estas distribuições podem, perfeitamente, estar correlacionadas; e o mais certo é estarem, já que, a mesma trama gera valores para todos os parâmetros. Ora, não existindo qualquer interesse em fornecer ao classificador informação repetida, interessa descorrelacionar estes parâmetros. Até porque, uma vez descorrelacionados, o número de parâmetros resultante pode ser menor que o original.

Algebricamente falando, o que se faz é considerar que cada parâmetro é um eixo de um determinado espaço com dimensão igual ou menor ao número de parâmetros. Um ponto nesse espaço corresponde a uma trama, com cada coordenada a valer o que cada um dos seus parâmetros vale. O que a Análise de Componentes Principais (ACP) faz é definir um novo conjunto de eixos, que constitua uma base ortogonal daquele espaço. Estes novos eixos são chamados Componentes Principais (CP) e são, pela definição de base, completamente descorrelacionados.

Note-se que, o facto de se estar a calcular uma base de um espaço que foi definido, à custa de um conjunto de eixos correlacionados, pode significar uma redução do número de parâmetros só por si. Mas há ainda outra vantagem: ao fazer-se uma nova análise Kruskal Wallis ao conjunto de CP e categorias, é expectável que as pontuações da análise mudem, permitindo uma nova redução dos parâmetros com baixas pontuações.

No caso em estudo, ao realizar-se a análise de componentes principais, não se conseguiu obter uma imediata redução de parâmetros. Não obstante, partiu-se para a análise Kruskal Wallis entre os CP e os valores das categorias. O resultado encontra-se na tabela 7.2.

Tabela 7.2: Teste Kruskal Wallis entre os componentes principais e a categoria das tramas.

Parâmetro	**CP1**	**CP2**	**CP3**	**CP4**	**CP5**	**CP6**
Chi-quadrado	25367,414	116187,879	129,850	6151,614	630,705	2835,880

Usou-se novamente o limiar de 10.000 pontos. Assim, os componentes principais que se mantêm são: CP1 e CP2.

Para finalizar, executou-se a simulação da rede neuronal novamente e obteve-se uma AUC de 0.949. Considera-se este valor aceitavelmente próximo do original e, portanto, dá-se por terminado este processo.

7.2.5 Conclusão

Mantendo uma elevada eficácia, foi possível, com o método aqui descrito, reduzir o número de parâmetros calculados e usados na classificação de 16 para 6 e 2 respetivamente; o que, por sua vez, aumenta significativamente a eficiência do processo de classificação.

Sabe-se que a análise estatística realizada neste capítulo carece de variabilidade de oradores e, em menor grau, de conteúdo sonoro audível na fala sussurrada. Adicionalmente, o método de classificação aqui usado precisa de ser mais rápido para que, em conjunto com o processo de vozeamento artificial, seja concebida uma aplicação de tempo real. Considera-se, no entanto, que os procedimentos e os resultados obtidos neste capítulo não são de ser subestimados na sua relevância, podendo ser um excelente ponto de partida para um desenvolvimento futuro.

Capítulo 8

Conclusões

Com este capítulo, encerra-se a exposição dos conteúdos que compõem a dissertação. Contempla-se o caminho percorrido até aqui, discute-se até que ponto se atingiram os objetivos propostos desta dissertação, descrevem-se as principais dificuldades sentidas na sua elaboração e sugerem-se linhas por onde conduzir um projeto subsequente.

8.1 Satisfação dos objetivos

A proposta inicial era a de conseguir avaliar até que ponto a reconstrução de voz, através de um sussurro, era possível; e, para além disso, até que ponto a qualidade do sinal final podia ser aumentada, usando características do orador apenas existentes num sinal vozeado.

A abordagem foi a de produzir, por um lado, um sinal artificialmente vozeado auxiliado pela voz original do orador, e por outro, um completamente autónomo desta referência.

