

Faculdade de Engenharia da Universidade do Porto



Reconhecimento Robusto de Vogais Isoladas

Carlos Roberto de Oliveira Costa

Mestrado Integrado em Engenharia Eletrotécnica e de Computadores

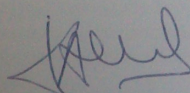
Orientador: Aníbal Ferreira (Prof. Dr. Eng.)
Coorientador: Ricardo Sousa (Dr. Eng.)

A Dissertação intitulada

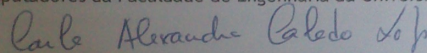
“Reconhecimento Robusto de Vogais Isoladas”

foi aprovada em provas realizadas em 22-07-2013

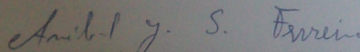
o júri



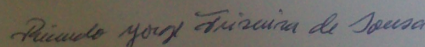
Presidente Professor Doutor Fernando Arménio da Costa Castro e Fontes
Professor Associado do Departamento de Engenharia Eletrotécnica e de
Computadores da Faculdade de Engenharia da Universidade do Porto



Professora Doutora Carla Alexandra Calado Lopes
Professora Adjunta do Departamento de Engenharia Eletrotécnica da Escola
Superior de Tecnologia e Gestão do Instituto Politécnico de Leiria

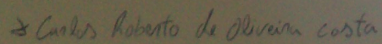


Professor Doutor Aníbal João de Sousa Ferreira
Professor Associado do Departamento de Engenharia Eletrotécnica e de
Computadores da Faculdade de Engenharia da Universidade do Porto



Doutor Ricardo Teixeira de Sousa
Bolseiro Pos - Doc do FC-UP

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 - Carlos Roberto de Oliveira Costa

Faculdade de Engenharia da Universidade do Porto

Resumo

Atualmente, sistemas de reconhecimento automático de vogais apresentam níveis de desempenho baixos quando comparados com a capacidade que o ouvido humano tem de perceber a mesma frase. Se tivermos em conta sujeitos do género feminino ou crianças, esta diferença de desempenhos acentua-se ainda mais.

No âmbito desta dissertação, serão estudados vários métodos de extração de características, alguns dos mais usados nesta área, para posteriormente e recorrendo à ajuda de um *software* estatístico com diversos classificadores, perceber quais as características ou combinação delas que conduz a melhores resultados.

Recorre-se ao uso de uma base de dados contendo vozes de diversos oradores de diferentes idades e géneros, sendo que as amostras são de curta duração pois é pretendido numa fase seguinte a este projeto ser possível fazer a identificação de vogais em tempo real.

A capacidade das características em estudo em reconhecer o orador de cada um dos diversos ficheiros de voz analisados é também testado, embora apenas o seja de uma forma exploratória.

Abstract

Currently, automatic vowel recognition systems present low levels of performance when compared to the capacity that the human hearing shows in understanding the same sentence. If we take into account female or children subjects, this difference in performance is even more notorious.

In this dissertation, various methods of features extractions will be studied, some of which are commonly used in the area of speech analysis, and subsequently using the aid of a statistical software with various classifiers realize what features or combination of them leads to better results.

We resort to the use of a database containing the voices of different speakers, of various ages and genders where voice samples are of short duration since it is desired for a next phase of development, to achieve vowel identification in real time.

The ability of the studied features to recognize the speaker of each of the different voice files is also tested, although only in an exploratory manner.

Agradecimentos

Gostaria de agradecer ao Prof. Doutor Aníbal Ferreira pela sua orientação e apoio dado ao longo da dissertação. Uma palavra de agradecimento também ao Doutor Ricardo Sousa pela orientação ajuda e disponibilidade constante durante a realização do trabalho. Não posso também deixar de agradecer ao Vítor Almeida pela ajuda prestada sempre que necessário.

Aproveito também para agradecer à minha família pelo apoio constante ao longo deste trajecto.

Aos meus amigos de longa data Luís Teixeira e João Aido, agradeço por terem estado sempre presentes, e aos meus amigos não de tão longa data, mas que serão concerteza daqui em diante e que partilham o gosto pelo *classic paleo* um grande obrigado por tudo.

Carlos Costa

Conteúdo

1 Introdução

1.1 Caracterização do problema	1
1.2 Objetivo	1
1.3 Motivação	2
1.4 Estrutura da dissertação	2

2 Estado da arte

2.1 Produção da fala	3
2.2 Vozeamento da fala	5
2.3 Modelo fonte-filtro	7
2.4 Sistema auditivo humano	8
2.5 Revisão da literatura	9
2.6 Extração de características	11
2.6.1 Linear predictive coding	11
2.6.2 Mel-frequency cepstral coding	12
2.6.3 Normalized relative coding	13
2.6.4 Perceptual linear prediction	15
2.7 Classificação	17
2.7.1 Seleção de características: Método de Krukal-Wallis	17
2.7.2 Classificadores	17
2.7.2.1 Método Nearest-Neighbour	18
2.7.2.2 Método Support vector machines	18

3 Procedimento experimental

3.1 Ferramentas usadas	21
3.1.1 Matlab	21
3.1.2 Weka	22
3.1.3 IBM SPSS Statistics	22
3.2 Base de dados	22

4 Resultados

4.1 Resultados individuais	25
4.1.1 MFCC	26
4.1.2 LPC	27

4.1.3 NRD	27
4.1.4 PLP	28
4.2 Resultado de conjuntos de características	28
4.3 Escolhas de características	30
4.3.1 Escolha de valores mais elevados	30
4.3.2 Escolha de valores segundo análise gráfica	32
4.4 Conclusões	33

5 Identificação do orador

5.1 Identificação com todos os oradores	36
5.2 Identificação de oradores por classes	36
5.2.1 Homens	37
5.2.2 Mulheres	37
5.2.3 Crianças	38
5.3 Conclusões	39

6 Conclusão

6.1 Conclusão	41
6.2 Trabalho futuro	42

A Anexo A - Seleção de características

A.1 Resultados dos testes Kruskal-Wallis	43
A.2 Seleção dos melhores coeficientes	48
A.2.1 Seleção dos valores mais altos	48
A.2.1.1 MFCC	48
A.2.1.2 LPC	48
A.2.1.3 LPC-WK	48
A.2.1.4 PLP	49
A.2.1.5 NRD	49
A.2.2 Seleção dos valores por análise gráfica.	49
A.2.2.1 MFCC	49
A.2.2.2 LPC	51
A.2.2.3 LPC-WK	52
A.2.2.4 PLP	54
A.2.2.5 NRD	55

Referências

57

Lista de figuras

Figura 2.1: Sistema vocal humano. [10]	4
Figura 2.2: Representação temporal e respetivo espectrograma da vogal /e/ emitidas por um orador masculino e um orador feminino.[10]	6
Figura 2.3: Modelo fonte-filtro	7
Figura 2.4: Computação dos coeficientes MFCC	12
Figura 2.5: Algoritmo de estimação dos NRDs e do deslocamento temporal n_0	15
Figura 2.6: Diagrama de blocos da análise de voz PLP	15
Figura A.1: Representação gráfica dos coeficientes da componente total dos sinais de voz da característica MFCC	49
Figura A.2: Representação gráfica dos coeficientes da componente harmónica dos sinais de voz da característica MFCC	50
Figura A.3: Representação gráfica dos coeficientes da componente de ruído dos sinais de voz da característica MFCC	50
Figura A.4: Representação gráfica dos coeficientes da componente total dos sinais de voz da característica LPC	51
Figura A.5: Representação gráfica dos coeficientes da componente harmónica dos sinais de voz da característica LPC	51
Figura A.6: Representação gráfica dos coeficientes da componente de ruído dos sinais de voz da característica LPC	52
Figura A.7: Representação gráfica dos coeficientes da componente total dos sinais de voz da característica LPC-WK	52
Figura A.8: Representação gráfica dos coeficientes da componente harmónica dos sinais de voz da característica LPC-WK	53
Figura A.9: Representação gráfica dos coeficientes da componente de ruído dos sinais de voz da característica LPC-WK	53
Figura A.10: Representação gráfica dos coeficientes da componente total dos sinais de voz da característica PLP	54

Figura A.11: Representação gráfica dos coeficientes da componente harmónica dos sinais de voz da característica PLP	54
Figura A.12: Representação gráfica dos coeficientes da componente de ruído dos sinais de voz da característica PLP	55
Figura A.13: Representação gráfica dos coeficientes da componente total dos sinais de voz da característica NRD	55
Figura A.14: Representação gráfica dos coeficientes da componente harmónica dos sinais de voz da característica NRD	56

Lista de tabelas

Tabela 4.1: Número de coeficientes por característica	26
Tabela 4.2: Classificação das vogais pela característica MFCC com diferentes componentes	26
Tabela 4.3: Classificação das vogais pela característica LPC com diferentes componentes .	27
Tabela 4.4: Classificação das vogais pela característica LPC-WK com diferentes componentes	27
Tabela 4.5: Classificação das vogais pela característica NRD com diferentes componentes .	28
Tabela 4.6: Classificação das vogais pela característica PLP com diferentes componentes .	28
Tabela 4.7: Classificação das vogais pelo conjunto de características e com diferentes componentes	29
Tabela 4.8: Classificação das vogais por diferentes combinações de características considerando todas as componentes	29
Tabela 4.9: Classificação das vogais pelas diferentes características após seleção dos melhores coeficientes segundo o valor mais elevado	31
Tabela 4.10: Classificação das vogais pelo conjunto de todas as características após seleção dos melhores coeficientes segundo o valor mais elevado	32
Tabela 4.11: Classificação das vogais pelas diferentes características após seleção dos melhores coeficientes segundo interpretação gráfica	32
Tabela 4.12: Classificação das vogais pelo conjunto de todas as características após seleção dos melhores coeficientes segundo interpretação gráfica	33
Tabela 4.13: Classificação dos oradores pelas diferentes características e componentes . .	36
Tabela 4.14: Classificação dos oradores masculinos pelas diferentes características e componentes	37
Tabela 4.15: Classificação dos oradores femininos pelas diferentes características e componentes	37
Tabela 4.16: Classificação das vozes de crianças pelas diferentes características e componentes	38
 Tabela A.1: Classificação de todos os coeficientes de todas as características segundo o método de Kruskal-Wallis	 43

Tabela A.2: Coeficientes selecionados segundo o valor mais alto da característica MFCC .	48
Tabela A.3: Coeficientes selecionados segundo o valor mais alto da característica LPC . .	48
Tabela A.4: Coeficientes selecionados segundo o valor mais alto da característica LPC-WK	48
Tabela A.5: Coeficientes selecionados segundo o valor mais alto da característica PLP . . .	49
Tabela A.6: Coeficientes selecionados segundo o valor mais alto da característica NRD . .	49
Tabela A.7: Coeficientes selecionados segundo análise gráfica da característica MFCC . . .	50
Tabela A.8: Coeficientes selecionados segundo análise gráfica da característica LPC . . .	52
Tabela A.9: Coeficientes selecionados segundo análise gráfica da característica LPC-WK	53
Tabela A.10: Coeficientes selecionados segundo análise gráfica da característica PLP . . .	55
Tabela A.11: Coeficientes selecionados segundo análise gráfica da característica NRD . . .	55

Abreviaturas e Símbolos

ANOVA	<i>One way analysis of variance</i>
DFT	<i>Discrete fourier transform</i>
F0	Frequência fundamental
F1	Primeira frequência formante
F2	Segunda frequência formante
F3	Terceira frequência formante
F4	Quarta frequência formante
F5	Quinta frequência formante
FFT	<i>Fast Fourier Transform</i>
LP	<i>Linear Prediction</i>
LPC	<i>Linear predictive coding</i>
Matlab	<i>MATrix LABoratory</i>
MFCC	<i>Mel-frequency cepstral coding</i>
NN	<i>Nearest neighbour</i>
NRD	<i>Normalized relative delay</i>
ODFT	<i>Odd-frequency discrete fourier transform</i>
PLP	<i>Perceptual linear prediction</i>
PSC	<i>Perceptual spectral clusters</i>
SAH	Sistema Auditivo Humano
SNR	Relação sinal-ruído
Weka	<i>Waikato environment for knowledge analysis</i>

Capítulo 1

Introdução

1.1 Caracterização do problema

Diversos estudos foram realizados, até hoje, no âmbito da percepção da fala humana através de uma aplicação informática. Esses estudos têm todos em comum a dificuldade em conseguir obter resultados que rivalizem com a capacidade que um humano tem de perceber claramente aquilo que ouve.

No âmbito desta dissertação acresce outro problema que é o de obter uma identificação em tempo real. A latência permitida entre entrada e saída é muito reduzida (cerca de 30 ms visto que este intervalo de tempo é considerado o máximo de atraso aceitável em problemas de sincronia labial). [1]

Este trabalho terá como foco a identificação robusta e isolada das vogais existentes na língua Portuguesa: /a/, /e/, /i/, /o/ e /u/.

1.2 Objetivo

Esta dissertação terá como objetivo o estudo de características que sejam capazes de distinguir diferentes vogais emitidas por diferentes oradores e posteriormente a criação de uma aplicação que identifique de forma robusta e isolada vogais emitidas por utilizadores. As aplicações do género existentes, apesar de apresentarem bons resultados no caso de utilizadores adultos do sexo masculino, desiludem quando é o caso de utilizadores do sexo feminino ou para crianças.

Será por isso de maior importância na elaboração desta dissertação, os resultados obtidos nestes últimos 2 casos.

1.3 Motivação

A motivação desta dissertação são duas áreas que poderão beneficiar da aplicação proposta: aprendizagem de línguas e terapia da fala. Aplicações para estas áreas requerem a identificação de sons falados proferidos em isolamento, particularmente vogais.

No primeiro caso a aprendizagem de uma nova língua passa inicialmente pela aprendizagem das vogais e percepção de interações vogal-consoante. Uma aplicação informática facilita a obtenção de *feedback* e facilita a aprendizagem da pronúncia correta.

