FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Reconhecimento de Orador em Dois Segundos

Diana Rocha Mendes

Mestrado Integrado em Engenharia Electrotécnica e de Computadores

Orientador: Aníbal Ferreira (Prof. Dr.)

Julho de 2011

© Diana Rocha Mendes, 2011

Resumo

Nas últimas décadas tem-se vindo a verificar um aumento da popularidade dos sistemas biométricos, assim como o aumento do interesse em outros sistemas para além dos já estabelecidos, como o reconhecimento da impressão digital. Um desses métodos é o de reconhecimento de orador. Actualmente esta tecnologia já é utilizada em diversos sectores, desde *telephone banking* e outros sistemas de autenticação remota a aplicações forenses. Apesar dos sistemas actuais apresentarem elevados níveis de desempenho, certas situações de identificação ainda constituem um desafio para estes, nomeadamente cenários de teste em que os segmentos de voz têm duração reduzida – na ordem de poucos segundos.

Esta dissertação tem por objectivo estudar o desempenho dos sistemas do estado da arte nas condições mencionadas e explorar diferentes métodos que melhorem a sua robustez. Com este fim são exploradas novas características da voz, com forte poder discriminatório de orador, que conjuntamente com outras características do estado da arte permitem obter superior desempenho nos cenários descritos. Estas novas características estudadas denominam-se *Normalized Relative Delays* e extraem informação relacionada com a fase dos harmónicos da voz.

Foi implementado um sistema de reconhecimento de orador baseado em *Gaussian Mixture Models*, que se utilizou como plataforma para o teste das características mencionadas. Adicionalmente foi também utilizado um segundo método de classificação, baseado no método *Nearest Neighbour*. Tendo reunido diversos resultados entre os sistemas, verificou-se um aumento da robustez da identificação de orador em algumas situações de teste. Por fim descreve-se na presente dissertação a aplicação do sistema baseado em *Gaussian Mixture Models* e *Mel-Frequency Cepstral Coefficients* a um caso real de identificação de orador proposto à Universidade do Porto.

Esta dissertação enquadrou-se nas actividades do projecto de investigação "Assistive Real-Time Technology in Singing", financiado pela Fundação para a Ciência e Tecnologia, com a referência PTDC/SAU-BEB/104995/2008.

Abstract

In the last decades there has been an increase in popularity of biometric systems, as well as an interest in other new biometric systems besides the standard ones, such as fingerprint recognition. One of these growing methods is speaker recognition. This technology is currently being used in several sectors, from telephone banking and other remote authentication systems to forensic applications. Although state-of-the-art systems achieve a high level performance, some identification scenarios still constitute a challenge, particularly when the voice segments available to perform the identification task are fairly short – a few seconds.

This dissertation aims to study the performance of state-of-the-art systems in the aforementioned conditions and to explore different methods in order to improve their robustness. In this work, novel highly discriminative voice features are studied, so that the combined use of these features with state-of-the-art ones, improves the performance of speaker recognition systems. The proposed features are Normalized Relative Delays, and they are based on the extraction of information on the voice harmonics' phase.

A speaker recognition system based on Gaussian Mixture Model was implemented during this dissertation, and it was used as a testing platform for the characteristics mentioned above. Another speaker classification technique, based on the Nearest Neighbour algorithm was also used. Having collected results from both solutions, it can be shown an improvement of speaker identification robustness in certain identification scenarios. Finally, the application of a Gaussian Mixture Models and Mel-Frequency Cepstral Coefficients based system on a speaker identification case study (proposed to University of Porto) is presented.

This dissertation was incorporated in the research project "Assistive Real-Time Technology in Singing", supported by Fundação para a Ciência e Tecnologia, with reference PTDC/SAU-BEB/104995/2008.

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 Ricardo Sousa pela disponibilidade constante e pela ajuda prestada durante a realização do trabalho. Não posso também deixar de agradecer a todos os presentes todos os dias na sala de trabalho, pelo ambiente proporcionado e pelo espírito de colaboração criado.

A todos os meus amigos e colegas de curso por me animarem nos momentos que precisei e pelo intercâmbio de ideias e informação para a elaboração desta dissertação. Aproveito também para agradecer aos meus pais e à minha família pelos seus continuados esforços ao longo da realização do curso e pelo apoio que prestaram.

Por último um grande obrigado ao Telmo pela motivação ao longo destes últimos anos e no período de dissertação, tanto na ajuda dada para a escrita desta como todo o apoio emocional.

Diana Mendes

Conteúdo

1 Introdução **1**
1.1 Reconhecimento de Orador . . . 1
1.2 Aplicações . . . 1
1.3 Contextualização e Objectivos . . . 2
1.4 Estrutura da Dissertação . . . 3

2 Estado da Arte **5**
2.1 Conceitos Fundamentais . . . 5
2.1.1 Tracto vocal e produção de fala . . . 5
2.1.2 Modelo fonte-filtro . . . 7
2.2 Reconhecimento de Orador . . . 7
2.2.1 Medidas de desempenho . . . 10
2.3 Extracção de Características . . . 10
2.3.1 *Linear Predictive Coding* . . . 10
2.3.2 *Mel-frequency Cepstral Coefficients* . . . 11
2.3.3 Alternativas e Conclusões . . . 13
2.4 *Pattern Matching* . . . 14
2.4.1 *Nearest Neighbour* . . . 15
2.4.2 *Gaussian Mixture Models* . . . 15
2.4.3 *Support Vector Machines* . . . 17
2.4.4 Alternativas e Conclusões . . . 19
2.5 Desempenho dos Sistemas Actuais . . . 20
2.6 Conclusões . . . 24

3 Sistema de Reconhecimento de Orador **27**
3.1 Introdução . . . 27
3.2 *Gaussian Mixture Models* . . . 28
3.2.1 O algoritmo *Expectation Maximization* . . . 28
3.3 *Normalized Relative Delays* . . . 30
3.3.1 *Normalized Relative Delays* . . . 30
3.3.2 Estudos preliminares . . . 32
3.4 Implementação . . . 32
3.4.1 Ferramentas utilizadas . . . 33
3.4.2 Extracção de características MFCC . . . 35
3.4.3 Modelação e classificação GMM . . . 35
3.5 Conclusão . . . 36

4 Testes e Resultados **39**
4.1 Apresentação dos testes realizados 39
4.1.1 Organização dos testes 39
4.2 Bases de Dados 40
4.2.1 TIMIT 40
4.2.2 Vogais Cantadas 41
4.2.3 Vogais Faladas 41
4.3 Desempenho das características MFCC 41
4.3.1 Normalização *Cepstral Mean Subtraction* e MFCCs diferenciais 42
4.3.2 Desempenho do sistema utilizando voz completa 45
4.3.3 Desempenho do sistema utilizando segmentação da voz de acordo com vozeamento 46
4.4 Desempenho das características NRD 53
4.4.1 Classificador *Nearest Neighbour*, ambiente Weka 53
4.4.2 Classificador *Gaussian Mixture Model* 56
4.5 Aplicação num caso prático 56
4.5.1 Base de dados e pré-processamento 56
4.5.2 Métodos Weka 58
4.5.3 Método GMM 63
4.6 Conclusões 65