Conseguiu-se uma ótima inteligibilidade do sinal final, independentemente da ajuda da amostra original. No entanto, a naturalidade da voz, a identidade do orador, bem como a qualidade do sinal de uma forma geral, não só são medianas, como a versão independente pontua, em média, aproximadamente 1 valor abaixo nos 3 parâmetros. Estes resultados permitiram perceber que, para se conseguir uma fala vozeada agradável ao ouvido humano, é necessário continuar a melhorar a algoritmia por de trás do vozeamento; levam também a crer que, se se conseguir prever fielmente o comportamento da curva de F_0, obter-se-á uma versão autónoma de qualidade suficientemente boa para superar os métodos atuais de terapia da fala. Contudo, apesar da versão independente aqui gerada produzir um resultado sensivelmente tão desagradável como uma eletrolaringe, se a primeira fosse usada num aparelho mãos-livres não invasivo, superaria a segunda.

Adicionalmente, e tendo em mente o projeto de um transformador de fala sussurrada em vozeada, em tempo real, investigaram-se, superficialmente, métodos para a classificação automática de tramas, de acordo com a necessidade de implantar vozeamento. Percebeu-se que a classificação automática é suficientemente eficaz, se realizada com o auxílio de uma rede neuronal. Contudo, e embora se tenha feito um esforço nesse sentido, a redução do tempo de execução do classificador é um assunto premente na viabilização deste processo de classificação automática.

8.2 Principais dificuldades

As principais dificuldades sentidas no desenvolvimento desta dissertação prenderam-se, essencialmente, com a necessidade de aprendizagem de conceitos ora esquecidos, ora novos. Particularmente, toda a área da fonética foi quase cem por cento uma novidade; por sua vez, a aprendizagem da análise através do cepstrum, embora menos demorada, foi de raiz; o estudo de algoritmos de aprendizagem de máquina foi igualmente inovador; para terminar, todo o processamento estatístico, à exceção das análises descritivas, contribuiu aditivamente para o desafio.

Foram, igualmente, postas à prova, não só a capacidade de conceber soluções para problemas de processamento de sinal, como a capacidade de as implementar em Matlab.

De salientar, ainda, a intrínseca necessidade constante de gestão do tempo disponível, para a realização de todos os procedimentos. De facto, os problemas mencionados anteriormente apenas o foram pela necessidade iminente de os ver ultrapassados num curto espaço de tempo. De facto, crê-se que, relaxando parcamente a condição temporal, obter-se-iam resultados encorajadoramente superiores.

8.3 Trabalho futuro

Existem vários rumos que a evolução deste trabalho pode tomar.

A estimação e implementação de uma componente harmónica variável no tempo é um processo tão complexo quanto influente, caso se pretenda aumentar a qualidade do sinal sintetizado.

Por outro lado, se o objetivo for tornar o sinal mais fiel às características acústicas do orador, deve-se melhorar o processo de deslocamento do filtro do trato vocal, bem como aumentar a magnitude do módulo do espectro às frequências formantes.

Ou ainda, se o importante for conceber um classificador mais rápido, poder-se-á experimentar sacrificar a sobreposição de N-1 amostras por um valor menor.

8.4 Observações finais

Apesar da pressão resultante da necessidade de produzir bons resultados num curto espaço de tempo, prevalece um sentimento de gratidão e satisfação pela oportunidade do desenvolvimento deste projeto, o qual proporcionou, para além da engenharia evolvida, o contacto com outras áreas de conhecimento, nobres e de grande utilidade humana, permitindo alcançar resultados palpáveis e esperançosos no domínio da terapia de fala e do processamento de sinal.

Anexo A

Vogais da língua portuguesa e suas divisões

Este anexo contém as tabelas relativas às vogais da língua portuguesa. Estas encontram-se divididas em 2 grupos: orais e nasais.

A.1 Orais

	Anterior ou palatal	Central	Posterior ou velar
Alta	[i]	[i]	[u]
Média	[e]	[ɐ]	[o]
Baixa	[ɛ]	[a]	[ɔ]

Figura A.1: Conjunto das vogais orais da língua portuguesa [2].

A.2 Nasais

	Anterior ou palatal	Central	Posterior ou velar
Alta	[ĩ]		[ũ]
Média	[ẽ]	[ɐ̃]	[õ]
Baixa			

Figura A.2: Conjunto das vogais nasais da língua portuguesa [2].

Anexo B

Sinais e espectrogramas das amostras recolhidas

Este anexo contém os sinais das amostras recolhidas e os seus respetivos espectrogramas. As duas primeiras secções referem-se à fala vozeada e sussurrada do orador, respetivamente. E as duas últimas secções representam a situação análoga no caso da oradora.