No caso da terapia da fala, será mais motivante, para o paciente, trabalhar com este género de aplicação, pois poderá disponibilizar uma grande diversidade de exercícios práticos, e também poderá acelerar programas de reabilitação visto que permitirá ao paciente praticar sozinho, e em qualquer lugar, sempre que desejar.

1.4 Estrutura da dissertação

Para além da introdução, esta dissertação inclui 5 capítulos e uma secção em anexo.

No capítulo 2 é explicado o processo da produção da fala, bem como o processo de audição. É feita também uma revisão da literatura existente sobre o tema do trabalho e descrevem-se as principais características acústicas consideradas no nosso estudo. Por fim são explicitados os métodos usados para se proceder à classificação dos resultados.

No terceiro capítulo são listadas e explicadas as principais ferramentas às quais se recorreu no decurso desta dissertação. A base de dados utilizada também é abordada sendo descritas as modificações feitas na já existente.

No capítulo 4 são apresentados os resultados dos testes efetuados relativamente à identificação correta das vogais. É também feita uma discussão dos mesmos.

No capítulo 5 são demonstrados os resultados dos testes feitos no sentido de identificar o orador, sendo também estes analisados.

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

Há lugar por fim a uma secção em anexo onde são listados os gráficos e resultados dos testes Kruskal-Wallis, bem como a seleção dos coeficientes de cada característica que são usados nos testes analisados no capítulo 4.

Capítulo 2

Estado da arte

Neste capítulo serão estudados tópicos importantes a compreender no sentido de realizar o trabalho proposto.

Será importante compreender o funcionamento do sistema auditivo humano (SAH) bem como o sistema vocal de forma a projetar uma estratégia de desenvolvimento que seja adequada tomar no sentido de obter os melhores resultados.

Serão também estudados avanços existentes na área de estudo da voz, em particular os algoritmos de extração de características que permitem caracterizar a informação obtida aquando da captura da fala de um orador.

2.1 Produção da fala

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

Para que possa haver comunicação entre o orador e o ouvinte, o primeiro tem que converter as suas ideias para uma língua que ambos percebam, utilizando palavras e formando frases, que respeitem as regras gramaticais da língua escolhida. O orador pode

também utilizar entoação ou acentuar palavras para enfatizar determinados segmentos da mensagem.

Na grande maioria das línguas existentes, a produção do sinal de fala inicia-se com o diafragma a comprimir os pulmões, obrigando os alvéolos pulmonares e os brônquios a expelirem ar através da traqueia, que faz a interligação entre o pulmão esquerdo e o pulmão direito, para a laringe. O fluxo de ar passa depois da cavidade faríngea para a cavidade oral e/ou nasal, acabando por sair pela boca e/ou narinas. Na figura 2.1 está representado o sistema vocal humano com os principais órgãos e articuladores envolvidos na produção da fala, tal como explicado até aqui.

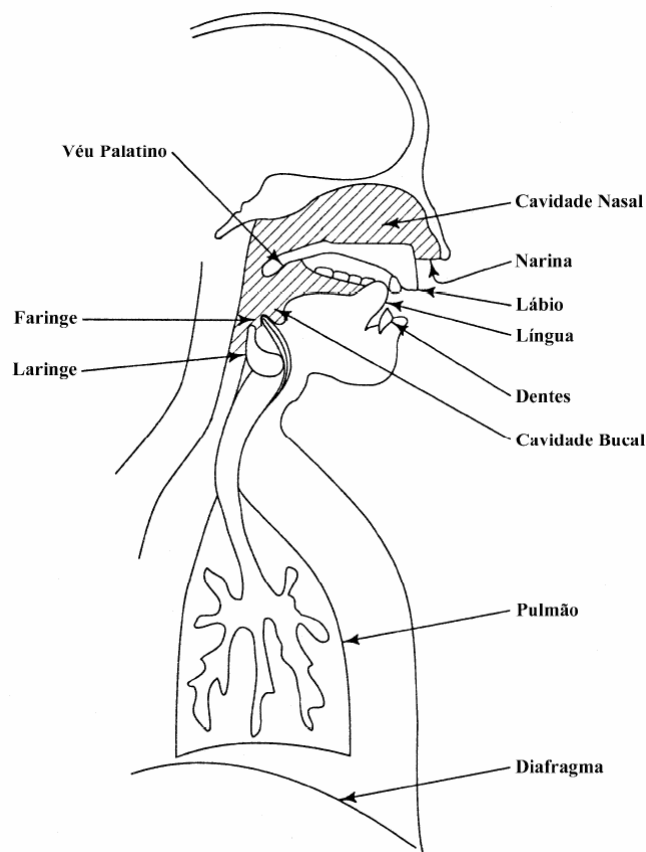


Figura 2.1: Sistema vocal humano [10]

O comprimento médio do trato vocal num homem adulto é de cerca de 17 cm, enquanto que numa mulher adulta é de cerca de 14 cm e numa criança de cerca de 10 cm. Estas diferenças terão influência na frequência e tom do discurso produzido por cada um. Ao longo do trato vocal a secção do mesmo pode variar entre 20 cm^2 e 0 cm^2 . Esta variação da secção do trato vocal é conseguida devido à mobilidade das pregas vocais, do palato mole ou véu palatino, da língua, da mandíbula e dentes e dos lábios que, com alterações das suas

posições, conseguem alterar as dimensões do trato vocal e assim alterar as propriedades acústicas do som emitido.

2.2 Vozeamento da fala

Os sons de fala são gerados com ou sem vibração das pregas vocais. Os sons produzidos sem vibração das pregas vocais são designados de não vozeados, enquanto os sons produzidos com vibração das pregas vocais, ou seja, através da abertura e fecho da glote (espaço entre as pregas vocais), são designados de vozeados. O vozeamento acontece quando o fluxo de ar vindo dos pulmões passa pela laringe e as pregas vocais interrompem esse fluxo de uma forma quase periódica, excitando assim o trato vocal. [10]

Nas zonas vozeadas, à medida que as pregas vocais vibram, estas fazem variar o grau de abertura da glote e consequentemente o volume de ar proveniente dos pulmões que passa através dela. [10]

A energia do som produzido pelas pregas vocais é um tom periódico complexo com uma frequência fundamental baixa, também conhecida como frequência fundamental da fonação (F_0). [2]

A frequência fundamental depende da dimensão e espessura das pregas vocais, o que faz com que a voz de uma pessoa varie com a idade. Para oradores do género masculino, a gama de vibração das pregas vocais situa-se entre os 50 e os 250 Hz, enquanto para oradores do género feminino essa gama situa-se entre os 120 e os 300 Hz. No caso das vozes de crianças a gama de vibração pode chegar aos 500 Hz.

Um orador pode ser caracterizado através da sua frequência fundamental média, com variações naturais dependentes da entoação, *stress* e emoção. É normal um orador apresentar uma variação que pode atingir em fala natural uma oitava (por exemplo entre 80 e 160 Hz para um orador masculino), podendo atingir 2 oitavas no caso de fala forçada ou cantada. [10]

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

Posteriormente, o trato vocal filtra o som produzido introduzindo ressonâncias, (designadas também por frequências formantes ou simplesmente formantes) a certas frequências. Estes formantes são numerados: o primeiro com a frequência mais baixa é chamado de primeiro formante (F_1), o seguinte é chamado de segundo formante (F_2) e assim por diante. As frequências centrais dos formantes diferem com a forma do trato vocal. [10]

Vogais são os sons da fala que são mais facilmente caracterizáveis visto que os seus formantes e outras características acústicas são relativamente estáveis ao longo do tempo, quando falados em isolamento. [2]

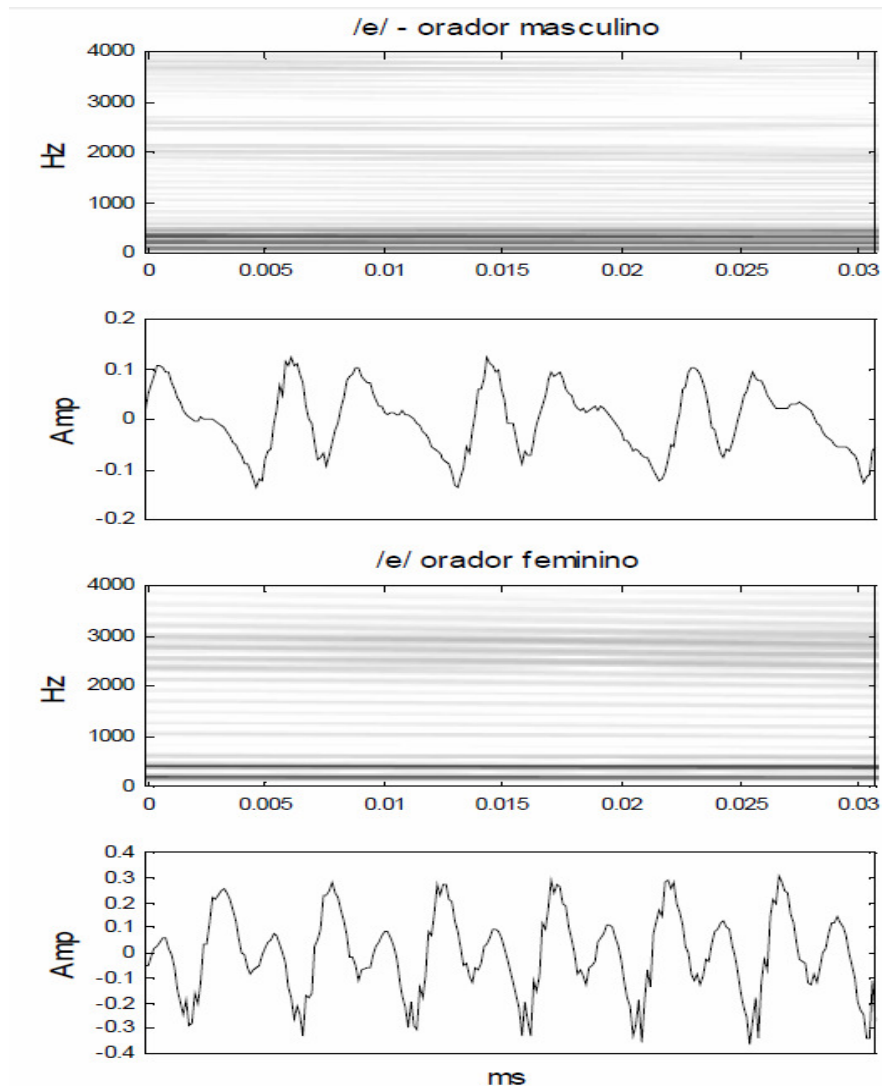


Figura 2.2: Representação temporal e respetivo espectrograma da vogal /e/ emitidas por um orador masculino e um orador feminino.[10]

A figura 2.2 apresenta a forma de onda de um segmento vozeado, a vogal /e/, dito por um orador do género masculino e por um orador do género feminino. Está ainda representado o respetivo espectrograma para cada orador, ou seja gráficos tempo versus frequência, em que a intensidade em cada ponto dá informação da energia associada a cada frequência num instante de tempo determinado. Pode-se verificar o maior valor da frequência fundamental para o orador feminino em relação ao orador masculino e a correspondente melhor definição em frequência, já que as harmónicas se encontram mais espaçadas.

2.3 Modelo fonte-filtro

O modelo fonte-filtro foi originalmente proposto por Fant [13] e consiste numa simplificação do sistema de produção de voz, sobre as perspectivas física e de processamento de sinal.

Este modelo descreve a produção de fala como um processo constituído por duas etapas independentes, envolvendo a geração de som de fonte, com a sua própria forma espectral e estrutura espectral fina, que é depois modelada ou filtrada pelas propriedades ressonantes do trato vocal. [2]

A primeira etapa consiste na produção de som através da vibração das pregas vocais no caso da voz vozeada e ruído resultante nomeadamente de turbulência no caso da voz não-vozeada. Na segunda etapa, o trato vocal modela a fonte quando o sinal desta passa pelas várias cavidades e formas que o constituem.

O trato vocal é modelado por um filtro caracterizado por uma função de transferência $H(z)$.

Na figura 2.3 está de forma simples e resumida representado o que se acabou de descrever.

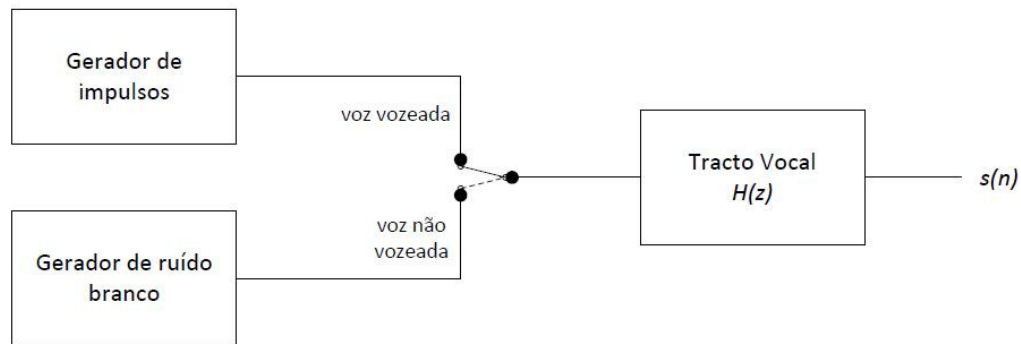


Figura 2.3: Modelo Fonte-filtro

Um gerador de impulsos é responsável pela emissão da voz vozeada que após aplicação de um filtro ($H(z)$) fornece uma versão sintética de voz vozeada.

No caso da voz não vozeada, a sua emissão está ao cargo de um gerador de ruído branco que é igualmente filtrado pelo trato vocal para a obtenção do sinal de voz final.

2.4 Sistema Auditivo Humano

Diversos estudos têm sido feitos no sentido de perceber o SAH, embora sem nunca ter sido possível obter conclusões definitivas sobre o seu funcionamento. No entanto, é geralmente aceite que os primeiros dois ou três formantes são usados para perceber vogais ouvidas. [3]

Segundo o estudo de Baumann e Berlin [5], F1 apresenta grande importância para diferenciar vozes femininas, enquanto que para vozes masculinas a dispersão entre F4 e F5 se revela mais importante.