5 Conclusão **67**
5.1 Trabalho Futuro 68

Referências **71**

Lista de Figuras

2.1 Aparelho Fonador Humano. [1] 6
2.2 Modelo fonte-filtro. [2] 7
2.3 Áreas de reconhecimento de orador. 8
2.4 Estrutura genérica de um sistema de reconhecimento de orador. 9
2.5 Computação dos coeficientes MFCC [2] [3]. 12
2.6 Representação de um conjunto de dados constituído por duas classes, do hiperplano que as separa e dos vectores de suporte encontrados (*support vectors*). . . . 18
2.7 Duas classes de dados não linearmente separáveis no espaço original bidimensional (*esquerda*), são separáveis num espaço tridimensional (*direita*). Figura retirada de [4] 19
2.8 Sistema LIA SVM 512, NIST SRE '08. Figura adaptada de [5]. 22
2.9 Sistemas LIA, NIST SRE '08. Figura adaptada de [5]. 23

3.1 Algoritmo de estimação dos NRDs e do deslocamento temporal n_0. Figura adaptada de [6]. 31
3.2 Classificação de vogais cantadas através das características: NRD, características do fluxo glótico (GLOT_F) e MFCC. Figura adaptada de [7]. 33
3.3 Classificação de cantor através das características: NRD, características do fluxo glótico (GLOT_F) e MFCC. Figura adaptada de [7]. 34

4.1 Percentagem de identificações correctas num cenário de 8 oradores, com variação de diversos parâmetros – número de componentes gaussianas, uso de normalização CMS e inclusão de MFCCs diferenciais. 43
4.2 Percentagem de identificações correctas num cenário de 20 oradores, com variação de diversos parâmetros – número de componentes gaussianas, uso de normalização CMS e inclusão de MFCCs diferenciais. 43
4.3 Percentagem de identificações correctas num cenário de 40 oradores, com variação de diversos parâmetros – número de componentes gaussianas, uso de normalização CMS e inclusão de MFCCs diferenciais. 44
4.4 Comparação entre desempenho do sistema usando tempo de teste de 2 segundos e 6 segundos. Resultados apresentados para diferente número de gaussianas utilizadas e diferente número de oradores considerados. 46
4.5 Fonemas incluídos nas anotações que acompanham a TIMIT. 47
4.6 Percentagem de identificações correctas utilizando apenas vogais, para universo de oradores de diferente dimensão (8, 20 e 40 oradores) e teste de diferente duração. 49
4.7 Percentagem de identificações correctas utilizando partes vozeadas, para universo de oradores de diferente dimensão (8, 20 e 40 oradores) e teste de diferente duração. 50

4.8 Percentagem de identificações correctas utilizando partes não vozeadas da fala, para universo de oradores de diferente dimensão (8, 20 e 40 oradores) e teste de diferente duração. 51
4.9 Comparação do desempenho do sistema considerando diferentes partes da voz, para 40 oradores e tempo de treino de 301 vectores MFCC. 52
4.10 Comparação do desempenho do sistema considerando diferentes partes da voz, para 40 oradores e tempo de treino de 578 vectores MFCC. 52
4.11 Matrizes de confusão obtidas para os testes realizados com o algoritmo NNge. . . 60
4.12 Matrizes de confusão obtidas para os testes realizados com o algoritmo SMO (SVMs). 62
4.13 Visualização da projecção dos coeficientes MFCC de dois segmentos do mesmo orador. A projecção mostra os coeficientes de ordem 2 e de ordem 5. 63

Lista de Tabelas

2.1 EER obtido em sistemas GMM-UBM, sistemas GMM-UBM com compensação FA (GMM UBM+FA), sistemas SVM-UBM (GMM UBM+SVM), SVM-UBM com FA (GMM UBM+SVM+FA) e NAP (GMM UBM+SVM+NAP) [8]. 19
2.2 EER obtido para sistema baseado em JFA, duração de treino de 2.5 minutos [9]. . 24

4.1 Tabela resumo dos testes realizados ao sistema de reconhecimento de orador implementado. 40
4.2 Percentagem de identificações correctas para cenário de 8, 20 e 40 oradores e diferente número de componentes gaussianas. 45
4.3 Percentagem de identificações correctas obtidas através de características MFCC e NRD, em ambiente Weka, com segmentos de entrada pertencentes à base de dados Vogais Cantadas. 54
4.4 Percentagem de identificações correctas obtidas através de características MFCC e NRD, em ambiente Weka, com segmentos de entrada pertencentes à base de dados TIMIT. 55
4.5 Percentagem de identificações correctas obtidas através de características MFCC e NRD, em ambiente Weka, com segmentos de entrada pertencentes à base de dados Vogais Faladas. 55
4.6 Exemplo de regras que definem um hiperrectângulo. 59

Abreviaturas e Símbolos

Abreviaturas

ANN	*Artificial Neural Network*
BM	*Background Model*
CMS	*Cepstral Mean Subtraction*
DFT	*Discrete Fourier transform*
EER	*Equal Error Rate*
EM	*Expectation Maximization*
FA	*Factor Analysis*
GMM	*Gaussian Mixture Model*
GSM	*Global System for Mobile Communications*
HMM	*Hidden Markov Model*
IDFT	*Inverse Discrete Fourier transform*
IPA	International Phonetic Alphabet
JFA	*Joint Factor Analysis*
LIA	Laboratoire Informatique d'Avignon
LFA	*Latent Factor Analysis*
LP	Linear Preditivo
LPC	*Linear Predictive Coding*
LPCC	*Linear Predictive Cepstral Coefficients*
MAP	*Maximum A Posteriori*
MIT	Massachusetts Institute of Technology
MFCC	*Mel-Frequency Cepstral Coefficients*
NAP	*Nuisance Attribute Projection*
NIST	National Institute of Standards and Technology
NN	*Nearest Neighbour*
NNge	*Nearest Neighbour generalised*
NRD	*Normalized Relative Delay*
ODFT	*Odd-frequency Discrete Fourier Tranform*
PCM	*Pulse Code Modulation*
SMO	*Sequential Minimal Optimization*
SRE	*Speaker Recognition Evaluation*
SVM	*Support Vector Machines*
TPR	*True Positive Rate*
UBM	*Universal Background Model*
VQ	*Vector Quantization*
Weka	*Waikato Environment for Knowledge Analysis*

Capítulo 1

Introdução

1.1 Reconhecimento de Orador

Reconhecimento de orador trata-se da tarefa computacional de estabelecer ou verificar a identidade de um orador através da sua voz [10]. Sistemas de reconhecimento de orador encontram-se no âmbito de sistemas biométricos, mais especificamente em biometria de *performance*, em que o indivíduo deve executar uma tarefa para ser reconhecido [11].

Existem duas áreas principais em reconhecimento de orador: identificação de orador e verificação de orador. Nesta última pretende-se confirmar que o segmento de voz em análise foi produzido por determinada pessoa, cuja identidade é conhecida de antemão, tomando-se apenas uma decisão binária de confirmação ou rejeição. Em identificação de orador, por contraste, o objectivo é seleccionar o orador de um universo de oradores conhecidos, sem qualquer indicação prévia da sua identidade. O reconhecimento de orador abrange também outros dois métodos distintos: dependente e independente de texto, conforme as gravações de voz usadas correspondem ou não a uma frase específica (texto) que todos os oradores proferiram.