B.1 Amostra de fala vozeada do orador

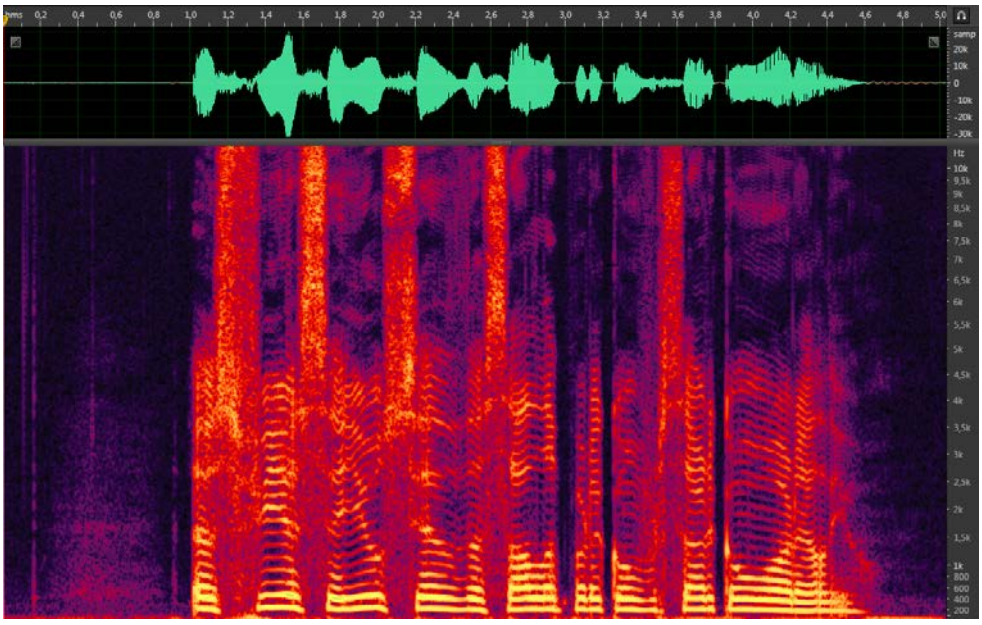

Figura B.1: Conjunto sinal e espectrograma relativo à amostra vozeada enunciada pelo orador.

B.2 Amostra de fala sussurrada do orador

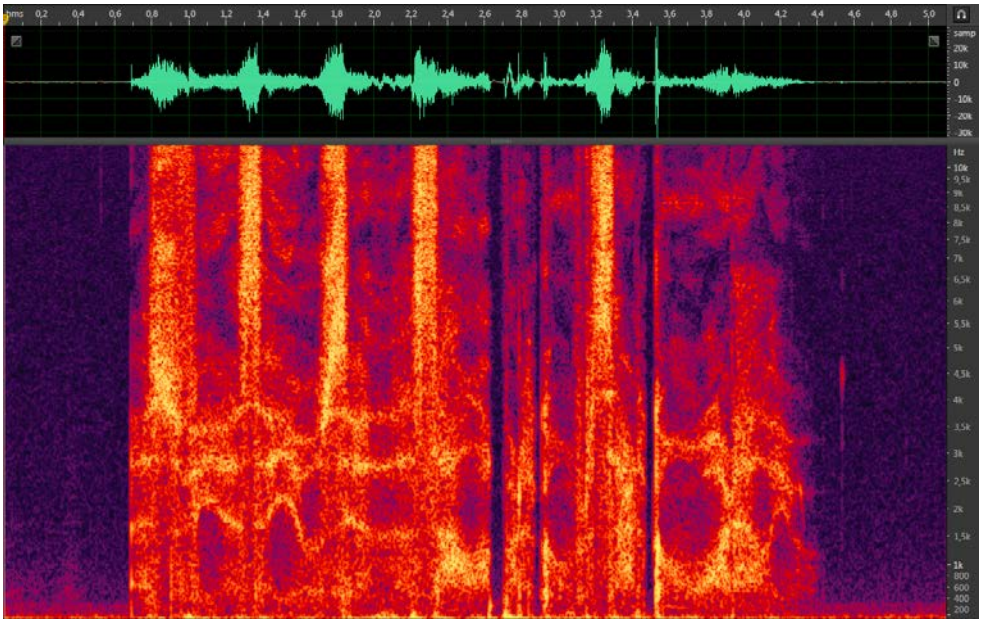

Figura B.2: Conjunto sinal e espectrograma relativo à amostra sussurrada enunciada pelo orador.