As frequências dos primeiros três formantes (F1, F2 e F3) são geralmente considerados os mais proeminentes relacionais acústicos de uma determinada vogal. O contraste espectral (ou seja, a magnitude relativa dos picos e vales do espectro) e frequência são também fatores importantes no reconhecimento de vogais pelos humanos. [4]

Os formantes F4 e F5 são conhecidos por serem mais independentes da vogal falada mas possuem tipicamente muito menos energia em espectrogramas de voz feminina quando comparadas com os masculinos. Assim, apesar destes dois formantes serem mais adequados para distinção de falantes, o nível de energia poderá ser na maioria dos casos demasiado baixo para ser usado na identificação de vozes femininas. [5]

Um estudo efetuado por Assmann e Terrance [12] compara o impacto da frequência fundamental F0 relativamente à identificação de vogais, emitidas por humanos quando existe uma alteração das suas frequências originais. Neste estudo, a identificação de vogais apresentou um declínio gradual quando a envolvente espectral era alterada quer em vogais faladas por homens, mulheres ou crianças. O aumento da envolvente espectral levou a uma diminuição da identificação de vogais emitidas por crianças comparativamente a adultos, enquanto que a diminuição da envolvente espectral teve maior impacto em vogais faladas por homens comparativamente a mulheres e crianças. Uma interpretação possível do observado é que o declínio de performance é um reflexo da experiência dos ouvintes relativamente a discurso natural e a falta de exposição a padrões formantes extremos como neste caso, o que é consistente com estudos de Rosen et al. 1999; Nogaki et al. 2007 que concluem que a inteligibilidade do discurso melhora com exposição prolongada ao mesmo. [12]

Alterações coordenadas (F0 e envolvente espectral aumentadas ou diminuídas em simultâneo) produziram de forma geral melhores resultados do que quando F0 e a envolvente espectral eram alterados de forma oposta. A performance é mais reduzida quando ocorrem deslocamentos opostos pois quando F0 é mais elevado que a frequência do primeiro formante a qualidade da vogal é deteriorada. Assim, é sugerido que uma representação melhorada de F1 poderá levar a melhores resultados. [12]

Segundo [5] é sugerido que será uma melhor estratégia usar *shimmer*, *jitter*, o desvio padrão de F0, F5, a dispersão entre F1 e F5, a dispersão entre F4 e F5, ou a duração das amostras vocais para diferenciar oradores.

2.5 Revisão da literatura

Enquanto humanos apresentam poucas ou nenhuma dificuldade em lidar com o reconhecimento de voz falada em situações acústicas desafiantes, sistemas de reconhecimento automático de discurso regra geral não conseguem apresentar a mesma robustez do sistema auditivo humano. Esta observação tem motivado estudos que comparam a performance de reconhecimento de discurso humano com sistemas de reconhecimento automático, com o objetivo de os melhorar aprendendo os princípios de funcionamento da audição humana. Estudos reportam que o fosso entre a capacidade de reconhecimento humano e sistemas de reconhecimento automático aumenta com a complexidade da tarefa de reconhecimento, por exemplo, com o aumento de ruído ou o aumento do vocabulário a identificar. Um estudo de Sroka and Braida (2005) concluiu que um sistema automático usando características cepstrais alcançou resultados próximos do humano apenas quando a relação sinal-ruído era aumentada 10 dB. Observaram ainda que filtros passa-alto e passa-baixo reduziram (e em certos casos, chegaram a eliminar) a diferença. Assim Sroka e Braida concluem que humanos e classificadores automáticos fazem uso de diferentes pistas. [11]

A maioria das abordagens para reconhecer vogais emitidas por uma pessoa são baseadas em estimação de formantes fazendo uso de técnicas de *linear prediction* (LP). Estas técnicas assumem que a produção de sons pelo sistema fonético humano podem ser modelados como um filtro só de pólos (“*all-pole*”) que é excitado por um fluxo periódico de impulsos glotais. A taxa de repetição destes impulsos corresponde à frequência fundamental (F0), ou tom, e os pólos do filtro correspondem às ressonâncias do trato vocal ou formantes. [4]

Este método tem falhas graves no caso das vozes femininas e de crianças, pois quando F0 é comparável ou superior a F1, o que acontece neste caso, ou no caso de ser voz cantada, técnicas de LP não são de confiança pois a magnitude do espectro torna-se *undersampled*, isto é, só é amostrada em múltiplos inteiros da frequência fundamental. Uma observação frequente é que as frequências formantes estimadas estão “trancadas” para harmónicas no espectro de magnitude, o que representa um problema de robustez. Ainda assim, o sistema auditivo humano não é afetado nestas circunstâncias o que denota que este realiza normalização do discurso usando provavelmente outros parâmetros do sinal acústico, nomeadamente F0. [4]

Uma abordagem diferente é proposta por [6] em que o sinal do discurso é modelado usando intervalos de distribuição *zerocrossing* no domínio do tempo. São usados os padrões

de distribuição, (que se verificou serem similares para uma mesma vogal, mas variantes de vogal para vogal) para reconhecer as vogais usando redes neuronais artificiais de múltiplas camadas sem realimentação. Verifica-se que este método resulta de forma efetiva para o objetivo proposto (identificação de vogais), embora só tenha sido testado com um único sujeito. Também é estudada a adaptação ao ruído, adicionando ruído aditivo gaussiano branco com diferentes relações sinal-ruído. Verifica-se que quando a relação sinal-ruído é de 0 dB, 3dB e 10dB a precisão é fraca. Acima de 20 dB, os resultados melhoram significativamente.

A grande vantagem desta abordagem é a reduzida complexidade computacional quando comparada com as técnicas mais convencionais como *fast Fourier transform* (FFT) e métodos de agrupamentos.

Outra alternativa é apresentada por [7], onde uma visão geral de diversos preditores lineares e codificação é obtida introduzindo restrições de esparsidade no *framework* de LP. Em análise de discurso estes preditores mostraram conseguir obter um desacoplamento mais eficiente entre tons harmónicos e a envolvente espectral. Isto traduz-se em preditores que não são corrompidos pela fina estrutura da excitação do tom e oferecem propriedades interessantes tais como a janela invariante e invariância do tom. A codificação robusta baseada em predição forneceu uma aproximação competitiva para codificação de discurso com uma abordagem de diversos níveis energéticos e também um mais lento decaimento de qualidade para relação sinal ruído (SNR) mais reduzido.

Uma técnica diferente é proposta por [8] onde o tom é tido como elemento central. O conceito de perceção de vogais assume que o SAH discrimina vogais ao executar um agrupamento percetivo espectral (*perceptual spectral clusters* - PSC) de parciais na estrutura harmónica da vogal. O PSC tenta identificar agrupamentos de parciais harmónicos cujas características dão origem a pistas percetivas relevantes usadas pelo SAH em discriminação e identificação de vogais.

O conceito PSC é fortemente enraizado na ideia que o reconhecimento humano de uma vogal sustentada resulta da identificação do tom e timbre, sendo ambos sensações percetivas. É conhecido que os parciais de uma estrutura harmónica são fundidos (ou integradas) na perceção de um único tom, mesmo que alguns parciais estejam em falta. Por outro lado o timbre é comumente visto como a “cor” do som e, no caso do som harmónico tal como a emissão de vogais, depende da energia espectral dos seus parciais, como acontece em sons musicais. Assim, para uma vogal, a análise do timbre requer a identificação da estrutura harmónica subjacente. O conceito PSC é construído na integração percetiva dos parciais pertencentes à mesma estrutura harmónica, e tentam identificar agrupamentos de parciais harmónicos e as suas características que ajudam a explicar a habilidade do SAH para discriminar vogais de forma robusta. Deste modo, admite-se que um segundo nível de integração percetiva envolvendo parciais harmónicos em cada PSC é realizado pelo SAH. [4]

Os resultados obtidos do estudo desta técnica situaram-se entre 82% e 84% de vogais reconhecidas corretamente, o que apesar de ainda ser distante do valor que o SAH consegue reconhecer, perto dos 100%, é positivo e encoraja a serem estudadas alternativas a características de PSC, ou num uso mais apropriado da informação contida no tom. [8]

2.6 Extração de características

2.6.1 Linear Predictive Coding

Como já foi referido na secção 2.5, em análise de sinal de voz é muito utilizada a predição linear (LP).

Assume-se que o sinal de som é a combinação linear dos seus valores passados e de uma entrada atual. Em condições ideais, a aplicação deste método a uma região de voz vozeada resulta num filtro IIR *all-pole* e um trem de impulsos (que correspondem ao resíduo).

A equação 2.1 representa a saída atual. O valor da saída de índice m é obtido usando uma combinação linear dos P valores passados onde m é o índice do tempo discreto, $\hat{x}(m)$ a predição de $x(m)$ e a_k os coeficientes de predição.

$$\hat{x}(m) = \sum_{k=1}^P a_k x(m-k) \quad (2.1)$$

O erro de predição é dado pela equação 2.2, que é a diferença entre o sinal $x(m)$ e o sinal estimado $\hat{x}(m)$.

$$e(m) = x(m) - \sum_{k=1}^P a_k x(m-k) \quad (2.2)$$

O que se procura obter através do modelo LPC são os coeficientes a_k expressos num vector de p dimensões, para uma predição de ordem p . Estes coeficientes são determinados de forma a minimizar uma medida do erro $e(m)$, tipicamente o seu valor quadrático médio. Tendo em conta que $e(m)$ contém toda a informação da voz que não é modelada pelos coeficientes de predição, minimizar o erro significa maximizar a informação expressa pelo modelo LPC.

Se a ordem P da predição linear for muito baixa, ressonâncias importantes do trato vocal podem ser descartadas, enquanto que se esta for muito alta, a fonte do sinal é modulada e não as características do trato vocal, o que revela a importância da ordem para as características do modelo do trato vocal a analisar.

2.6.2 Mel-Frequency Cepstral Coding

Os *Mel-Frequency Cepstral Coefficient* (MFCC) são parâmetros que descrevem o perfil da magnitude e que consideram as características do ouvido humano [14]. Estes podem ser definidos como o cepstro de uma janela de análise determinada a partir de uma DFT numa escala de frequência e magnitude que é característica da audição humana.

O cepstro é uma representação do sinal de voz em que um sinal de fonte glótica, de variação temporal rápida, e a resposta do trato vocal, de variação lenta, são desacoplados e transformados em dois componentes aditivos. O cepstro é obtido através da seguinte fórmula:

$$\text{Cepstro}(\text{frame}) = \text{IDFT}(\log(|\text{DFT}(\text{frame})|)) \quad (2.3)$$

Ao calcular-se a DFT da *frame* obtém-se uma multiplicação dos termos, e ao calcular-se o logaritmo transforma-se essa multiplicação numa soma. Após aplicar a DFT inversa obtém-se uma representação das duas componentes do sinal de voz em que estas se encontram perfeitamente distintas. Na figura seguinte (2.4) é apresentado o diagrama de blocos que conduz ao cálculo dos coeficientes MFCC



Figura 2.4: Computação dos coeficientes MFCC

Em primeiro lugar é aplicada uma janela de análise a uma *frame* do sinal, tipicamente uma janela de Hamming, de forma a reduzir o efeito de *leakage* e obter a

magnitude do espectro. É aplicado um banco de filtros passa-banda triangulares em escala Mel com o objetivo de reduzir as características e realçar as frequências que, de um modo perceptivo, são mais importantes. São consideradas mais significativas as componentes de frequência mais baixa.

As bandas dos filtros passa-banda crescem com o aumento da frequência. A escala linear de frequência é convertida na escala Mel através da equação 2.4. Esta transformação é feita de forma a dar menos ênfase às altas frequências, pois o sistema auditivo humano apresenta menor sensibilidade a estas frequências.

$$M(f) = 2595 \log_{10} \left(1 + \frac{f}{700} \right) \quad (2.4)$$

Finalmente, é aplicada a transformada *Discrete Cosine Transform* do tipo II (DCT-II) para converter os valores obtidos do banco de filtros em coeficientes cepstrais.

2.6.3 Normalized Relative Delay

Outra característica também a estudar neste trabalho serão os *Normalized Relative Delay* (NRD).

Características de sinal baseadas na magnitude do espectro são usadas de forma comum em processamento de fala. Por outro lado, a fase do espectro raramente é utilizada devido à dificuldade de a interpretar e manipular. A fase é então frequentemente ignorada apesar de existirem evidências que mostram que sinais com a mesma magnitude mas diferente fase de espectro são perceptivamente distintas para um ouvinte humano.

Os coeficientes NRDs fazem a modelação dos atrasos relativos entre os diferentes harmónicos, refletindo a contribuição da fonte glótica e do atraso de grupo do filtro formado pelo trato vocal.

Deste modo torna-se interessante estudar o impacto que as características NRDs irão apresentar na identificação de vogais.