1.2 Aplicações

A tecnologia de reconhecimento de orador oferece várias aplicações na área de segurança. Há mais de uma década que se encontram em funcionamento sistemas de reconhecimento de orador como parte integrante de sistemas de segurança de organizações a nível mundial. Empresas como Allianz Dresdner, Banco Santander, VISA, IBM Europa e Morgan Stanley utilizam esta tecnologia como forma de redefinição periódica de *passwords* das contas de acesso dos funcionários [12]. Outras aplicações, que representam uma grande parte do mercado para biometria de voz, inserem-se no âmbito de aplicação de penas judiciais. A autenticação da voz pode substituir a utilização de PINs para controlo das chamadas efectuadas pelos reclusos. É também utilizada para monitorização de indivíduos em liberdade condicional, prisão domiciliária e outras situações em que é

necessário confirmar a localização do indivíduo. Para tal, é feita automaticamente uma chamada para o local onde é previsto estar o orador, e a identidade deste é confirmada [12].

As aplicações mencionadas aproveitam uma das características que distinguem a biometria de voz de outros métodos de biometria: o uso de equipamento não especializado para recolha dos dados biométricos. No caso da voz, a maioria das soluções é implementada de forma a serem usados microfones comuns ou telefones, enquanto que outros métodos biométricos exigem utilização de equipamento proprietário ou equipamento adaptado à tecnologia em causa [12]. Esta característica traduz-se em vantagens a nível de variedade de aplicações e também em autonomia dos sistemas implementados. Como exemplo deste último, ao ser usado o telefone para fazer o reconhecimento de orador, esse reconhecimento pode ser autónomo, não sendo necessário alocar quaisquer recursos humanos a esta tarefa. Há assim uma diminuição de custos e aumento da flexibilidade do sistema (por exemplo, em termos de horário de funcionamento).

Por último, uma das áreas também receptiva aos sistemas de segurança por biometria de voz é a automação dos serviços *self-service* por telefone. Como uma forma de prevenir e diminuir a taxa de fraude nestes sistemas, várias empresas e organizações adicionaram à verificação por PIN ou password o reconhecimento por voz. Uma das aplicações com maior aderência a esta tecnologia é *telephone banking*, serviço que disponibiliza as operações de consulta de saldo, transferência bancária e pagamentos através do telefone. Reconhecimento de orador é já utilizado nesta área desde 1996, data em que o Glenview State Bank of Illinois implementou pela primeira vez esta tecnologia no seu serviço de *telephone banking* [12].

1.3 Contextualização e Objectivos

O presente trabalho consiste no desenvolvimento de uma solução de reconhecimento de orador, especificamente no âmbito de identificação de orador independente de texto. A implementação do sistema tem em vista o aumento da robustez na identificação de orador dos sistemas do estado da arte, nos cenários de teste em que os segmentos de voz utilizados apresentam durações reduzidas – cerca de dois segundos. Para isto serão estudadas formas de optimizar os métodos já existentes e extensamente utilizados actualmente, e serão também explorados novos métodos, como novas características de voz que indiquem potencial capacidade de alcançar o objectivo pretendido.

É também objectivo do trabalho em causa extrapolar conclusões para um projecto em desenvolvimento paralelo à presente dissertação. Este projecto resulta da colaboração entre a Universidade do Porto e a Polícia Judiciária, e pressupõe a implementação de um método de reconhecimento de orador independente de texto. Este deverá ser aplicado numa base de dados fornecida, com características que condicionam tanto os métodos escolhidos para implementação do sistema, como outras variantes de parametrização do sistema. Assim, a solução desenvolvida ao longo da dissertação foi, em grande parte, desenhada com este objectivo em mente.

1.4 Estrutura da Dissertação

Neste subcapítulo é descrita a organização geral do presente documento. Este encontra-se dividido em seis capítulos principais. No primeiro foi feita uma breve introdução ao problema de reconhecimento de orador e às principais variantes que este compreende; foram apresentadas as aplicações mais comuns desta tecnologia, e por fim apresentaram-se os objectivos e a motivação do trabalho desenvolvido ao longo da dissertação.

No segundo capítulo são abordados alguns conceitos fundamentais à compreensão dos métodos utilizados, principalmente relacionados com o sistema fonador humano e com características da voz.

No terceiro capítulo deste documento é feita uma análise do estado da arte, com restrição ao âmbito em que se insere o trabalho a efectuar. É apresentada a estrutura base dos sistemas de identificação de orador, através da descrição dos módulos funcionais que geralmente os constituem. Para cada um desses módulos é feita uma breve caracterização dos métodos mais frequentemente utilizados e extensivamente analisados na literatura.

Posteriormente, é apresentado no quarto capítulo o sistema de reconhecimento de orador implementado. São descritos em maior profundidade alguns dos métodos escolhidos e são descritas as ferramentas utilizadas para a implementação e estudo da solução desenvolvida.

Os testes realizados à solução desenvolvida assim como os resultados obtidos através destes encontram-se no capítulo 5. É também abordado o trabalho resultante da colaboração da Universidade do Porto e da Polícia Judiciária, pelo que se descrevem os métodos de reconhecimento de orador aplicados nesse projecto e os resultados obtidos.

Por último são retiradas conclusões gerais sobre os resultados obtidos. Para além disto são enunciadas as principais dificuldades sentidas e reflecte-se sobre trabalho futuro.

Capítulo 2

Estado da Arte

No presente capítulo faz-se um estudo do estado da arte em sistemas de reconhecimento de orador, com especial foco em sistemas de identificação e verificação independente de texto. Este levantamento de tecnologias e algoritmos utilizados serviu de base para a escolha dos métodos a implementar no sistema desenvolvido nesta dissertação. Para além disto, foi possível determinar níveis de desempenho actualmente alcançados pelas soluções do estado da arte, o que permite estabelecer valores de comparação para o sistema implementado.

Numa primeira parte deste capítulo são apresentados alguns conceitos fundamentais que serão utilizados ao longo da dissertação. Seguidamente é aprofundado o enquadramento do estudo realizado na temática de reconhecimento de orador. É também descrito o funcionamento geral dos sistemas actuais e são apresentados os principais desafios inerentes à tarefa de identificação e verificação. Os principais blocos funcionais apresentados nesta última parte são descritos em detalhe nos seguintes subcapítulos, assim como os algoritmos mais utilizados em cada um destes. Por fim analisa-se o desempenho de alguns sistemas do estado da arte que incorporam as técnicas apresentadas.

2.1 Conceitos Fundamentais

Neste subcapítulo será feito um breve sumário dos principais conceitos relacionados com o tracto vocal humano e com a produção de fala de forma a auxiliar a compreensão de algumas metodologias descritas ao longo da dissertação.

2.1.1 Tracto vocal e produção de fala

A forma do tracto vocal é um dos factores mais importantes na distinção de diferentes oradores através da fala. De facto, os sistemas de reconhecimento de orador usam geralmente características da voz derivadas do tracto vocal apenas [11].

A geração de fala inicia-se nos pulmões, os quais expelem ar para a traqueia através da glote, área localizada entre as pregas vocais. Estas últimas são parte constituinte da laringe. Os outros órgãos principais do tracto vocal, como visto na figura 2.1, são: faringe, cavidade oral e cavidade nasal. Estes são também denominados "articuladores" em contexto de processamento da fala. A passagem do ar no tracto vocal provoca alterações à onda acústica formada, através das ressonâncias do tracto vocal (as quais são denominadas formantes). Assim, o espectro do sinal é alterado, o que permite estimar através da voz a forma do tracto vocal de um indivíduo.