B.3 Amostra de fala vozeada da oradora

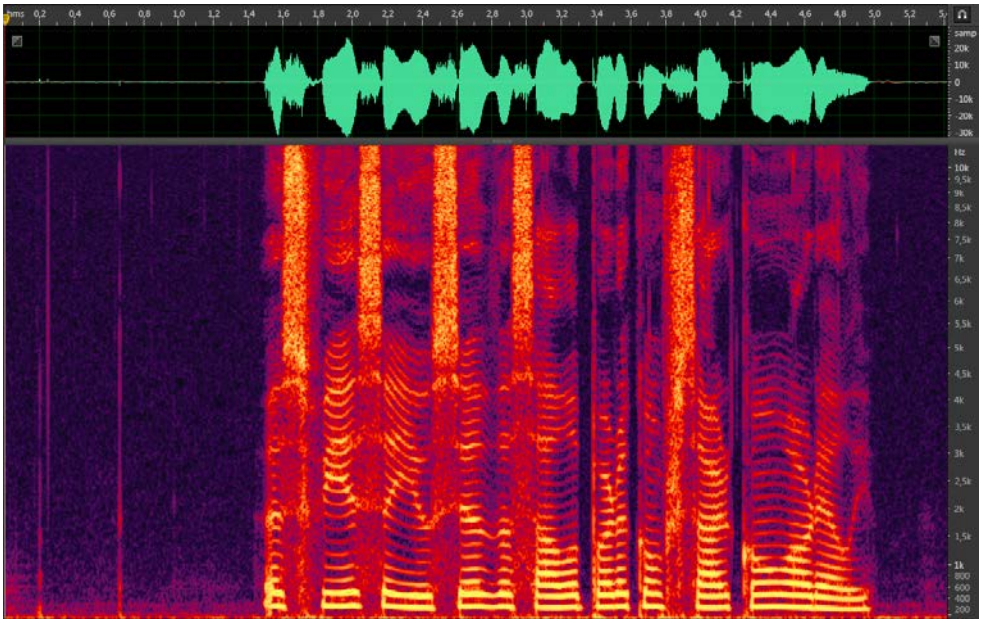

Figura B.3: Conjunto sinal e espectrograma relativo à amostra vozeada enunciada pela oradora.