Passando a descrever o algoritmo de extração de NRDs:

Considere-se um sinal quasi-periódico constituído por M sinusóides relacionadas entre si através da frequência fundamental ω_0 :

$$\begin{aligned} s[n] &= A_0 \sin(n\omega_0 + \varphi_0) + \sum_{i=1}^{M-1} A_i \sin(n\omega_i + \varphi_i) \\ &= A_0 \sin \omega_0(n + n_0) + \sum_{i=1}^{M-1} A_i \sin(\omega_i(n + n_i)) \end{aligned} \quad (2.5)$$

em que A_i representa a magnitude, ω_i a frequência, φ_i a fase e n_i o atraso temporal do harmónico i. Ao multiplicar $s[n]$ por uma janela temporal $h[n]$ e efetuar uma transformação para o domínio das frequências, através da DFT, resultam os coeficientes espectrais $X[k] = \text{DFT}\{x[n]\}$, onde $x[n] = s[n]h[n]$. As fases dos picos harmónicos em $X[k]$ dependem do índice de tempo n , pois correspondem ao atraso de grupo da janela de tempo $h[n]$. Devido à simetria par que esta janela tem, o atraso de grupo é constante e independente da frequência. Sendo N o comprimento da janela temporal (e também o comprimento da DFT) o atraso de grupo é simplesmente $(N-1)/2$ amostras. Assim, ao analisar os picos harmónicos em $X[k]$, o índice temporal n toma o valor da constante de tempo de referência correspondente ao atraso de grupo afetando de forma igual todas as frequências. De modo a focar a análise nos atrasos dos harmónicos relativos à mesma referência temporal poderemos ignorar a influência de n na fórmula anterior. Para simplificar, considerando a referência temporal igual a zero :

$$\begin{aligned}
 s &= A_0 \sin(n_0 \omega_0) + \sum_{i=1}^{M-1} A_i \sin(n_i \omega_i) \\
 &= A_0 \sin\left(2\pi \frac{n_0}{P_0}\right) + \sum_{i=1}^{M-1} A_i \sin\left(2\pi \frac{n_i}{P_i}\right) \\
 &= A_0 \sin\left(2\pi \frac{n_0}{P_0}\right) + \sum_{i=1}^{M-1} A_i \sin\left(2\pi \frac{n_0 + n_i - n_0}{P_i}\right) \\
 &= A_0 \sin\left(2\pi \frac{n_0}{P_0}\right) + \sum_{i=1}^{M-1} A_i \sin\left[2\pi \left(\frac{n_0}{P_i} + \frac{n_i - n_0}{P_i}\right)\right] \\
 &= A_0 \sin\left(2\pi \frac{n_0}{P_0}\right) + \sum_{i=1}^{M-1} A_i \sin\left[2\pi \left(\frac{n_0}{P_i} + NRD_i\right)\right]
 \end{aligned} \tag{2.6}$$

em que P_i representa o período da sinusoide i.

Na figura seguinte ilustra-se o algoritmo NRD:

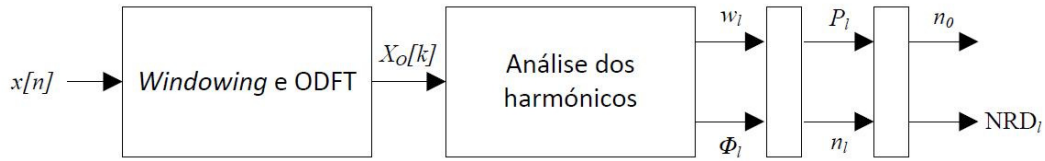


Figura 2.5: Algoritmo de estimação dos NRDs e do deslocamento temporal n_0 .

O sinal de fala $s[n]$ é multiplicado por uma janela de seno e transformado para o domínio das frequências através da *Odd-frequency Discrete Fourier Transform* (ODFT). De seguida é feita uma análise dos harmónicos para estimar a amplitude, a frequência e a fase destes. Destes dados são usados a frequência e a fase para calcular o período ($P_i = 2\pi / \omega_i$) e o atraso ($n_i = \phi_i / \omega_i$), que depois permitem calcular os NRDs, como explicado anteriormente.

2.6.4 Perceptual linear prediction

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

Na figura 2.6 está representado o diagrama de blocos correspondente à análise PLP

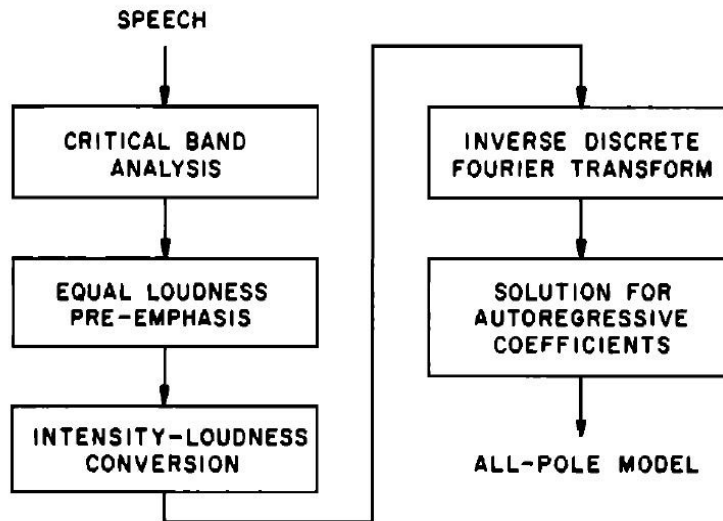


Figura 2.6: Diagrama de blocos da análise de voz PLP

A análise PLP consiste em segmentar o sinal de fala usando um janelamento de Hamming, com janelas de aproximadamente 20ms. A transição para o domínio das frequências é feita usando a transformada discreta de Fourier, normalmente a FFT de 256 pontos para uma frequência de amostragem de 10kHz. O espectro de potência de termo curto é obtido somando o quadrado das componentes real e imaginária do espectro de sinal de termo curto, como demonstra a equação 2.7.

$$P(\omega) = \text{Re}[S(\omega)]^2 + \text{Im}[S(\omega)]^2 \quad (2.7)$$

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

$$\Omega(\omega) = 6 \ln \left\{ \frac{\omega}{1200\pi} + \left[\left(\frac{\omega}{1200\pi} \right)^2 + 1 \right]^{0.5} \right\} \quad (2.8)$$

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

$$\Psi(\Omega) = \begin{cases} 0 & \text{se } \Omega < -1.3, \\ 10^{2.5(\Omega+0.5)} & \text{se } -1.3 \leq \Omega \leq -0.5, \\ 1 & \text{se } -0.5 < \Omega < 0.5, \\ 10^{-1.0(\Omega-0.5)} & \text{se } 0.5 \leq \Omega \leq 2.5, \\ 0 & \text{se } \Omega > 2.5 \end{cases} \quad (2.9)$$

A convolução dos dois espectros permite uma reamostragem a intervalos de aproximadamente um Bark. O sinal reamostrado é depois pré-enfatizado pela curva de igualização de sonoridade, que simula a sensibilidade auditiva humana para valores de aproximadamente 40 dB. Essa aproximação é dada pela equação 2.10

$$E(\omega) = \left[(\omega^2 + 56.8 \cdot 10^6) \omega^4 \right] / \left[(\omega^2 + 6.3 \cdot 10^6)^2 \cdot (\omega^2 + 0.38 \cdot 10^9) \right] \quad (2.10)$$

e o resultado da pré-ênfase é dado pela equação 2.11

$$\Xi[\Omega(\omega)] = E(\omega) \Theta[\Omega(\omega)] \quad (2.11)$$

Por fim utiliza-se uma compressão de raiz cúbica, de acordo com a equação 2.12, para simular a relação não-linear entre a intensidade de um determinado som e a percepção da *loudness* e do mesmo pelo ouvido humano.

$$\Phi(\Omega) = \sqrt[3]{E(\Omega)} \quad (2.12)$$

O sinal $\Phi(\Omega)$ é aproximado pelo espectro de um modelo autorregressivo só com pólos usando o método da autocorrelação sendo que no final é possível extrair determinadas características, como é o caso dos coeficientes do modelo autorregressivo ou coeficientes cepstrais.

2.7 Classificação

Após extrair diferentes características dos sinais de áudio a estudar, será necessário analisa-los e perceber quais os mais discriminativos e com maior peso na distinção das vogais emitidas. Assim será de seguida apresentado o classificador usado bem como o método estatístico que serviu para distinguir essas mesmas características.

2.7.1 Seleção de características: Método de Krukal-Wallis

O método de Kruskal-Wallis é uma extensão do método de Wilconxon-Mann-Whitney. É um método não paramétrico de teste, que permite concluir se as amostras têm origem na mesma distribuição. É usado para comparar três ou mais amostras independentes. O equivalente paramétrico deste método é o teste de *one way analysis of variance* (ANOVA), com a diferença de que enquanto a análise de variância dos testes dependem da hipótese de que todas as populações em confronto são independentes e normalmente distribuídas, o teste de Kruskal-Wallis não coloca nenhuma restrição sobre a distribuição.

Quando o teste de Kruskal-Wallis apresenta bons resultados, isso significa que pelo menos uma das amostras é diferente das restantes. O teste não identifica onde é que essa diferença se verifica ou exatamente qual ou quais são as diferenças, apenas nos diz que existe.

O teste de Kruskal-Wallis também é usado quando o grupo de amostras a examinar é de tamanhos diferentes.

A hipótese nula de que todas as populações possuem funções de distribuição iguais é testada por este método, contra a hipótese alternativa de que ao menos duas das populações possuem funções de distribuição diferentes. No âmbito desta dissertação a hipótese nula será

se as distribuições dos parâmetros são semelhantes para a mesma vogal ou, como hipótese alternativa, pertencem a vogais diferentes.

2.7.2 Classificadores

2.7.2.1 Nearest Neighbour

O algoritmo *Nearest Neighbour* (NN) é um método de *machine learning* em que se procura classificar dados desconhecidos ao sistema. Esta classificação tem como base a comparação de dados conhecidos previamente com novos dados. O algoritmo parte de um conjunto de observações, acompanhado das respetivas anotações com informação sobre a classe a que cada observação pertence. Este tipo de aprendizagem, que parte de observações classificadas, denomina-se *supervised learning*. Quando são dados ao algoritmo observações novas, este calcula a distância entre cada ponto novo e os pontos armazenados, de forma a saber qual o vizinho mais próximo (*nearest neighbour*). O ponto é classificado de acordo com a classe do vizinho mais próximo.

Uma das medidas de distância mais utilizadas é a distância Euclidiana:

$$d(x, y) = \|x - y\| = \sqrt{\sum_{i=1}^m (x_i - y_i)^2} \quad (2.13)$$

2.7.2.2 Support Vector Machines

Um outro algoritmo usado no âmbito deste trabalho foi o *Support Vector Machine*. Esta é uma ferramenta de *machine learning*, tal como o NN apresentado atrás e enquadra-se no tipo de aprendizagem supervisionada.

Os dados de treino são compostos por um conjunto de vetores x_i , em que cada um deles contém um certo número de características. Estes vetores são acompanhados por anotações (vetores y_i) que contêm a classificação dos dados de entrada. No caso de os dados pertencerem a duas classes, estas anotações tomam geralmente a forma de $y = +1$ ou $y = -1$.

Considerando ainda o caso mais simples de classificação, em que os dados são separáveis linearmente, a tarefa de aprendizagem consiste em encontrar um hiperplano orientado, tal que os dados $y = +1$ e $y = -1$ se encontrem em lados opostos do hiperplano. Este deve ser escolhido de forma a maximizar a distância relativa a cada uma das classes de pontos. Os vetores de suporte são os pontos do conjunto de dados que se encontram mais próximos do hiperplano. O hiperplano é dado pela expressão:

$$g(x) = w \cdot x + b = 0 \quad (2.14)$$

onde w é o vetor de pesos que determina a orientação do plano e b é o desvio do plano relativamente à origem do espaço de entrada.

Capítulo 3

Procedimento experimental

Neste capítulo serão apresentados de forma sucinta os passos que foram tomados no sentido de se poder fazer o estudo que era proposto como fundamento deste trabalho.

Serão descritas as ferramentas usadas, bem como a base de dados que serviu de base para os diversos testes.

3.1 Ferramentas usadas

3.1.1 Matlab

A principal tecnologia usada no desenvolvimento deste trabalho, no que toca à extração de informação e aplicação de algoritmos nos ficheiros de voz foi o Matlab (*MATrix LABoratory*), que é um software interativo de alta performance voltado para o cálculo numérico. Esta ferramenta integra análise numérica com cálculo de matrizes e processamento de sinais, o que é de relevante importância no âmbito desta dissertação. Tendo em conta que este é um programa de enorme aceitação tanto em ambiente académico como empresarial para estudo e desenvolvimento de algoritmos, torna-se lógica a escolha deste programa.

3.1.2 Weka

O Weka (*Waikato Environment for Knowledge Analysis*) é um *software open-source* de *data mining*, ou exploração de dados para classificação. Este disponibiliza diferentes algoritmos de *machine learning* e reconhecimento de padrões.

Permite importar uma base de dados e fazer o pré-processamento dos dados, classificação, *clustering* e seleção de atributos entre outras funcionalidades. Dado ser um *software open-source*, este vai recebendo atualizações regulares, tendo-se vindo a tornar cada vez mais uma referência em *data mining*.

Este *software* foi usado para obter a classificação das diferentes vogais a partir de diferentes características bem como para perceber quais as mais discriminativas.

3.1.3 IBM SPSS Statistics

IBM SPSS *Statistics* é um *software* de análise estatístico, tal como o Weka. Possui algumas funções idênticas ao programa anteriormente descrito, embora não seja *open-source* e tenha como principal interesse no desenvolvimento deste trabalho o facto de possuir o método de *Kruskal-Wallis* já implementado.

Apresenta-se igualmente como uma referência em *data mining* e permite também importar uma base de dados para análise.

Foi usado para fazer a classificação segundo o método de *Kruskal-Wallis* das amostras a estudar.

3.2 Base de dados

Para se elaborar a base de dados a utilizar durante os testes, recorreu-se a uma base de dados já existente e cedida pelo professor Doutor Aníbal Ferreira usada em estudos anteriores [8]. Esta contém as diferentes vogais presentes na língua portuguesa - /a/, /e/, /i/, /o/ e /u/ - emitidas por um conjunto alargado de oradores (27 crianças e 17 adultos, 11 do sexo feminino e 6 do sexo masculino). Os segmentos presentes nesta base de dados incluem apenas o intervalo estável do sinal, sendo excluídos os intervalos iniciais e finais em que pode haver regiões de transição entre fonemas, entre outros fatores. É seguro assumir que a identificação de vogais presentes nesta base de dados por um ouvinte humano seria de 100% de correção. Cada vogal tem a duração exata de 100 milissegundos.