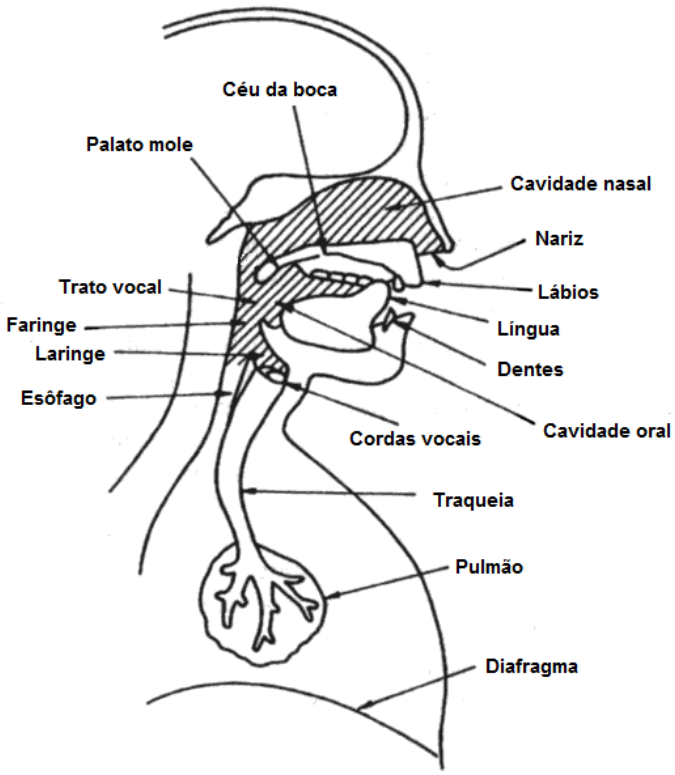

Figura 2.1: Aparelho Fonador Humano. [1]

A excitação gerada pelo ar expelido dos pulmões resulta em diferentes fenómenos, tais como: *phonation* (fonação), *whispering*, *frication*, *compression* e *vibration*. Estes podem ocorrer separadamente ou em simultâneo [11]. No presente documento apenas o fenómeno fonação é pertinente para o estudo em causa, e como tal os restantes não serão explorados em detalhe.

Fonação ocorre quando o ar é modulado pelas pregas vocais, isto é, quando ocorre vibração das mesmas. A frequência desta oscilação denomina-se frequência fundamental e é uma das características físicas que contribui para a distinção de oradores, pois depende do comprimento, da tensão e da massa das pregas vocais. Um som sem vozeamento é produzido, em contrapartida, pela passagem turbulenta de ar através de uma abertura estreita no tracto vocal. Fonação pode ocorrer em conjunto com alguns dos outros fenómenos mencionados. *Frication*, que ocorre quando o ar é forçado através de um canal estreito formado pela colocação de dois articuladores próximos, pode ser acompanhado por fonação, como no caso da consoante "z", ou pelo contrário, sem fonação,

como no caso da consoante “s” [11]. Para designação de fala acompanhada por fonação utiliza-se também o termo voz “vozeada”.

2.1.2 Modelo fonte-filtro

Na área de reconhecimento de fala e de orador um modelo de representação do sistema fonador humano frequentemente utilizado é o modelo fonte-filtro. Este modelo faz a aproximação do sistema de produção de fala ao assumir que este é constituído por uma fonte, que produz a excitação do tracto vocal, e um filtro, que representa o tracto vocal. O sinal de excitação é modulado por um trem de impulsos periódico (para produção de voz vozeada) ou ruído branco (voz não vozeada). Por sua vez, o tracto vocal é modulado por um filtro digital, caracterizado por uma função de transferência $H(z)$ [2].

Nesta aproximação assume-se geralmente a independência da fonte e do filtro, e devido a este factor de simplificação este modelo de representação é frequentemente adoptado. É utilizado nomeadamente em predição linear, em que se assume que o filtro é um filtro *all-pole*, e os coeficientes são estimados de forma a minimizar o sinal de erro. A fala, segundo este modelo, é produzida através da convolução do sinal de excitação com a resposta do filtro.

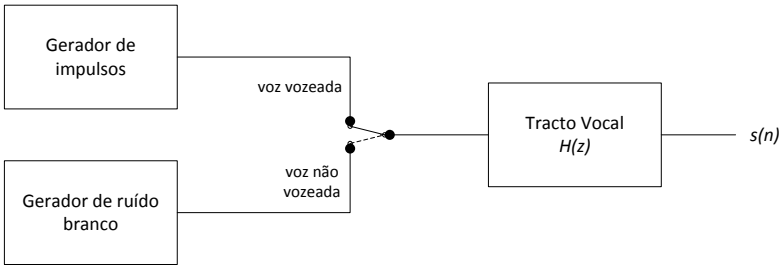

Figura 2.2: Modelo fonte-filtro. [2]

Ao longo do estado da arte o modelo fonte-filtro será abordado, tanto como referência para os métodos que o adoptam como para os métodos que propõem outros modelos e expõem as limitações deste.

2.2 Reconhecimento de Orador

Revisitando as noções já introduzidas no capítulo 1, reconhecimento de orador abrange duas áreas principais: identificação de orador e verificação de orador. A identificação de orador pode ser feita ainda em *close-set* ou *open-set*. No primeiro caso, é considerado um universo limitado e conhecido de oradores, isto é, os oradores inscritos no sistema. Assim, uma nova amostra de um sinal de voz é sempre classificada como um dos oradores da base de dados. Em tarefas *open-set* existe, no entanto, mais uma hipótese de classificação: o sistema pode considerar que a amostra testada não corresponde a nenhum dos oradores conhecidos. Diz-se que o conjunto de oradores é aberto (*open-set*) [13]. Enquanto que a tarefa de verificação é inerentemente *open-set*,

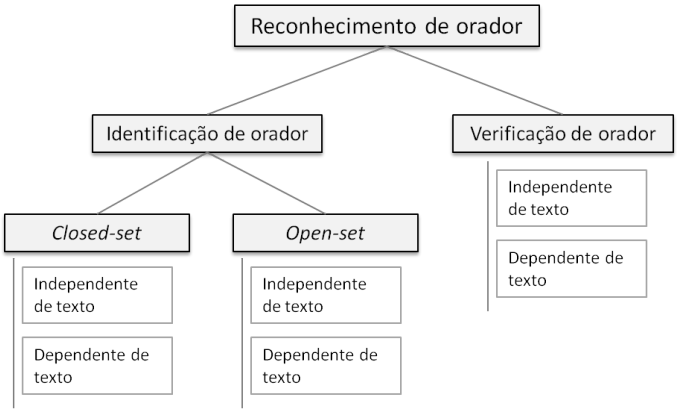

Figura 2.3: Áreas de reconhecimento de orador.

a identificação pode tomar as duas variantes, segundo algumas fontes [2]. Outras, pelo contrário, apenas consideram identificação *close-set* [13], o caso mais comum.