B.4 Amostra de fala sussurrada da oradora

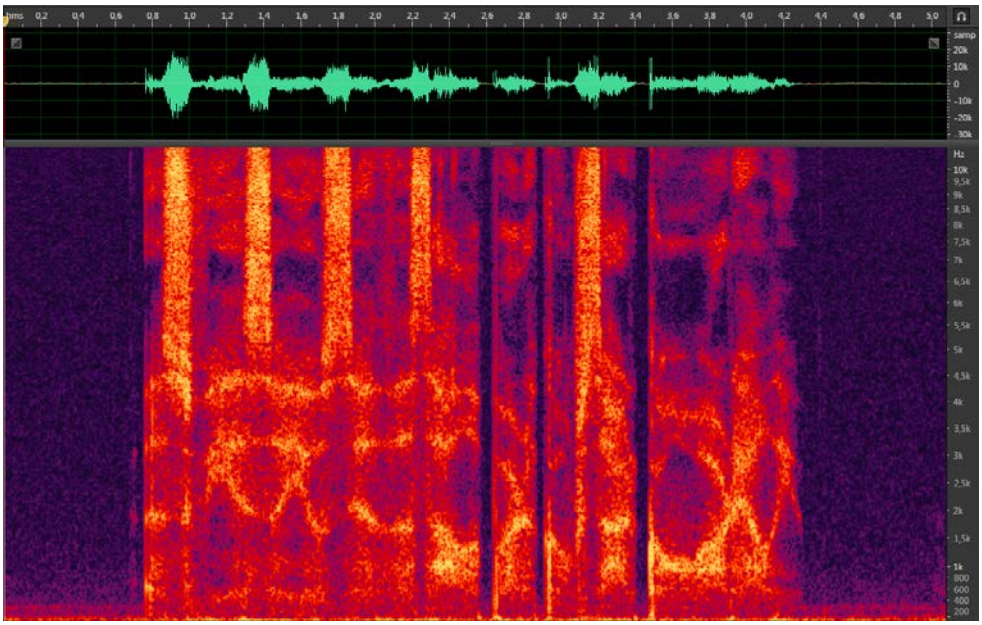

Figura B.4: Conjunto sinal e espectrograma relativo à amostra sussurrada enunciada pela oradora.

Anexo C

Sinais de fala vozeada artificialmente e seus espectrogramas

Neste anexo encontram-se outros 4 conjuntos sinal e espectrograma. Os dois primeiros são referentes às versões dependente e independente, respectivamente, geradas a partir do sinal de sussurro e vozeado do orador. Os dois últimos são, analogamente, referentes às versões dependente e independente no caso da oradora.

C.1 Versão dependente relativa ao orador

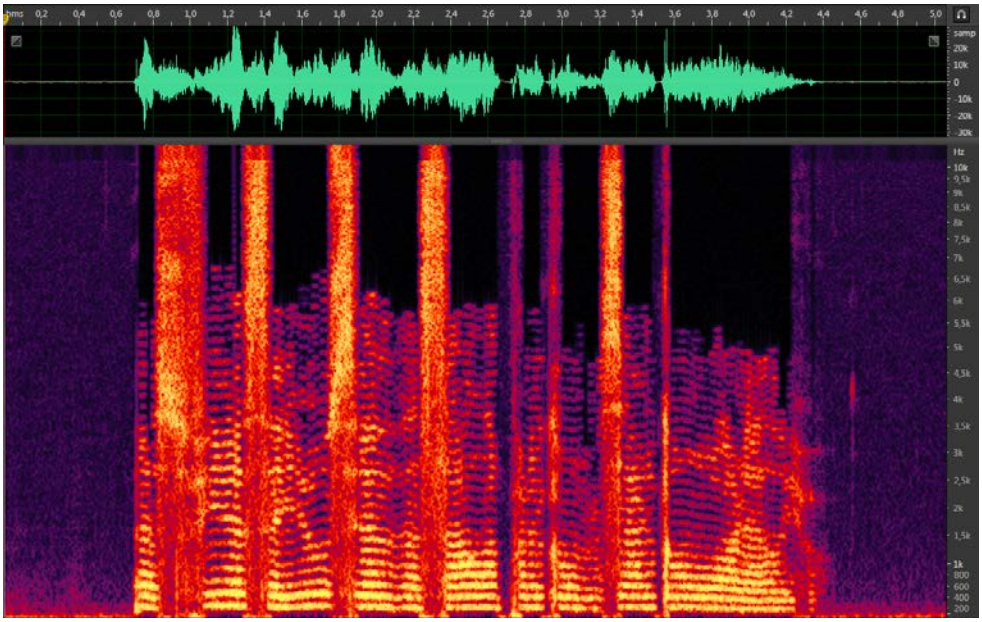

Figura C.1: Sinal vozeado artificialmente sintetizado com recurso ao sinal vozeado original do orador, e seu espectrograma.