Para além das vogais presentes na base de dados cedida, foram obtidos mais dois novos sinais a partir destes. Cada sinal de voz foi dividido na sua componente harmónica e

componente de ruído. Este processo foi realizado recorrendo ao uso do algoritmo *Searchtonal*, que é um detetor de *pitch* (frequência fundamental) e da respetiva componente harmónica, e que foi desenvolvido pelo professor Doutor Aníbal Ferreira e tem sofrido melhorias ao longo do tempo pela sua equipa de investigação, de forma a estimar de uma forma robusta e precisa a frequência fundamental e os seus respetivos harmónicos. Este detetor de *pitch* implementa análise cepstral e um banco de regras heurísticas de seleção dos candidatos à frequência fundamental e componentes harmónicas respetivas. Uma vez obtida a componente harmónica, uma simples subtração ao sinal total por este novo sinal dá-nos a componente de ruído.

Note-se ainda que um ficheiro de voz apresentava perturbações e foi por isso excluído da base de dados a utilizar, daí que o total de sinais diferentes a usar seja de 219 ($[44 \text{ oradores} * 5 \text{ vogais}] - 1$) por cada componente (sinal total, sinal harmónico e sinal de ruído).

Capítulo 4

Resultados

Neste capítulo serão apresentados os resultados obtidos dos testes realizados. Foram extraídas diferentes características descritas anteriormente e classificadas recorrendo ao uso do *software* WEKA. Usaram-se dois algoritmos diferentes para proceder à classificação no sentido de obter maior segurança na validade dos resultados.

Além de diferentes características, estas também correspondem a diferentes componentes de sinal de áudio, assim quando uma característica tiver no seu nome um “t”, “h” ou “n” a acompanhá-lo isso corresponde à componente do sinal total, harmónico ou de ruído respetivamente.

Importa referir que os dois classificadores apresentam resultados em termos da taxa de sucesso de classificação das vogais. A diferença dos resultados entre eles explica-se pelas duas perspetivas diferentes que têm, como é referido na secção 2.7, e devido ao diferente comportamento relativamente aos dados.

Para além de um estudo individual de diferentes características, procedeu-se depois ao estudo das mesmas de um modo global e também duas a duas no sentido de perceber a existência, ou não, de alguma combinação que pudesse conduzir a melhores resultados de classificação.

4.1 Resultados individuais

De seguida são apresentados de forma individual os resultados da classificação de cada uma das características. São classificadas também segundo combinações diferentes das

componentes descritas atrás, no sentido de perceber a importância de cada uma destas relativamente à característica em questão no que toca à classificação de vogais.

É usado um número diferente de coeficientes por característica, como se pode observar na tabela 4.1.

Tabela 4.1: Número de coeficientes por característica.

Característica	Número de coeficientes
MFCC	16
LPC	16
LPC-WK	8
NRD	14
PLP	8

O número de coeficientes diz respeito ao número existente por cada componente: total, harmónico e ruído. No caso dos NRDs não existirá obviamente componente de ruído.

O tamanho das frames é igual em todos os casos, 1024 amostras, sendo também a frequência de amostragem, de 32kHz, comum a todos.[8]

4.1.1 MFCC

De seguida são apresentados os resultados da extração e classificação dos coeficientes MFCC.

Tabela 4.2: Classificação das vogais pela característica MFCC com diferentes componentes

Característica	NNGE(%)	SMO(%)
MFCC(t+h+n)	93.6079	99.0868
MFCC(t)	91.7808	98.1735
MFCC(h+n)	91.7808	99.5434
MFCC(h)	91.3242	99.5434
MFCC(n)	88.5845	94.5205

No caso dos MFCCs o conjunto das componentes harmónicas e de ruído é melhor do que cada uma isoladamente, sendo também possível observar que o resultado é próximo da componente do sinal total, como seria expectável visto que ambas as componentes têm origem na divisão desse mesmo sinal.

Os resultados da classificação das vogais presentes na tabela são elevados, o que reforça as indicações da literatura de que os MFCC serão uma das melhores características para este efeito, e daí o seu elevado uso.

4.1.2 LPC

Nas tabelas 4.3 e 4.4 podem-se observar os resultados das classificações dos coeficientes LPC e LPC-WK.

Tabela 4.3: Classificação das vogais pela característica LPC com diferentes componentes

Característica	NNGE(%)	SMO(%)
LPC(t+h+n)	56.621	73.0594
LPC(t)	53.8813	64.3836
LPC(h+n)	50.6849	70.7763
LPC(h)	54.3379	60.274
LPC(n)	39.726	50.2283

Tabela 4.4: Classificação das vogais pela característica LPC-WK com diferentes componentes

Característica	NNGE(%)	SMO(%)
LPC-WK(t+h+n)	85.8447	91.3242
LPC-WK(t)	85.3881	89.0411
LPC-WK(h+n)	84.4749	90.8676
LPC-WK(h)	80.8219	91.3242
LPC-WK(n)	68.0365	59.8174

No estudo dos LPC foram obtidos resultados relativamente pobres, no entanto a classificação a partir dos LPC-WK é francamente superior o que vem reforçar o estudo de [8] onde também o uso do teorema de Wiener-Khintchine resultou numa melhoria dos resultados de classificação. A ideia de que a informação contida no *pitch* tem influência elevada na classificação segundo LPC é aqui reforçada pela melhoria significativa que se observa facilmente nas tabelas (4.3) e (4.4)

4.1.3 NRD

Na classificação dos coeficientes NRD os resultados foram os seguintes

Tabela 4.5: Classificação das vogais pela característica NRD com diferentes componentes

Característica	NNGE(%)	SMO(%)
NRD(t+h)	57.9909	68.0365
NRD(t)	52.0548	63.9269
NRD(h)	61.1872	66.6667

Os NRDs apresentam resultados baixos de correta classificação de vogais, o que pode indicar que a sua vocação não será esta, ou pelo menos sem auxílio de outras características que considerem aspetos diferentes. Os resultados contrariam o que é sugerido por [15] mas também estamos perante condições diferentes. As amostras de áudio têm duração mais curta e o número de frames em análise é também menor. Assim pode-se encontrar justificação para os NRDs não apresentarem valores competitivos relativamente a por exemplo aos MFCCs.

A componente de sinal total apresenta o resultado mais baixo visto conter ruído, e este afetar o seu desempenho. Nesta característica seria esperado que a componente harmónica fosse a que conseguisse obter melhor resultado, como se verificou.

4.1.4 PLP

Na tabela 4.6 podem-se observar os resultados relativos às classificações dos coeficientes PLP.

Tabela 4.6: Classificação das vogais pela característica PLP com diferentes componentes

Característica	NNGE(%)	SMO(%)
PLP(t+h+n)	29.2237	37.4429
PLP(t)	28.7671	32.8767
PLP(h+n)	33.79	36.0731
PLP(h)	25.5708	31.5068
PLP(n)	32.8767	34.2466

No caso dos PLP, os resultados são visivelmente pobres. Foram feitas tentativas de melhorar os resultados, como por exemplo a alteração do número de frames a analisar por

ficheiro, mas que se revelaram infrutíferas. Deste modo fica a ideia de que os PLP não serão minimamente adequados para a tarefa de identificação isolada de vogais em tempo real.

4.2 Resultado de conjuntos de características

Na tabela 4.7 pode-se observar os resultados da classificação usando todas as características ao mesmo tempo, com diferentes componentes em jogo, ou seja, foi classificado MFCC+LPC+LPC-WK+PLP+NRD. Estão discriminadas em cada linha da tabela quais as componentes destas características que entraram no teste.

Tabela 4.7: Classificação das vogais pelo conjunto de características e com diferentes componentes

Componentes	NNGE(%)	SMO(%)
t+h+n	92.6941	99.0868
t	92.2374	99.0868
h	93.6073	98.6301
n	89.0411	94.9772

Só alguns casos dos MFCCs apresentam valores melhores o que indica que haverá de certeza redundância que irá fazer com que estes resultados não sejam os melhores comparativamente a esses casos. Tirando essas exceções, as classificações superam todas as outras características, o que encoraja a ideia de que a combinação de diferentes características irá resultar em melhor aproveitamento.

No caso do ruído os valores sofrem uma descida notória o que em certa parte se pode justificar pelo facto dos NRDs não possuírem essa componente, o que fará com que existam menos amostras para o classificador treinar e testar, resultando nessa diferença.

Tabela 4.8: Classificação das vogais por diferentes combinações de características considerando todas as componentes

Característica	NNGE(%)	SMO(%)
MFCC+LPC	91.3242	99.5434
MFCC+LPC-WK	92.2374	99.5434
MFCC+PLP	93.6073	98.6301
MFCC+NRD	95.4338	99.5434
LPC+LPC-WK	81.7352	90.8676

LPC+PLP	59.8174	74.4292
LPC+NRD	67.1233	77.1689
LPC-WK+PLP	86.3014	89.9543
LPC-WK+NRD	85.8447	93.1507
PLP+NRD	59.8174	67.5799

Na tabela 4.8 estão presentes as combinações duas a duas de todas as características alvo de estudo neste trabalho.

Pode-se observar que os MFCC são de facto a característica com maior peso na classificação das vogais, visto nas combinações onde esta se encontra a classificação ser no mínimo de 91%, significativamente acima da melhor combinação que se encontra a seguir. Note-se também que a combinação MFCC+NRD é a que obtém o melhor resultado. Isto vai de encontro às conclusões obtidas por [15] em que se verifica o mesmo. Este facto pode-se explicar pois estas duas características são ortogonais entre si o que leva a que não tenham redundância de informação e se consiga então obter maior proveito da mesma.

4.3 Escolhas de características

Obtidos os resultados que foram apresentados até agora, houve uma seleção de características no sentido de perceber quais os coeficientes mais discriminativos e tentar melhorar os resultados.

Para proceder à seleção das características fez-se uso do método de Kruskal-Wallis descrito no capítulo 2. Classificaram-se todos os coeficientes extraídos segundo este método e recorreu-se a dois tipos de seleção.

No primeiro caso foi a escolha dos valores mais elevados, ou seja, dado um valor máximo de coeficiente para uma dada característica selecionaram-se os coeficientes que tivessem classificação no máximo 10% abaixo desta. De seguida fez-se a classificação recorrendo aos classificadores já usados anteriormente.

No segundo caso de seleção, ordenaram-se os valores dos coeficientes por cada característica segundo a classificação de Kruskal-Wallis por ordem decrescente, foram desenhados os gráficos correspondentes (que estão presentes em anexo) e selecionaram-se os coeficientes que se encontravam antes de ser notória uma descida abrupta nos gráficos.

Os resultados destes testes são apresentados de seguida.

4.3.1 Escolha de valores mais elevados

Como foi explicado, foram selecionados os coeficientes que, segundo a classificação Kruskal-Wallis, seriam à partida mais discriminativos. Os resultados dessa seleção estão presentes na tabela 4.9.

Tabela 4.9: Classificação das vogais pelas diferentes características após seleção dos melhores coeficientes, segundo o valor mais elevado.

Característica	NNGE(%)	SMO(%)
MFCC(t+h+n)	92.2374	93.6073
MFCC(t)	92.2374	92.6941
MFCC(h)	94.0639	93.1507
MFCC(n)	85.8447	89.0411
LPC(t+h+n)	49.3151	52.5114
LPC(t)	53.4247	49.7717
LPC(h)	50.2283	43.8356
LPC(n)	42.4658	40.6393
LPC-WK(t+h+n)	82.1918	84.0183
LPC-WK(t)	79.9087	78.5388
LPC-WK(h)	79.4521	78.9954
LPC-WK(n)	44.7489	44.2922
PLP(t+h+n)	30.5936	35.6164
PLP(t)	28.7671	32.8767
PLP(h)	24.2009	32.8767
PLP(n)	26.9406	28.3105
NRD(t+h)	40.1826	47.9452
NRD(t)	38.8128	39.726
NRD(h)	32.4201	44.7489

Mais uma vez é visível que os resultados dos MFCCs superam todos os outros. É notório que o classificador SMO é afetado pelo número de amostras que lhe são fornecidas para classificar visto que tem uma queda acentuada nos seus resultados enquanto que o NNGE apresenta melhorias em alguns casos.

Num olhar mais abrangente sobre os resultados verifica-se que na maioria dos casos os resultados pioram com este processo de escolha de coeficientes.

Na tabela 4.10 estão os resultados segundo a mesma seleção mas para o caso em que se consideram o conjunto das características todas ao mesmo tempo, discriminadas as componentes em análise.

Tabela 4.10: Classificação das vogais pelo conjunto de todas as características após seleção dos melhores coeficientes, segundo o valor mais elevado.

Componentes	NNGE(%)	SMO(%)
t+h+n	91.342	97.7169
t	90.8676	96.347
h	90.8676	96.347
n	86.758	89.4977

De forma consistente com o obtido no estudo individualizado, os resultados pioram, e no caso do classificador SMO, os resultados sofrem uma perda mais ligeira, o que reforça a ideia de que este classificador terá dificuldades quando se diminui o número de amostras.

4.3.2 Escolha de valores segundo análise gráfica

Os gráficos usados para esta seleção e os coeficientes escolhidos encontram-se em anexo.

Na tabela 4.11 apresenta-se o resultado da classificação dessa mesma seleção.

Tabela 4.11: Classificação das vogais pelas diferentes características após seleção dos melhores coeficientes, segundo interpretação gráfica.

Característica	NNGE(%)	SMO(%)
MFCC(t+h+n)	92.2374	93.6073
MFCC(t)	92.2374	92.6941
MFCC(h)	94.0639	93.1507
MFCC(n)	85.8447	89.0411
LPC(t+h+n)	56.1644	55.7078
LPC(t)	51.5982	48.8584
LPC(h)	49.3151	45.2055
LPC(n)	42.4658	40.6393
LPC-WK(t+h+n)	84.4749	88.1279
LPC-WK(t)	81.7352	81.2785
LPC-WK(h)	79.4521	78.9954

LPC-WK(n)	69.863	57.9909
PLP(t+h+n)	34.7032	33.79
PLP(t)	26.0274	31.5068
PLP(h)	29.6804	32.4201
PLP(n)	28.7671	28.7671
NRD(t+h)	40.1826	47.9452
NRD(t)	38.8128	39.726
NRD(h)	32.4201	44.7489

De uma forma geral, os resultados obtidos com este tipo de seleção ficam próximos dos obtidos com a seleção dos coeficientes de valor mais elevado analisados em 3.3.1.