Como explicitado também no capítulo 1, existem outras duas variantes a sistemas de reconhecimento de orador: dependente e independente de texto. Considera-se neste trabalho, como demonstrado na figura 2.3, que tanto verificação como identificação de orador podem funcionar nestes dois modos, no entanto sistemas dependentes de texto são predominantemente sistemas de verificação. De facto, as aplicações que requerem ao utilizador dizer uma frase pré-determinada são sistemas de validação de acesso de indivíduos, os quais se apresentam ao sistema alegando uma certa identidade. Como tal, os sistemas têm por objectivo verificar uma identidade conhecida *a priori*. Certas fontes, como [11], não reconhecem assim a existência de identificação de orador dependente de texto, apenas verificação.

Um sistema de reconhecimento de orador é geralmente constituído pelos seguintes componentes: extracção de características, *pattern matching* e decisão, como ilustrado na figura 2.4. Neste capítulo será feita uma explicação breve do funcionamento destes componentes e das principais técnicas que são utilizadas em cada um deles.

Existem duas fases na identificação de um orador. Na fase de registo de oradores (*enrollment*), são extraídas as características do sinal de fala e é construído um modelo para cada orador, que o representa. Este modelo é guardado na base de dados. Na fase de identificação, as características são extraídas da mesma forma, e é feita uma comparação entre estas e os modelos armazenados, resultantes da fase de registo. Com base nessa comparação é feita uma decisão quanto à identidade do orador. Na prática estas duas fases estão bastante relacionadas entre si, sendo que os algoritmos usados dependem do sistema como um todo. Por exemplo, o método utilizado para o cálculo de correspondência e de identificação está relacionado com o algoritmo de modelação [2].

A fase de registo de oradores é também designada por fase de "treino", pois a construção dos modelos dos oradores equivale ao "treino" destes. A fase de identificação é comummente referida como fase de "teste", pois é testado o conjunto de dados contra os modelos de oradores armazenados [13].

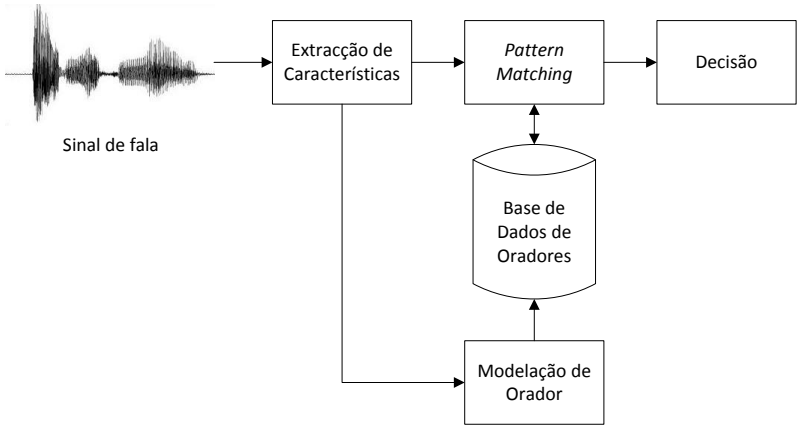

Figura 2.4: Estrutura genérica de um sistema de reconhecimento de orador.

A quantidade de dados de teste e treino, isto é, a duração total dos segmentos de voz utilizada para treinar os modelos e para posterior teste, é um factor determinante no desempenho do sistema de reconhecimento de orador. De facto, há geralmente um compromisso entre a precisão atingida e a duração do segmento de teste [11]. Verifica-se também um *trade-off* entre a precisão e a duração do treino [11]. O uso de uma quantidade de treino três vezes superior pode levar à diminuição do *Equal Error Rate* (EER) (ver medidas de desempenho - 2.2.1). numa proporção aproximadamente idêntica, como já apurado em alguns estudos [8]. A diminuição de EER em três vezes revela o quão significativa é a quantidade de dados disponível ao sistema de reconhecimento.

Outros factores, não relacionados com os algoritmos utilizados mas sim com a base de dados a que são aplicados, afectam a precisão dos sistemas de reconhecimento de orador. Entre os principais factores, a nível acústico, encontram-se: ruído, utilização de diferentes canais e microfones em diferentes gravações (denominado frequentemente na literatura como *channel mismatch*), variação da posição do microfone ao longo de uma gravação e más condições acústicas da sala, que provoquem ecos e outros artefactos. Factores relacionados com o indivíduo devem ser também considerados: estado emocional, estado patológico e até o próprio envelhecimento natural afectam a capacidade de reconhecimento de orador. Em alguns estudos foi obtida uma diferença em termos de probabilidade de falsos negativos de duas vezes, quando as sessões de gravação de vozes para treino e para teste foram efectuadas com 30 dias ou mais de intervalo [8], comparativamente ao cenário em que este intervalo era de apenas 5 dias ou menos. A influência do tempo entre a sessão de treino e de teste não está bem caracterizada, pelo facto de os estudos referidos terem sido efectuados com outros factores de variabilidade. Por outro lado esta diferença acentuada não foi verificada noutros estudos [8]. Como tal, o envelhecimento da voz e outros factores relacionados com o estado do orador são ainda alvo de estudo, e como tal devem ser considerados no desenho e avaliação de sistemas de reconhecimento de orador.

2.2.1 Medidas de desempenho

O desempenho de sistemas biométricos, incluindo sistemas de verificação de orador, é frequentemente medido em termos de *Equal Error Rate*. A EER é o ponto em que a probabilidade de falso negativo e probabilidade falso positivo são iguais. A probabilidade de falso negativo corresponde à probabilidade de o sistema rejeitar o orador em teste, quando na realidade o orador se encontra inscrito na base de dados. Designa-se frequentemente na literatura como *false rejection*, *miss probability* e *type I error*. Falso positivo trata-se do caso em que o sistema aceita um orador como sendo um dos oradores inscritos quando não o é na realidade. Designa-se também *false acceptance*, *false alarm* e *type II error*.

2.3 Extracção de Características

Um sinal de fala apresenta uma enorme quantidade de informação. Como tal, é importante seleccionar criteriosamente a informação relevante para a identificação de um indivíduo, isto é, identificar quais as características de um sinal de fala que possuem poder discriminatório sobre o orador.

Podemos dividir estas características em duas categorias principais: informação de alto-nível e informação de baixo-nível. A primeira está relacionada com o estilo da fala e hábitos oratórios do orador, como o seu dialecto, enquanto que a segunda refere-se a características resultantes das propriedades físicas do tracto vocal, como a frequência fundamental e frequências das formantes [10] [2]. A abordagem geralmente praticada incide sobre a informação de baixo-nível, devido à elevada complexidade da informação de alto-nível e consequente dificuldade na sua medição [2].

As características essenciais à identificação de orador num sinal de fala variam de forma relativamente lenta ao longo do tempo. Assim, se analisarmos o sinal em intervalos de tempo suficientemente curtos (entre 10 e 30 milissegundos), este apresenta características acústicas aproximadamente estáveis. Ao modelar o sinal a partir destas características, é possível reduzir significativamente a quantidade de dados necessária para o descrever. Este processo de redução do volume de dados, mantendo ao mesmo tempo a informação útil para classificação, encontra-se no domínio de extracção de características. A análise descrita denomina-se *short-term analysis* e as características em que se baseia pertencem ao conjunto da informação de baixo-nível da fala [2].