C.2 Versão independente relativa ao orador

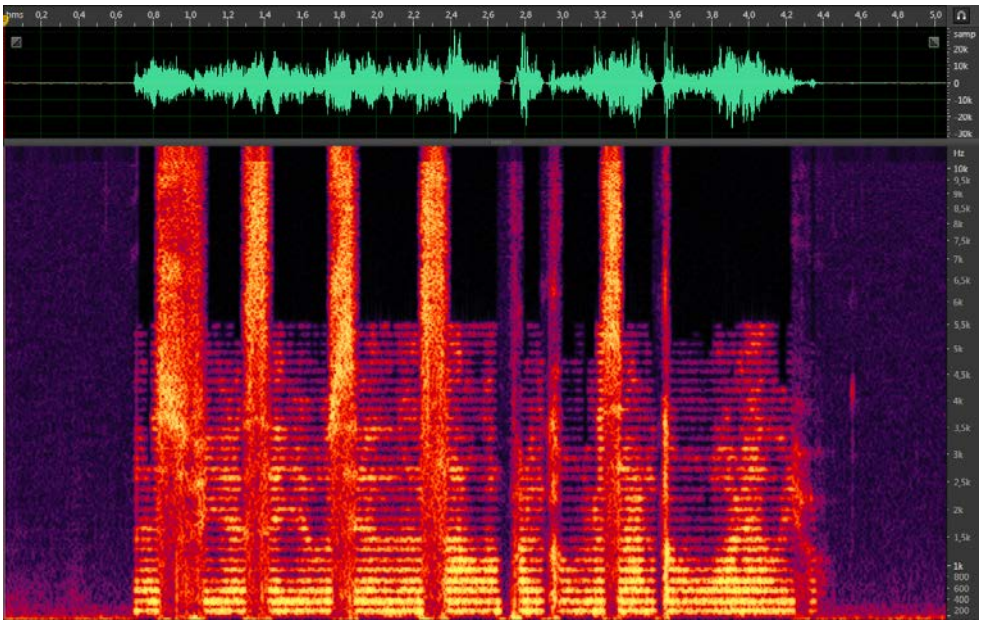

Figura C.2: Sinal vozeado artificialmente sintetizado sem recurso ao sinal vozeado original do orador, e seu espectrograma.

C.3 Versão dependente relativa à oradora

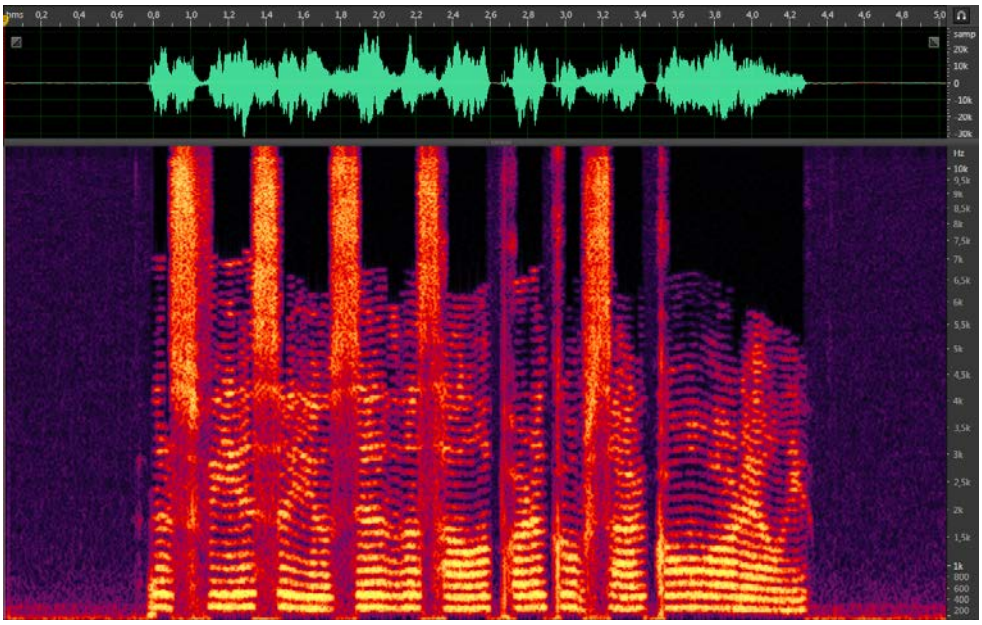

Figura C.3: Sinal vozeado artificialmente sintetizado com recurso ao sinal vozeado original da oradora, e seu espectrograma.