Os resultados ficam de forma geral abaixo do observado sem seleção de características e novamente os MFCCs se destacam relativamente às outras.

Na tabela 4.12 apresentam-se os resultados que consideram o conjunto das características todas, discriminadas segundo a componente em análise

Tabela 4.12: Classificação das vogais pelo conjunto de todas as características após seleção dos melhores coeficientes, segundo interpretação gráfica.

Componentes	NNGE(%)	SMO(%)
t+h+n	91.7808	98.1735
t	90.8676	95.4338
h	91.3242	97.2603
n	85.3881	89.9543

Os resultados são novamente semelhantes aos anteriores. Contudo é de notar uma ligeira melhoria, o que tendo em conta os valores individuais obtidos, que neste caso são na globalidade inferiores aos selecionados segundo os valores 10% mais próximos do máximo, parece indicar que alguns coeficientes com valores mais baixos deverão ter influência na classificação das vogais.

4.4 Conclusões

Neste capítulo foram apresentados os resultados obtidos da extração e classificação de diferentes características com o objetivo de identificar vogais.

Foram usadas 5 características diferentes, com particular destaque para a performance dos MFCCs que são os mais frequentes no estado de arte e que neste estudo

saem reforçados como sendo de facto os que possuem melhor capacidade de discriminação de vogais. Nota ainda para o facto da combinação de MFCCs e NRDs ter sido onde se obteve melhores resultados de entre as várias combinações efetuadas e que vai de encontro ao que é sugerido por [15] de que esta combinação poderá ser de facto a melhor para este objetivo.

Quanto à escolha de características, uma vez mais os MFCCs destacam-se dos demais embora exista um decréscimo de performance quando combinados com outros. Importa notar que os MFCCs foram os únicos a beneficiar da escolha de características o que sugere que existe uma forte preponderância de um número reduzido de coeficientes sobre os restantes no que toca à classificação de vogais. Dado a classificação piorar com o agrupamento de características, fica a ideia de que poderá não haver vantagem em termos de eficácia ao reduzir o número de coeficientes, ou que então deve ser encontrada outra forma de seleção.

Capítulo 5

Identificação do orador

Para além da identificação de vogais, foi feito mais um estudo exploratório no sentido de perceber a capacidade discriminatória da combinação dos MFCCs e dos NRDs, bem como conhecer o benefício da separação das componentes harmónicas e de ruído, relativamente ao caso em que temos de classificar oradores.

Nos testes realizados no capítulo 4 foi notório que os MFCCs eram os que mais se destacavam e que quando combinados com os NRDs, obtiveram o melhor resultado de todos os recolhidos. Assim surgiu o interesse em perceber como se comportariam estes se a sua missão não fosse a de identificar vogais mas sim de identificar o orador.

Este estudo foi facilitado pelo facto de a base de dados usada nos testes anteriores para a classificação de vogais já ter a distinção dos oradores presente nela proporcionando a oportunidade para os testes desejados.

Assim foram realizados dois testes: o primeiro com todos os oradores presentes na base de dados (6 homens, 11 mulheres e 27 crianças) e um segundo teste em que foram analisados nas 3 classes diferentes, de forma individual. No segundo teste foram usados apenas 6 oradores para os 3 casos visto que só existiam 6 homens na base de dados e para criar condições idênticas de teste para as 3 classes.

Por se ter verificado de forma consistente que o classificador SMO era afetado fortemente pela redução do número de amostras, decidiu-se proceder daqui para a frente à classificação recorrendo apenas ao classificador NNGE.

5.1 Identificação com todos os oradores

Feita a extração das características associadas a cada orador, ao invés de cada vogal como vinha a ser feito até aqui, procedeu-se à classificação das mesmas como se pode observar na tabela 4.13

Tabela 4.13: Classificação de todos os oradores pelas diferentes características e componentes.

Característica	NNGE(%)
MFCC(t+h+n)+NRD(t+h)	65
MFCC(t)+NRD(t)	62.5571
MFCC(h)+NRD(h)	64.9924
MFCC(t+h+n)	84.4749
MFCC(t)	84.9315
MFCC(h+n)	80.2131
MFCC(h)	88.2801
MFCC(n)	51.1416
NRD(t+h)	26.0274
NRD(t)	31.1644
NRD(h)	29.9087

Os MFCCs destacam-se sem surpresa como os mais discriminativos, embora não sejam tão capazes de identificar o orador como foram anteriormente de identificar a vogal. Comparativamente também aos testes anteriores, os MFCC+NRD desta vez não se revelam como os melhores, o que parece indicar que o propósito desta combinação será mais a identificação da vogal do que orador. Os NRDs seguem a mesma linha dos MFCCs e também obtêm piores resultados.

5.2 Identificação de oradores por classes

Feito o teste com todas as classes, pretende-se agora perceber se o facto de estarem presentes diferentes classes de oradores afeta negativamente os resultados obtidos. Deste modo foram testados individualmente 6 oradores masculinos, femininos e crianças.

5.2.1 Homens

Na tabela 4.14 estão os resultados para a classificação masculina

Tabela 4.14: Classificação dos oradores masculinos pelas diferentes características e componentes.

Característica	NNGE(%)
MFCC(t+h+n)+NRD(t+h)	75.5556
MFCC(t)+NRD(t)	78.8889
MFCC(h)+NRD(h)	76.6667
MFCC(t+h+n)	75.5556
MFCC(t)	83.3333
MFCC(h+n)	80
MFCC(h)	82.2222
MFCC(n)	66.6667
NRD(t+h)	35.5556
NRD(t)	33.3333
NRD(h)	31.1111

Observando os resultados, verifica-se que o principal beneficiado desta divisão foi a junção MFCC+NRD. Os MFCCs continuam a obter a melhor classificação mas sofrem uma perda ainda que pequena. Os NRDs melhoram o que com certeza terá contribuído para a melhoria na junção das características.

5.2.2 Mulheres

Na classificação das oradoras femininas os resultados foram os seguintes

Tabela 4.15: Classificação dos oradores femininos pelas diferentes características e componentes.

Característica	NNGE(%)
MFCC(t+h+n)+NRD(t+h)	77.778
MFCC(t)+NRD(t)	68.8889
MFCC(h)+NRD(h)	73.3333
MFCC(t+h+n)	86.6667
MFCC(t)	88.8889

MFCC(h+n)	84.4444
MFCC(h)	82.2222
MFCC(n)	50
NRD(t+h)	27.778
NRD(t)	30
NRD(h)	30

A tendência da evolução dos resultados mantém-se dos homens, com uma diferença nos MFCCs onde os valores melhoram relativamente ao conjunto das classes.

5.2.3 Crianças

Na tabela 4.16 podem-se observar as classificações das características relativas às vozes de crianças

Tabela 4.16: Classificação das vozes de crianças pelas diferentes características e componentes.

Característica	NNGE(%)
MFCC(t+h+n)+NRD(t+h)	67.778
MFCC(t)+NRD(t)	73.3333
MFCC(h)+NRD(h)	68
MFCC(t+h+n)	77.778
MFCC(t)	81.1111
MFCC(h+n)	78.8889
MFCC(h)	80
MFCC(n)	54.4444
NRD(t+h)	25.5556
NRD(t)	24.4444
NRD(h)	24.4444

De novo os MFCCs se destacam, embora com resultados significativamente mais baixos do que anteriormente. O facto de estarmos a lidar com vozes de crianças, que são por natureza irregulares e de difícil identificação, não será alheio ao decréscimo da classificação de forma global.

5.3 Conclusões

Neste capítulo estudou-se a classificação de oradores ao invés da classificação de vogais que era objetivo inicial desta dissertação e ponto principal de foco.

Começou-se por analisar de forma indiscriminada os oradores existentes na base de dados segundo a característica que tinha obtido melhores resultados nas vogais, os MFCCs, e também a característica que em conjunto com esta tinha conseguido obter a percentagem de correção mais elevada, os NRDs.

De seguida, foi feita a classificação segundo as diferentes classes de oradores: homens, mulheres e crianças.

Pode-se observar que os resultados nos homens e nas crianças pioraram de forma global em relação ao conjunto das classes. Nas mulheres os MFCCs conseguiram obter classificação ainda mais alta.

O maior beneficiado com a divisão das classes e análise individual foi a combinação MFCC+NRD onde se verificou uma melhoria bastante significativa.

Concluiu-se que os NRDs são sensíveis à rarefação dos dados, ou seja, são prejudicados quando o treino é realizado com poucos exemplos de cada orador, não conseguindo formar clusters bem definidos. Esta dificuldade de treino verificada nos NRDs foi responsável pelos baixos resultados obtidos na combinação das características.

Testes para comprovar ou fazer cair estas ideias poderão ser realizados tais como a redução do número de coeficientes extraídos, pois fica a ideia que poderá haver má seleção da informação durante a extração dos coeficientes do NRDs, que poderá estar a introduzir dispersão nos dados levando aos resultados de desempenho fraco que se verificaram.

Capítulo 6

Conclusão

O trabalho realizado ao longo da presente dissertação teve por objetivo a identificação das diferentes vogais existentes na língua Portuguesa: /a/, /e/, /i/, /o/ e /u/.

Pretendia-se conseguir chegar a um bom resultado de identificação mas também estudar a importância de diferentes características neste processo bem como da relevância de dividir um sinal de voz nas suas componentes harmónica e de ruído.

Para se conseguir chegar à identificação das vogais, procedeu-se ao estudo de diferentes características que são indicadas no estado da arte como capazes de o fazer. Foi obtida uma base de dados com diferentes oradores e as diferentes vogais emitidas pelos mesmos, sendo que se procedeu ao tratamento desta base de dados no sentido de obter as componentes de sinal harmónico e de ruído.

Recorrendo a um *software* de análise estatística, o Weka, obteve-se a classificação das diferentes vogais. Observou-se que os MFCCs obtiveram os melhores resultados o que justifica o facto de serem os mais populares neste tipo de análises. Outro facto importante a reportar foi que a combinação de MFCCs com NRDs obteve um poder de discriminação superior e que vai de encontro a estudos já citados e que era uma das expectativas deste estudo.

Foi feita uma seleção de coeficientes por cada característica no sentido de perceber o peso de cada um, e tentar obter uma classificação mais elevada. Esta seleção teria particular relevância numa aplicação que fizesse identificação das vogais em tempo real, permitindo desde logo filtrar a informação mais importante a reter. As classificações não sofreram melhorias na maioria dos casos. Algumas razões para isto poderão ser os métodos

escolhidos para esta seleção ou então a dificuldade que os próprios classificadores podem enfrentar ao não lhes serem fornecidas tantas amostras para treino.

Num aspeto mais exploratório e derivado de estudos anteriores que indicavam a capacidade dos NRDs de identificação de orador, tirou-se partido da base de dados obtida para tentar perceber a combinação dos NRDs com os MFCCs no caso da identificação do orador, e também identificar a existência de vantagem, ou não, da separação da componente harmónica da de ruído.

Os resultados não apresentaram vantagens, e a principal conclusão a retirar foi que os NRDs apresentam dificuldade em obter boa classificação quando não lhes são fornecidos amostras suficientes de treino. Nas condições presentes neste estudo, não será possível obter vantagem na combinação de MFCCs e NRDs, para a identificação de orador.

6.1 Trabalho futuro

Um dos resultados mais positivos a retirar deste estudo são os resultados superiores apresentados pela combinação de MFCCs com NRDs. Seria portanto interessante observar o comportamento desta combinação na situação de identificação de vogais em tempo real, nomeadamente, na criação de uma aplicação para identificação de vogais em tempo real onde seria possível comparar e possivelmente comprovar a superioridade desta combinação relativamente a outras combinações de características, bem como em comparação com a individualidade das características.

No que toca ao estudo feito para a identificação do orador, dados os resultados obtidos, uma hipótese de os melhorar será a recolha de um maior número de frames por ficheiro, bem como de mais exemplos de voz por cada orador de modo a fornecer mais exemplos de treino para os NRDs. É esperado que a classificação dos NRDs melhore e, por consequência, a combinação com MFCCs também evolua positivamente.

Anexo A

Anexo A - Seleção de características

A.1 Resultados dos testes Kruskal-Wallis

Tabela A.1: Classificação de todos os coeficientes de todas as características segundo o método de Kruskal-Wallis.