2.3.1 *Linear Predictive Coding*

O modelo linear preditivo (LP) assume que o sinal de voz resulta de uma combinação linear dos seus valores passados e de uma entrada actual:

$$s_n = -\sum_{k=1}^{p} a_k \cdot s_{n-k} + G \cdot u_n. \tag{2.1}$$

Na expressão 2.1, s_n representa a saída actual, p é a ordem de predição, a_k são os parâmetros do modelo denominados coeficientes de predição, s_{n-k} são saídas passadas, G é um factor de ganho escalar, e u_n é a entrada actual. Este último valor, u_n, representa na realidade a fonte do aparelho fonador, isto é, o impulso glótico. Como o valor da fonte é geralmente desconhecido, o modelo linear preditivo ignora esta entrada u_n e faz apenas a modelação do filtro, correspondente ao tracto vocal, como descrito no ponto 2.1.2.

$$\hat{s_n} = -\sum_{k=1}^{p} a_k \cdot s_{n-k}. \tag{2.2}$$

A diferença entre o sinal s_n e a sua aproximação $\hat{s_n}$ corresponde ao erro de predição e_n:

$$e_n = s_n + \sum_{k=1}^{p} a_k \cdot s_{n-k}. \tag{2.3}$$

Deduz-se a partir de 2.3 que e_n corresponde ao sinal de entrada $G \cdot u_n$.

O que se procura obter através do modelo LP 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 o erro e_n. Tendo em conta que e_n contém toda a informação da voz que não é modelada pelos coeficientes de predição, minimizar o erro significa maximizar a informação expressa pelo modelo LP.

É comum efectuar-se uma transformação não-linear sobre os coeficientes de predição para um domínio de características com maior significado perceptivo no contexto de modelação do filtro, como Rácios *Log Area* [11]. Uma transformação que tem sido muito usada é a transformação para Coeficientes Linear Preditivos Cepstrais (*Linear Predictive Cepstral Coefficients* – LPCC), pelo facto de o cepstro se ter vindo a provar como a melhor representação do sinal de fala para reconhecimento de orador. Os coeficientes cepstrais são calculados directamente a partir dos coeficientes de predição [2].

2.3.2 *Mel-frequency Cepstral Coefficients*

A voz pode ser descrita como a convolução de um sinal de fonte (fonte glótica) de variação temporal rápida com a resposta do tracto vocal, de variação lenta, representada como um filtro linear [2].

O cepstro é uma representação do sinal de voz em que estes componentes são desacoplados e transformados em dois componentes aditivos, facilitando a tarefa de separação dos dois e posterior análise. O cepstro é obtido através da seguinte expressão:

$$\text{Cepstro}(\textit{frame}) = \text{IDFT}(\log(\,|\,\text{DFT}(\textit{frame})\,|\,)). \tag{2.4}$$

Segue-se uma breve explicação da expressão 2.4: ao calcular-se a DFT da *frame* obtém-se uma multiplicação dos termos, ao invés de uma convolução, e ao calcular 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 uma da outra. Os principais passos realizados na computação dos coeficientes são apresentados na figura 2.5.

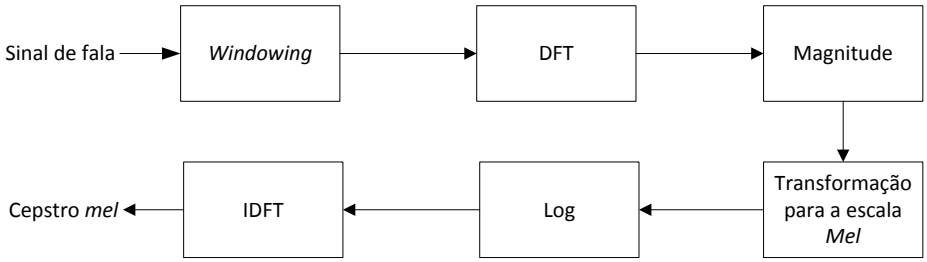

Figura 2.5: Computação dos coeficientes MFCC [2] [3].

Os coeficientes cepstrais *Mel* distinguem-se dos coeficientes cepstrais descritos acima pelo facto de a sua obtenção incluir um passo extra: a transformação das frequências segundo a escala *Mel* (daí existir também a designação *Mel-Warped Cepstrum*, pois faz-se *warp* no domínio das frequências). Esta transformação, expressa por 2.5 [3] [14], é feita de forma a dar menos ênfase às altas frequências, devido ao facto de o sistema auditivo humano apresentar menor sensibilidade a estas.

$$M(f) = 2595 \log_{10}(1 + \frac{f}{700}). \tag{2.5}$$

Uma das formas de obter os *Mel-frequency Cepstrum Coefficients* (MFCCs), após os cálculos indicados, é através de um banco de filtros. Cada filtro deste tem uma resposta em frequência triangular, adaptada à frequência desejada, e calcula a média do espectro em volta dessas frequências.

Normalmente são usados apenas entre 12 e 20 coeficientes, aos quais se atribuem diferentes pesos de acordo com a quantidade de informação sobre o orador que cada um contém. Este método de representação de características tem a vantagem de ser facilmente utilizado em conjunto com o método de classificação *Gaussian Mixture Model*, pelo facto da densidade do cepstro ser bem modelada por combinações de curvas gaussianas, que este utiliza. Adicionalmente, estudos têm demonstrado que os MFCCs produzem bons resultados em sistemas de reconhecimento de orador e de fala [11], daí serem o método mais utilizado actualmente [15].

2.3.2.1 MFCCs diferenciais

Por vezes é feita a extracção de MFCCs diferenciais, para além dos coeficientes MFCC mencionados. Os MFCCs diferenciais tratam-se da primeira derivada (ΔMFCCs) e da segunda derivada (ΔΔMFCCs) dos coeficientes originais. A primeira derivada apresenta pouca correlação com o vector do qual foi extraída, e por este motivo foi já mostrado em alguns trabalhos experimentais que melhora o desempenho de reconhecimento de orador, assim como a segunda derivada [16]. O cálculo da derivada pode ser feito da seguinte forma [14]:

$$d_t = \frac{\sum_{\theta=1}^{\Theta} \theta(c_{t+\theta} - c_{t-\theta})}{2\sum_{\theta=1}^{\Theta} \theta^2} \tag{2.6}$$

onde d_t é o coeficiente delta no instante t, $c_{t-\theta}$ e $c_{t+\theta}$ são coeficientes estáticos nos instantes $t-\theta$ e $t+\theta$ respectivamente e Θ é o tamanho da janela de cálculo dos MFCCs diferenciais.

2.3.2.2 Técnicas de normalização

Como já indicado no ponto 2.2, *channel mismatch* é um dos factores que dificulta a tarefa de reconhecimento de orador. Apesar de a denominação apontar apenas para diferenças de canal, esta geralmente refere-se a um conjunto alargado de variantes entre duas gravações, tais como: ruído de fundo, ruído não estacionário, diferenças de acústica da sala, para além da diferença entre os canais de transmissão [17].

Várias técnicas são utilizadas actualmente ao nível das características de voz extraídas, com o intuito de reduzir os efeitos de *channel mismatch*; entre as mais comuns encontram-se *Cepstral Mean Subtraction* (CMS) e *Relative Spectra Filtering* (RASTA *filtering*).