C.4 Versão independente relativa à oradora

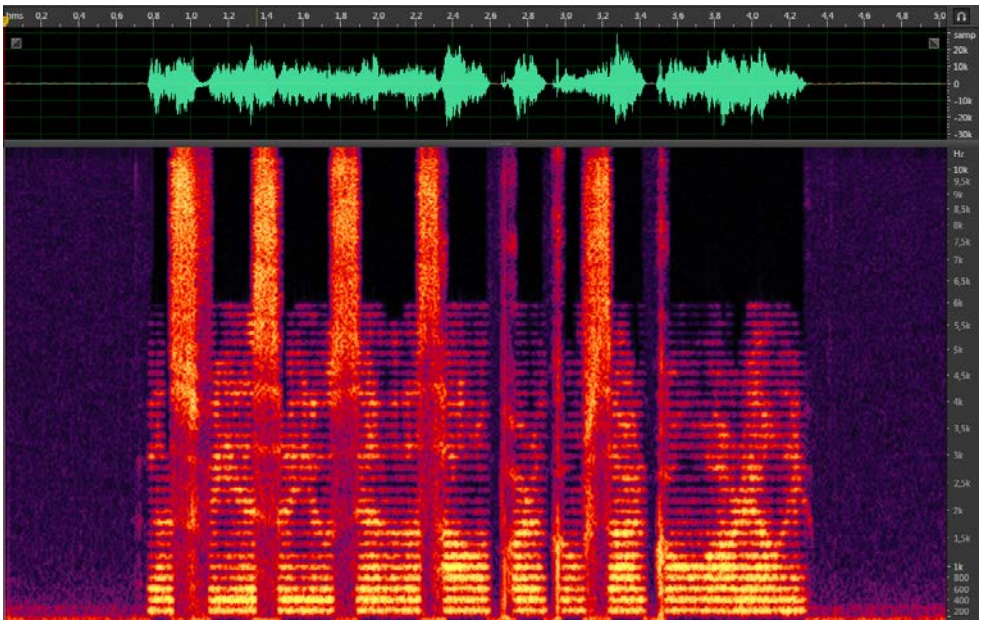

Figura C.4: Sinal vozeado artificialmente sintetizado sem recurso ao sinal vozeado original da oradora, e seu espectrograma.

Anexo D

Escalas de Likert

Este anexo contém as escalas de Likert respetivas ao teste subjetivo realizado. As tabelas encontram-se idênticas às que foram mostradas aos voluntários.

D.1 Degradação

Tabela D.1: Correspondência entre o número do nível de Likert e o seu significado, no que respeita a categoria degradação.

Degradação	
Quão grande foi a degradação de qualidade da primeira amostra para a segunda?	
Pontuação	**Significado**
5	Não detetada
4	Detetada mas não incomodativa
3	Um pouco incomodativa
2	Incomodativa
1	Muito incomodativa

D.2 Inteligibilidade

Tabela D.2: Correspondência entre o número do nível de Likert e o seu significado, no que respeita a categoria inteligibilidade.

Inteligibilidade	
Quão bem é que se percebeu o que foi dito?	
Pontuação	**Significado**
5	Percebi o que foi dito sem qualquer problema
4	Não percebi uma palavra OU percebi tudo mas com esforço
3	Não consegui perceber no máximo 3 palavras
2	Não consegui perceber 4 ou mais palavras
1	Impossível de perceber qualquer palavra

D.3 Naturalidade

Tabela D.3: Correspondência entre o número do nível de Likert e o seu significado, no que respeita a categoria naturalidade.

Naturalidade	
Quão natural(=humana) é que a voz soou?	
Pontuação	**Significado**
5	Perfeitamente natural
4	Ligeiramente robótica
3	A meio caminho entre humano e *robot*
2	Deu para perceber que era um humano a falar, mas por pouco
1	Completamente robótica

D.4 Identidade

Tabela D.4: Correspondência entre o número do nível de Likert e o seu significado, no que respeita a categoria identidade.

Identidade	
Quão bem é que deu para identificar quem estava a falar?	
Pontuação	**Significado**
5	Era claramente o/a Paulo/Mariana a falar
4	Era provavelmente o/a Paulo/Mariana a falar.
3	Não consegui perceber quem era mas se me dissessem quem era a falar eu concordaria
2	Podia ser qualquer pessoa do sexo masculino/feminino
1	Podia ser qualquer pessoa

Referências

[1] Source-filter model of speech production. Disponível em https://www.msu.edu/course/asc/232/Charts/Source-Filter%20Model%20of%20Speech%20Production.html, acedido a última vez em 10 de Setembro de 2012.

[2] Maria Helena Mira Mateus. *Fonética e Fonologia do Português*. Universidade Aberta, First edição, 2005.