Característica	Score
MFCC_t1	170.270
MFCC_t2	177.713
MFCC_t3	175.469
MFCC_t4	86.813
MFCC_t5	98.535
MFCC_t6	24.489
MFCC_t7	15.643
MFCC_t8	61.838
MFCC_t9	3.952
MFCC_t10	8.254
MFCC_t11	15.259
MFCC_t12	23.713
MFCC_t13	17.253
MFCC_t14	5.650
MFCC_t15	23.579
MFCC_t16	20.947
MFCC_h1	168.684
MFCC_h2	178.047
MFCC_h3	179.167
MFCC_h4	82.225

MFCC_h5	97.490
MFCC_h6	18.126
MFCC_h7	16.590
MFCC_h8	61.496
MFCC_h9	6.084
MFCC_h10	6.810
MFCC_h11	15.188
MFCC_h12	18.151
MFCC_h13	14.068
MFCC_h14	6.017
MFCC_h15	22.782
MFCC_h16	17.354
MFCC_n1	174.722
MFCC_n2	167.706
MFCC_n3	169.201
MFCC_n4	68.614
MFCC_n5	85.284
MFCC_n6	9.022
MFCC_n7	26.740
MFCC_n8	35.126
MFCC_n9	10.069
MFCC_n10	5.694
MFCC_n11	14.885
MFCC_n12	13.695
MFCC_n13	13.335
MFCC_n14	6.960
MFCC_n15	11.425
MFCC_n16	4.526
LPC_t1	137.791
LPC_t2	125.818
LPC_t3	113.193
LPC_t4	46.539
LPC_t5	89.871
LPC_t6	13.458
LPC_t7	74.851
LPC_t8	18.650
LPC_t9	66.714
LPC_t10	33.101

LPC_t11	23.008
LPC_t12	40.244
LPC_t13	37.671
LPC_t14	47.584
LPC_t15	54.000
LPC_t16	58.083
LPC_h1	125.793
LPC_h2	115.863
LPC_h3	102.942
LPC_h4	46.668
LPC_h5	82.479
LPC_h6	5.383
LPC_h7	75.876
LPC_h8	15.621
LPC_h9	40.300
LPC_h10	20.441
LPC_h11	23.869
LPC_h12	43.741
LPC_h13	16.656
LPC_h14	47.913
LPC_h15	56.495
LPC_h16	64.007
LPC_n1	108.663
LPC_n2	98.066
LPC_n3	65.660
LPC_n4	39.161
LPC_n5	56.700
LPC_n6	2.839
LPC_n7	54.891
LPC_n8	13.259
LPC_n9	34.266
LPC_n10	17.616
LPC_n11	11.646
LPC_n12	20.790
LPC_n13	26.866
LPC_n14	14.241
LPC_n15	9.660
LPC_n16	24.458

LPC-WK_t1	163.728
LPC-WK_t2	114.345
LPC-WK_t3	72.139
LPC-WK_t4	40.656
LPC-WK_t5	155.554
LPC-WK_t6	173.999
LPC-WK_t7	57.236
LPC-WK_t8	48.746
LPC-WK_h1	163.352
LPC-WK_h2	121.633
LPC-WK_h3	74.232
LPC-WK_h4	40.904
LPC-WK_h5	155.051
LPC-WK_h6	170.722
LPC-WK_h7	67.602
LPC-WK_h8	53.198
LPC-WK_n1	66.194
LPC-WK_n2	79.781
LPC-WK_n3	75.293
LPC-WK_n4	64.657
LPC-WK_n5	38.489
LPC-WK_n6	33.349
LPC-WK_n7	26.674
LPC-WK_n8	28.738
PLP_t1	33.563
PLP_t2	34.113
PLP_t3	34.716
PLP_t4	36.950
PLP_t5	39.287
PLP_t6	37.062
PLP_t7	36.723
PLP_t8	35.604
PLP_h1	32.347
PLP_h2	30.408
PLP_h3	30.043
PLP_h4	33.129
PLP_h5	35.028
PLP_h6	34.058

PLP_h7	33.940
PLP_h8	34.010
PLP_n1	17.571
PLP_n2	37.501
PLP_n3	48.713
PLP_n4	41.938
PLP_n5	42.990
PLP_n6	41.417
PLP_n7	39.351
PLP_n8	28.485
NRD_t1	57.058
NRD_t2	73.084
NRD_t3	79.848
NRD_t4	43.746
NRD_t5	50.599
NRD_t6	38.556
NRD_t7	37.170
NRD_t8	37.874
NRD_t9	30.855
NRD_t10	29.703
NRD_t11	29.801
NRD_t12	30.982
NRD_t13	32.441
NRD_t14	32.151
NRD_h1	70.541
NRD_h2	82.601
NRD_h3	88.752
NRD_h4	49.614
NRD_h5	60.270
NRD_h6	49.199
NRD_h7	49.069
NRD_h8	51.322
NRD_h9	42.634
NRD_h10	39.201
NRD_h11	38.655
NRD_h12	42.174
NRD_h13	39.598
NRD_h14	39.838

A.2 Seleção dos melhores coeficientes

A.2.1 Seleção dos valores mais altos

A.2.1.1 MFCC

Tabela A.2: Coeficientes selecionados segundo o valor mais alto da característica MFCC

Tipo de sinal	Coeficientes
Total	1,2,3
Harmónico	1,2,3
Ruído	1,2,3

A.2.1.2 LPC

Tabela A.3: Coeficientes selecionados segundo o valor mais alto da característica LPC

Tipo de sinal	Coeficientes
Total	1,2
Harmónico	1,2
Ruído	2,3

A.2.1.3 LPC-WK

Tabela A.4: Coeficientes selecionados segundo o valor mais alto da característica LPC-WK

Tipo de sinal	Coeficientes
Total	1,6
Harmónico	1,5,6
Ruído	2,3

A.2.1.4 PLP

Tabela A.5: Coeficientes selecionados segundo o valor mais alto da característica PLP

Tipo de sinal	Coeficientes
Total	4,5,6,7,8
Harmónico	1,4,5,6,7,8
Ruído	3

A.2.1.5 NRD

Tabela A.6: Coeficientes selecionados segundo o valor mais alto da característica NRD

Tipo de sinal	Coeficientes
Total	2,3
Harmónico	2,3

A.2.2 Seleção dos valores por análise gráfica

A.2.2.1 MFCC

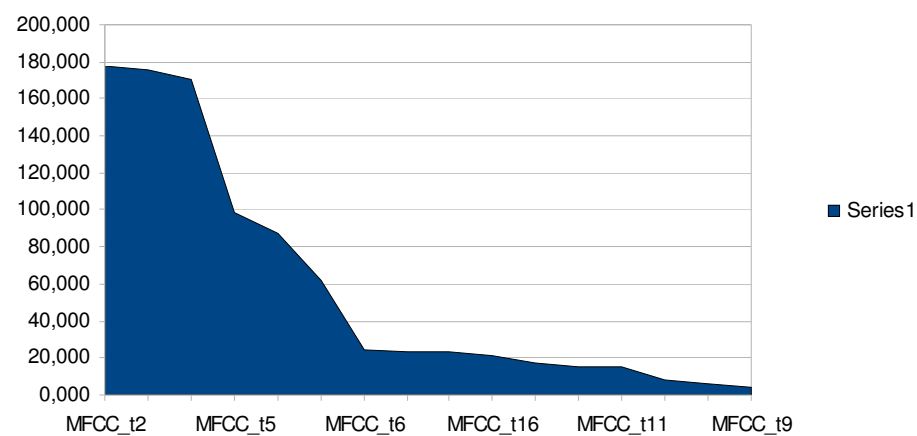


Figura A.1: Representação gráfica dos coeficientes da componente total dos sinais de voz da característica MFCC

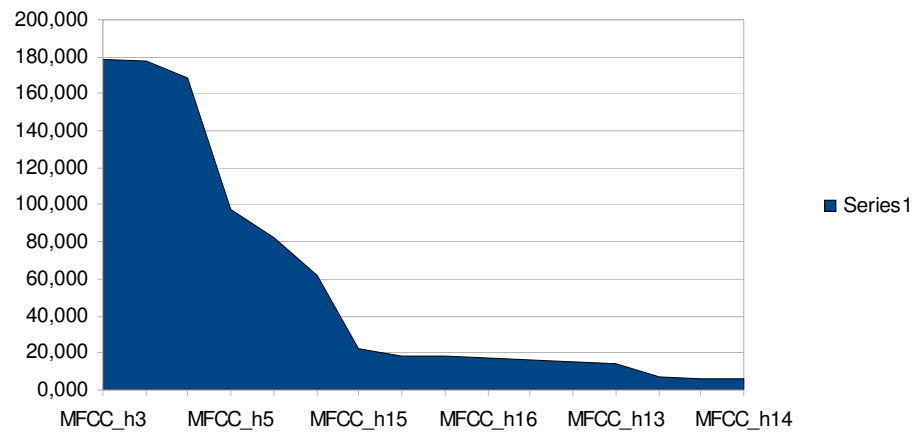


Figura A.2: Representação gráfica dos coeficientes da componente harmônica dos sinais de voz da característica MFCC

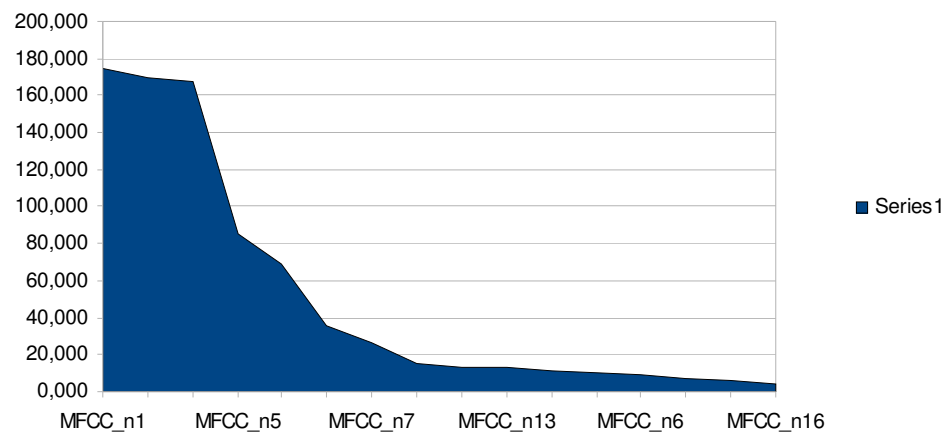


Figura A.3: Representação gráfica dos coeficientes da componente de ruído dos sinais de voz da característica MFCC

Tabela A.7: Coeficientes selecionados segundo análise gráfica da característica MFCC

Tipo de sinal	Coeficientes
Total	1,2,3
Harmónico	1,2,3
Ruído	1,2,3

A.2.2.2 LPC

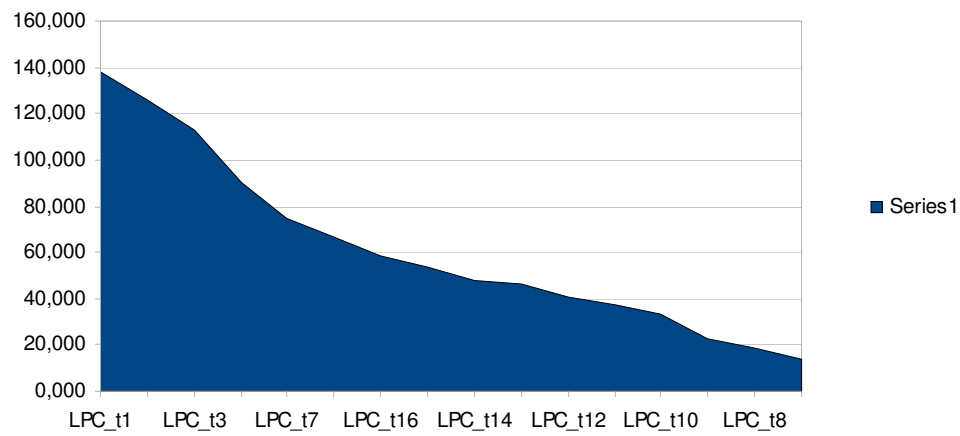


Figura A.4: Representação gráfica dos coeficientes da componente total dos sinais de voz da característica LPC

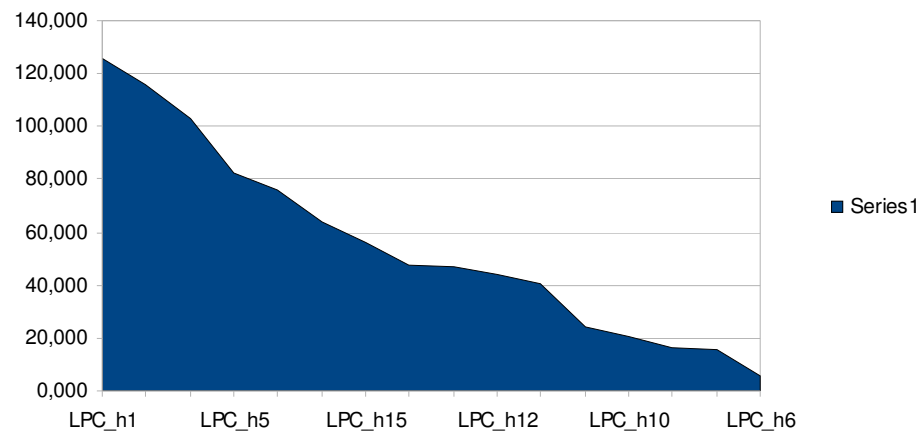


Figura A.5: Representação gráfica dos coeficientes da componente harmónica dos sinais de voz da característica LPC

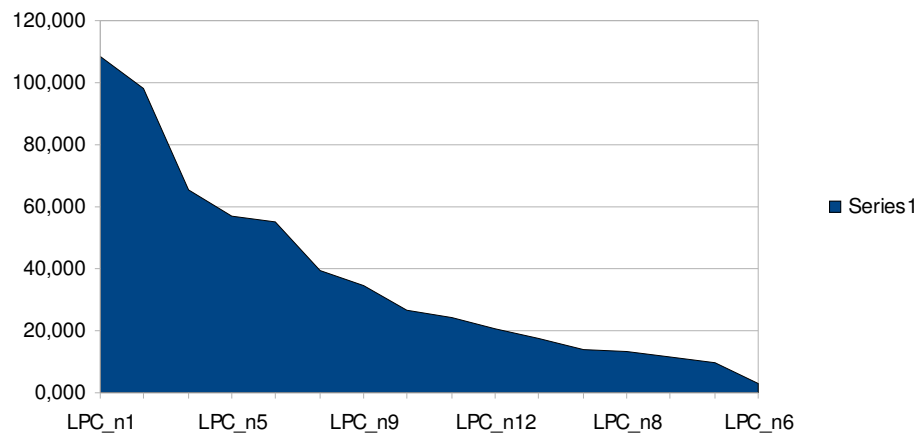


Figura A.6: Representação gráfica dos coeficientes da componente de ruído dos sinais de voz da característica LPC

Tabela A.8: Coeficientes selecionados segundo análise gráfica da característica LPC