CMS consiste em subtrair a média dos N vectores cepstrais extraídos aos vectores originais [18]:

$$\vec{c}_{cms;i} = \vec{c}_{y;i} - \vec{c}_{y;avg}, \quad i = 1,\ldots,N \tag{2.7}$$

onde $\vec{c}_{y;avg} = \frac{1}{N}\sum_{i=1}^{N} \vec{c}_{y;i}$. Este método é baseado no pressuposto que o filtro do canal não varia significativamente ao longo do segmento analisado. Como tal, o método CMS elimina as variações lentas do canal, o que enquanto reduz os efeitos de variação do canal, também elimina alguma informação relacionada com a fala e o orador [17].

Quanto à filtragem RASTA, esta aplica um filtro que elimina as componentes espectrais que variam demasiado lentamente ou rapidamente em comparação com a taxa típica de variação das componentes espectrais. CMS é uma das técnica mais simples, tanto em comparação com RASTA *filtering* como em comparação com a generalidade dos métodos de normalização. No entanto, tem sido indicado repetidamente que tem desempenho igual ou superior a técnicas mais complexas [17].

2.3.3 Alternativas e Conclusões

Foram descritas neste subcapítulo as duas técnicas de extracção de características de voz mais utilizadas actualmente em reconhecimento de orador - MFCCs e a variante *Linear Predictive Cepstral Coefficients* de LPCs [2] [19]. Ambos os métodos extraem eficazmente características do tracto vocal com propriedades discriminatórias de orador, embora por meios de computação diferentes. Apesar de não haver um consenso claro sobre a superioridade de um dos métodos, os MFCCs são geralmente indicados como os que atingem melhor desempenho [20] [21]. Existe, no

entanto, alguma reserva quanto a estes métodos. Por um lado, ambos atingem níveis de precisão tão ou mais elevados na tarefa de reconhecimento de fala, comparativamente à tarefa de reconhecimento de orador. É de facto de esperar que a suavização do espectro utilizada nos dois métodos cause alguma normalização quanto ao orador. Por outro lado, uma característica do sinal de fala utilizada em reconhecimento de orador deveria, idealmente, conter apenas informação relativa a este e conter o mínimo de informação relativa ao discurso contido no sinal de fala [19].

Outra desvantagem presente nestes métodos tradicionais prende-se no facto de não terem em consideração a dependência da fonte glótica e do tracto vocal [22]. Assim, estes métodos atingem bom desempenho porque fazem uma modelação bastante aproximada do tracto vocal apenas - no entanto, não aproveitam informação discriminatória de orador relacionada com a glote. Na última década têm sido conduzidos vários estudos com o intuito de incorporar esta informação nos sistemas de reconhecimento, sob a forma de várias características: *pitch*, estrutura harmónica e fluxo glótico (*glottal flow*), entre outros [16] [23].

2.4 *Pattern Matching*

Pattern Matching consiste em gerar uma pontuação de correspondência (*match score*) entre o modelo da voz do orador de entrada e modelos previamente conhecidos [11]. Existem portanto dois passos envolvidos na tarefa de *pattern matching*: modelação e *matching*. Modelação consiste em registar um orador no sistema de reconhecimento ao criar um modelo da sua voz, baseado nos vectores de características extraídos. Após ter sido obtido esse modelo, são calculadas medidas de semelhança com modelos já inscritos no sistema - *matching* [2].

Os métodos de modelação e *matching* são classificados em modelos *template* e modelos estocásticos. A abordagem por modelos *template* considera que a amostra em observação é uma réplica imperfeita do *template* e procura alinhar as duas de forma a minimizar a distância entre estas [11]. Em sistemas de reconhecimento de orador independentes de texto, são calculadas as médias dos vectores de características, obtidas a partir de períodos de tempo relativamente longos, para distinguir os oradores. São portanto ignoradas as variações temporais e são usadas apenas as médias globais – denominam-se métodos independentes do tempo [11] [2] [20].

Os modelos estocásticos baseiam-se numa abordagem diferente, denominada probabilística. O resultado do *matching* expressa-se numa medida de verosimilhança (*likelihood*) e não através de uma medida de distância entre modelos. Um orador é modelado através de distribuições de probabilidade que descrevem a variação das características ao longo do tempo [11] [2] [20].

Apesar dos primeiros trabalhos desenvolvidos na área de reconhecimento de orador, em especial reconhecimento dependente de texto, utilizarem maioritariamente métodos *template*, os métodos estocásticos rapidamente ganharam popularidade após se terem divulgado técnicas poderosas de modelação como *Gaussian Mixture Models*. Métodos baseados nas médias das características geralmente apresentam resultados sub-óptimos e são particularmente sensíveis a variações do

canal e a ruído de fundo [10] [20]. Adicionalmente, métodos estocásticos apresentam maior flexibilidade e as medidas de verosimilhança representam um resultado teoricamente mais significativo que as medidas de distância utilizadas em métodos *template* [11].

2.4.1 *Nearest Neighbour*

O algoritmo *Nearest Neighbour* (NN) é um método de *machine learning* em que se procura classificar dados desconhecidos ao sistema. Esta classificação baseia-se na comparação dos novos dados com dados conhecidos previamente. O algoritmo parte de um conjunto de observações (em reconhecimento de orador, um conjunto de características de voz), acompanhado das respectivas anotações com informação sobre a classe a que cada observação pertence (o orador). 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 [24]:

$$d(x,y) = \|x-y\| = \sqrt{\sum_{i=1}^{m}(x_i-y_i)^2}. \tag{2.8}$$

Uma variante comum ao método NN é o *K Nearest Neighbours*. Este considera não um vizinho mais próximo mas sim K, atribuindo assim à observação nova a classe mais frequente entre o conjunto dos K pontos. Este método é indicado para classificação de dados afectados por ruído, em que as concentrações de pontos (*clusters*) não se encontram perfeitamente delineados, pois previne que uma amostra seja classificada incorrectamente a partir de um ponto deslocado da maior aglomeração de pontos da classe a que pertence.

Este método é um método *template*. Em comparação com outros métodos deste tipo, como *Dynamic Time Warping*, atinge maior precisão, embora seja um método bastante exigente em termos computacionais.

2.4.2 *Gaussian Mixture Models*

Gaussian Mixture Models incluem-se nos modelos estocásticos. Ao longo da última década têm vindo a estabelecer-se como o método dominante em sistemas de identificação independente de texto [25].

Uma gaussiana multi-variada é expressa por:

$$f(\vec{x}|\vec{\mu},\Sigma) = \frac{1}{(2\pi)^{\frac{D}{2}}|\Sigma|^{\frac{1}{2}}}\exp\left(-\frac{1}{2}(x-\vec{\mu})^T\Sigma^{-1}(x-\vec{\mu})\right) \tag{2.9}$$

onde $\vec{x}$ é um vector de D dimensões, $\vec{\mu}$ o vector de média e Σ a matriz de covariância.

Um modelo de misturas gaussianas (*Gaussian Mixture Model* - GMM) consiste, como o próprio nome indica, numa soma pesada de várias gaussianas multi-variadas. Sendo M o número de gaussianas constituintes do modelo, este é descrito por:

$$f(x|\mu,\Sigma) = \sum_{k=1}^{M} c_k \frac{1}{\sqrt{2\pi|\Sigma_k|}} \exp\left((x-\mu_k)^T \Sigma^{-1} (x-\mu_k)\right) \tag{2.10}$$

onde c_k representa o peso de cada gaussiana.