[3] Hamid Reza Sharifzadeh. Reconstruction of natural sounding speech from whispers.

[4] Adobe Systems. Adobe audition cs6. Disponível em http://www.adobe.com/products/audition.html, acedido a última vez em 11 de Agosto de 2012.

[5] Webwhispers. Disponível em http://www.webwhispers.org/library/EsophagealSpeech.asp, acedido a última vez em 17 de Fevereiro de 2012.

[6] Eastern virgina medical school, department of otolaryngology. Disponível em http://www.evmsent.org/trachesoph.asp, acedido a última vez em 17 de Fevereiro de 2012.

[7] Webwhispers. Disponível em http://www.webwhispers.org/library/Electrolarynx.asp, acedido a última vez em 17 de Fevereiro de 2012.

[8] Paul Boersma e David Weenink. Praat: doing phonetics by computer. Disponível em http://www.fon.hum.uva.nl/praat/, acedido a última vez em 11 de Agosto de 2012.

[9] MathWorks. Matrix laboratory. Disponível em http://www.mathworks.com/products/matlab/, acedido a última vez em 13 de Agosto de 2012.

[10] Machine Learning Group at University of Waikato. Weka 3: Data mining software in java. Disponível em http://www.cs.waikato.ac.nz/ml/weka/, acedido a última vez em 11 de Agosto de 2012.

[11] IBM. Spss software: predictive analytics software and solutions. Disponível em http://www-01.ibm.com/software/analytics/spss/, acedido a última vez em 11 de Agosto de 2012.

[12] Sennheiser. Ear set 1. Disponível em http://en-de.sennheiser.com/ear-set-1, acedido a última vez em 13 de Agosto de 2012.

[13] Roland. Ua-25ex. Disponível em http://www.rolandus.com/products/productdetails.php?ProductId=970, acedido a última vez em 13 de Agosto de 2012.

[14] Arantza del Pozo e Steve Young. The linear transformation of lf glottal waveforms for voice conversion.

[15] Steelseries siberia v2. Disponível em http://steelseries.com/products/audio/steelseries-siberia-v2#specifications, acedido a última vez em 9 de Setembro de 2012.

[16] Svm separating hyperplanes, 2008. Disponível em http://en.wikipedia.org/wiki/File:Svm_separating_hyperplanes.png, acedido a última vez em 7 de Setembro de 2012.

[17] Polynomial neural networks. Disponível em http://ulcar.uml.edu/~iag/CS/Polynomial-NN.html, acedido a última vez em 8 de Setembro de 2012.

[18] Charles F. Hockett. The origin of speech. Setembro 1960.

[19] Aníbal Ferreira. *Comunicações Audiovisuais: tecnologias, normas e aplicações*. IST Press, First edição, 2009.

[20] Jacqueline Walker e Peter Murphy. A review of glottal waveform analysis.

[21] Ken J. Kallail e Floyd W. Emanuel. Formant-frequency differences between isolated whispered and phonated vowel samples produced by adult female subjects.

[22] Kazuya Takeda Taisuke Ito e Fumitada Itakura. Analysis and recognition of whispered speech. 2003.

[23] Maria Helena Borges Aguiar Vilarinho Machado Castro e Maria João Azevedo Padrão Ferreira Luís Miguel Teixeira de Jesus. Protocolo de anamnese vocal (jovens e adultos) da universidade de aveiro. Relatório té, Universidade de Aveiro, Março 2009.

[24] C. Gobl. A preliminary study of acoustic voice quality correlates.

[25] Methods for objective and subjective assessment of quality. Relatório té, Internation telecommunication union, 1996.

[26] Steven J. Sadoff LaDeana F. Weigelt e James D. Miller. Plosive/fricative distinction: The voiceless case.

[27] Tom Mitchell. *Machine learning*. McGraw-Hill Science/Engineering/Math, First edição, March 1997.

[28] Douglas O'Shaughnessy (1987). *Speech communication: human and machine*. Addison-Wesley, First edição, 1987.

[29] James A. Hanley e Barbara J. McNeil. A method of comparing the areas under receiver operating characteristic curves derived from the same cases.

[30] Simon Haykin. *Neural Networks: a comprehensive foundation*. Tom Robbins, Second edição, 1999.