Tipo de sinal	Coeficientes
Total	1,2,3
Harmónico	1,2,3
Ruído	1,2

A.2.2.3 LPC-WK

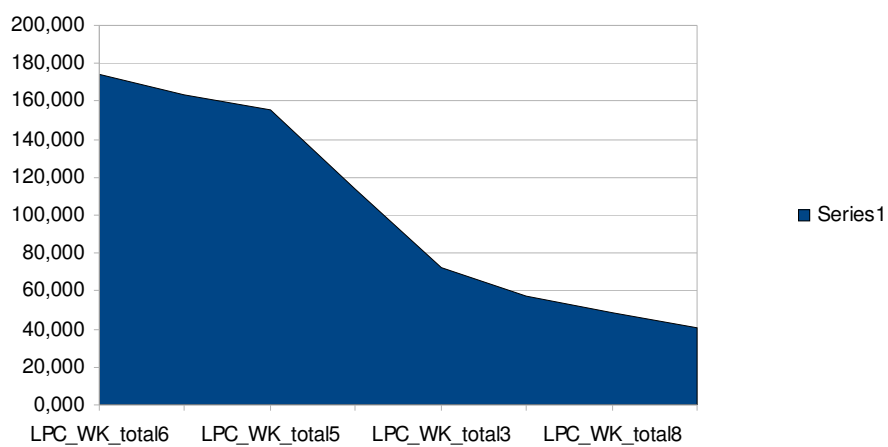


Figura A.7: Representação gráfica dos coeficientes da componente total dos sinais de voz da característica LPC-WK

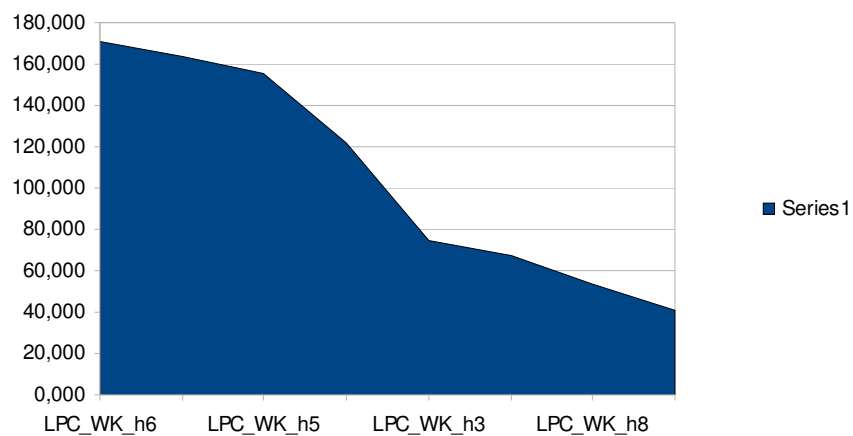


Figura A.8: Representação gráfica dos coeficientes da componente harmónica dos sinais de voz da característica LPC-WK

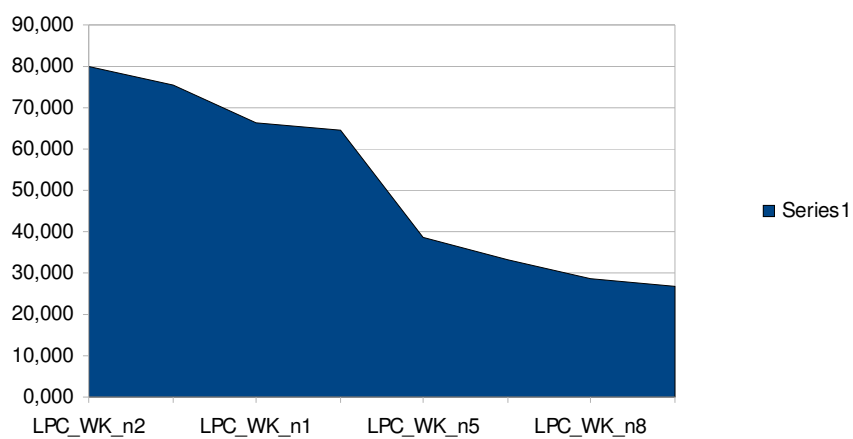


Figura A.9: Representação gráfica dos coeficientes da componente de ruído dos sinais de voz da característica LPC-WK

Tabela A.9: Coeficientes seleccionados segundo análise gráfica da característica LPC-WK

Tipo de sinal	Coeficientes
Total	1,5,6
Harmónico	1,5,6
Ruído	1,2,3,4

A.2.2.4 PLP

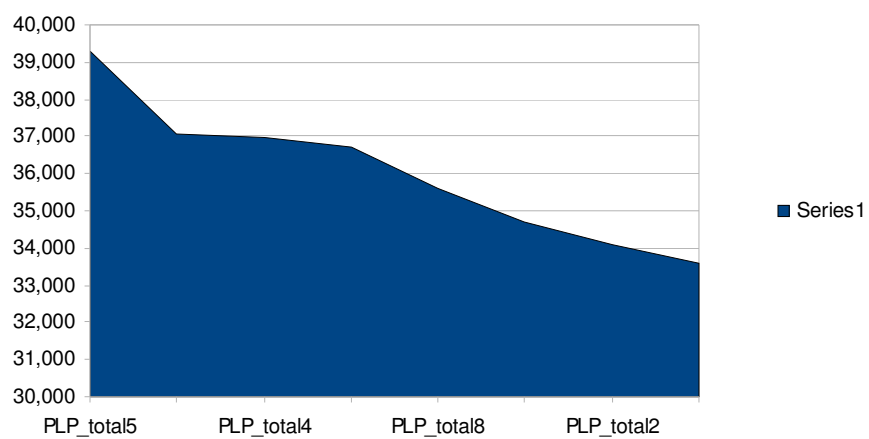


Figura A.10: Representação gráfica dos coeficientes da componente total dos sinais de voz da característica PLP

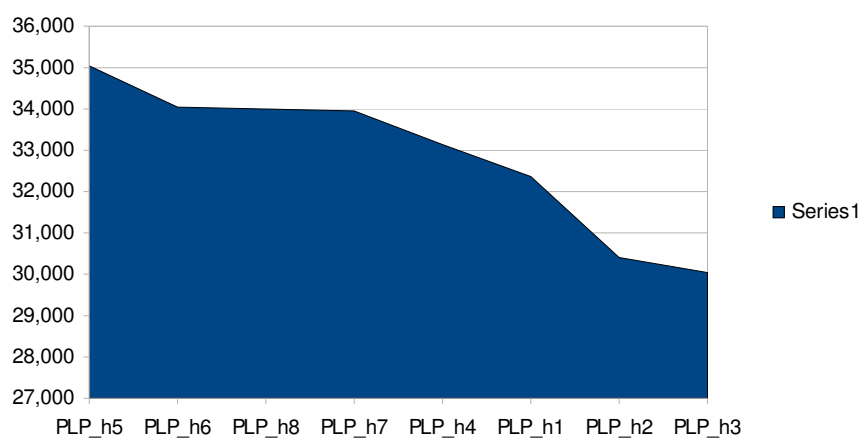


Figura A.11: Representação gráfica dos coeficientes da componente harmónica dos sinais de voz da característica PLP

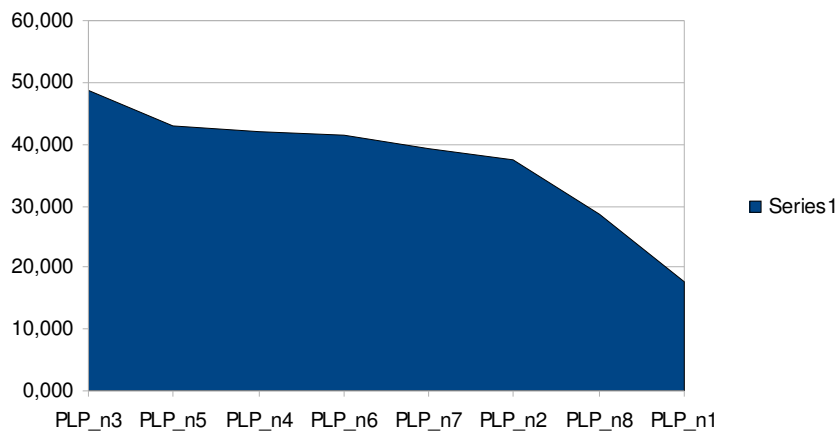


Figura A.12: Representação gráfica dos coeficientes da componente de ruído dos sinais de voz da característica PLP

Tabela A.10: Coeficientes selecionados segundo análise gráfica da característica PLP

Tipo de sinal	Coeficientes
Total	4,5,6,7
Harmónico	5,6,7,8
Ruído	2,3,4,5,6,7

A.2.2.5 NRD

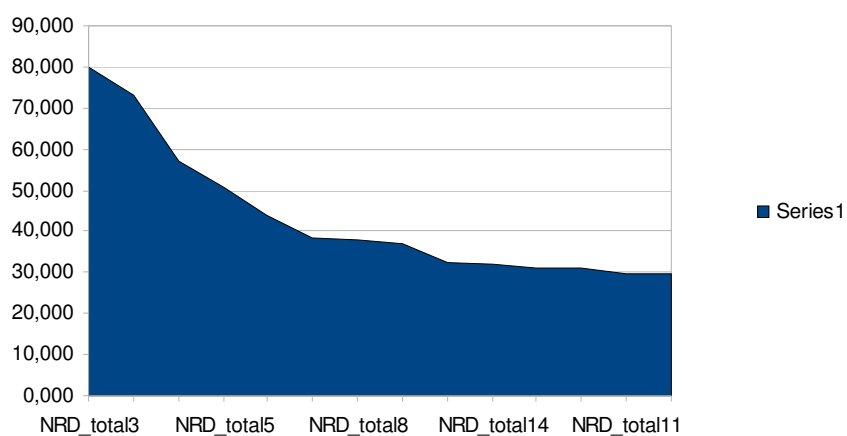


Figura A.13: Representação gráfica dos coeficientes da componente total dos sinais de voz da característica NRD

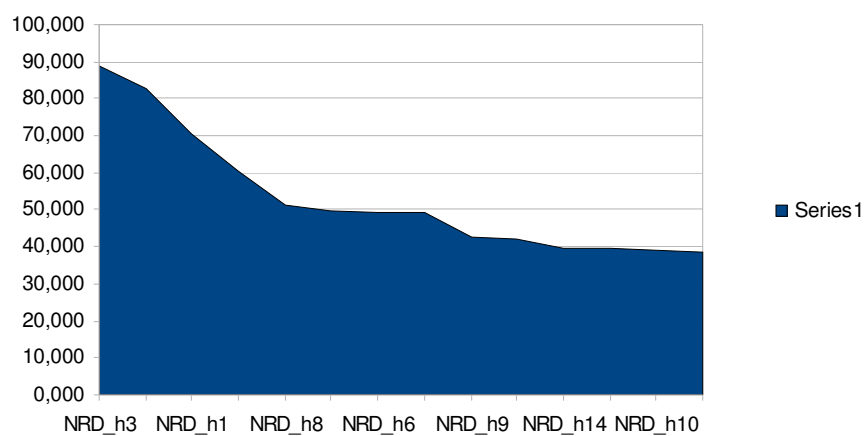


Figura A.14: Representação gráfica dos coeficientes da componente harmônica dos sinais de voz da característica NRD

Tabela A.11: Coeficientes selecionados segundo análise gráfica da característica NRD

Tipo de sinal	Coeficientes
Total	2,3
Harmónico	2,3

Referências

- [1] Escobar, J. and Partridge, C. . Flow synchronization protocol. IEEE/ACM Transactions on Networking 2(2): 111-121, 1994.
- [2] Moore, B. C. J. (2003). An introduction to the psychology of hearing. Amsterdam: Academic Press.
- [3] Ferrand, C. T. (2007). *Speech Science: An Integrated Approach to Theory and Clinical Practice* (Allyn and Bacon, Boston).
- [4] Ferreira, A.J.S. . Automatic Recognition of Isolated Vowels Using F0-Normalized Harmonic Features. ICETE 2008, CCIS 48: 307-317, 2009.
- [5] Baumann O, Berlin P. Perceptual scaling of voice identity: common dimensions for different vowels and speakers. Psychological Research 74(1): 110-120, 2009
- [6] Kumar Sunil R. K., Lajish V. L. (2012). Phoneme recognition using zerocrossing interval distribution of speech patters and ANN. Int J Speech Technol, 16: 125-131, 2013
- [7] Giacobello D., Christensen M. G., Murthi M. N., Jensen S. H., Moonen M. (2012). Sparse linear prediction and its applications to speech processing. IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL. 20, NO. 5: 1644-1657, 2012
- [8] Ferreira, A.J.S. . Static features in real-time recognition of isolated vowels at high pitch. Journal of the Acoustical Society of America, 122(4): 2389-2404, 2007
- [9] Klatt, D. H. (1982). Prediction of perceived phonetic distance from critical band spectra - a first step. Acoustics, Speech, and Signal Processing, IEEE International Conference: 1278-1281, 1982.
- [10] Sebenta Processamento digital da fala - INSTITUTO SUPERIOR DE ENGENHARIA DE LISBOA - disponível em: <http://www.deetc.isel.ipl.pt/comunicacoes/disciplinas/pdf/>
- [11] Bernd T. Meye, Thomas Brand, and Birger Kollmeier (2010). Effect of Speech-intrinsic variations on human and automatic recognition of spoken phonemes. Journal of the Acoustical Society of America, 129(1): 388-403, 2011
- [12] Peter F. Assmann, Terrance M. Nearey. Identification of frequency-shifted vowels. Journal of the Acoustical Society of America, 124(5): 3203-3212, 2008
- [13] G. Fant. Acoustical theory of speech production. Netherlands: Mouton and Co; 1960.

- [14] Huang X, Acero A, Hon H. Spoken Language Processing. A guide to theory, algorithm, and system development. New Jersey: Prentice-Hall Inc.; 2001.
- [15] Ricardo Sousa, Aníbal Ferreira. Singing voice analysis using relative harmonic delays. Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH: 1997-2000, 2011.