Através do modelo GMM é calculado o valor de *likelihood*, através da expressão:

$$b_j(o_t) = \sum_{m=1}^{M} c_{jm} \frac{1}{\sqrt{2\pi|\Sigma_{jm}|}} \exp\left((o_t-\mu_{jm})^T \Sigma_{jm}^{-1} (o_t-\mu_{jm})\right). \tag{2.11}$$

Na equação acima, $b_j(o_t)$ representa o *likelihood* do vector de observações acústicas o_t relativamente à gaussiana j. Ao criar o modelo GMM os parâmetros μ, Σ e c, que o descrevem, são escolhidos de forma a maximizar o valor de *likelihood*.

Os modelos GMM são descritos em mais detalhe no capítulo 3.2.

2.4.2.1 GMM-UBM

De forma a melhorar o desempenho dos sistemas baseados em *Gaussian Mixture Models*, os métodos actuais fazem frequentemente modelação de impostores (oradores não inscritos no sistema). Esta modelação é feita através de um *background model* (BM), cuja variante mais comum denomina-se *universal background model* (UBM), dai a designação GMM-UBM.

Para a tarefa de verificação de orador, a construção de um *background model* pode ser vista como a modelação da hipótese alternativa à hipótese colocada em cada teste: segmento Y pertence ao orador S. Considerando:

hipótese H_0 - segmento Y pertence ao orador S,

a hipótese contrária é definida como:

hipótese H_1 - segmento Y não pertence ao orador S.

A decisão entre as duas hipóteses é feita utilizando o valor de *likelihood* através do *likelihood ratio test*:

$$\frac{p(Y|H_0)}{p(Y|H_1)} \begin{cases} \geq \theta & \text{aceita } H_0 \\ < \theta & \text{aceita } H_1 \end{cases}$$

em que $p(Y|H_0)$ é o *likelihood* calculado para o segmento Y relativamente ao modelo criado para H_0 (modelo do orador S) e $p(Y|H_1)$ é *likelihood* calculado para o segmento Y relativamente ao modelo criado para H_1 (modelo dos oradores que não são o orador S - *background model*). O

valor θ é o *threshold* sobre o qual assenta a decisão. Geralmente os valores de *likelihood* são manipulados na forma logarítmica e o rácio acima é transformado em:

$$\Lambda(Y) = \log p(Y|H_0) - \log p(Y|H_1) \tag{2.12}$$

sendo $\Lambda(Y)$ o *log-likelihood ratio*.

A constituição dos modelos *background*, os modelos de hipótese alternativa, podem ser constituídos através de diferentes abordagens e com variação de vários parâmetros, incluindo número de oradores e quantidade de dados. Idealmente o número de oradores incluído deve ser o maior possível para modelar da melhor forma os impostores, no entanto implicações a nível de complexidade computacional e memória de armazenamento são também consideradas.

Por outro lado, o BM deve representar de forma equilibrada as subpopulações consideradas. Por exemplo, quando se incluem tanto oradores femininos como oradores masculinos, o *background model* deve ser treinado com a mesma quantidade de dados de vozes femininas e masculinas, caso contrário os resultados podem ser tendenciosos relativamente a uma das subpopulações. Os *background models* podem também ser formados com base em duas abordagens diferentes: construir um BM específico para cada orador, ou um único BM que pode ser utilizado para teste com todos os oradores. No primeiro caso, os impostores são escolhidos de acordo com o orador em teste. Este método traz geralmente maior precisão ao sistema, no entanto aumenta também a complexidade dos algoritmos. A segunda abordagem é portanto mais frequentemente utilizada [26] [27].

2.4.3 *Support Vector Machines*

Uma *Support Vector Machine* (SVM) é uma ferramenta de *machine learning*, que à semelhança do NN, enquadra-se no tipo de aprendizagem supervisionada (*supervised learning*). Segue-se uma breve descrição do seu funcionamento.

Os dados de treino são constituídos por um conjunto de vectores x_i, em que cada um destes compreende um certo número de características (*features*). Os vectores x_i são acompanhados por anotações, vectores y_i, que contêm a classificação dos dados de entrada. No caso de os dados pertencerem a duas classes, estas anotações tomam geralmente 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 com anotação $y = +1$ encontram-se de um lado do hiperplano e que os dados com anotação $y = -1$ se encontram no lado oposto do hiperplano. Este deve ainda ser escolhido de forma a maximizar a distância relativa a cada uma das classes de pontos. Os vectores de suporte (*support vectors*) são os pontos do conjunto de dados que se encontram mais próximos do hiperplano.

O hiperplano é dado pela expressão:

$$g(x) = w \cdot x + b = 0 \tag{2.13}$$

onde w é o vector de pesos que determina a orientação do plano e b é o desvio (*offset*) do plano relativamente à origem do espaço de entrada.

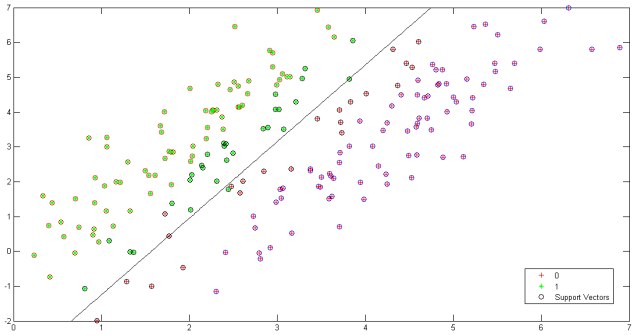

Figura 2.6: Representação de um conjunto de dados constituído por duas classes, do hiperplano que as separa e dos vectores de suporte encontrados (*support vectors*).

Usualmente é feita uma alteração de escala de forma a os pontos mais próximos de um lado do hiperplano sejam definidos por $w \cdot x + b = 1$ e os pontos do outro lado serem definidos por $w \cdot x + b = -1$.

Os planos definidos pelas expressões indicadas são chamados de planos canónicos e a região delimitada por estes *margin band*.

A linha perpendicular ao hiperplano é a menor linha que une os dois planos canónicos, e tem comprimento $2/||w||_2$ (sendo $||w||_2 = \sqrt[2]{w^T w}$). Pretende-se maximizar esta margem, o que equivale a minimizar:

$$\frac{1}{2}||w||_2^2 \tag{2.14}$$

com condição:

$$y_i(w \cdot x + b) \geq 1, \forall i. \tag{2.15}$$

O problema proposto acima pode ser reformulado através da função de Lagrange e utilização formulação dual:

$$W(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j). \tag{2.16}$$

Pretende-se maximizar $W(\alpha)$, com as variáveis α_i restritas a:

$$\alpha_i \geq 0 \qquad \sum_{i=1}^{m} \alpha_i y_i = 0. \tag{2.17}$$

Esta função de mapeamento, denominada *kernel*, pode ser escolhida de forma a obter melhor

separação dos dados. Caso estes não sejam separáveis no espaço em que se encontram inicialmente, pode ser feito um mapeamento num espaço de maior dimensão. Um exemplo deste procedimento é ilustrado na figura 2.7.

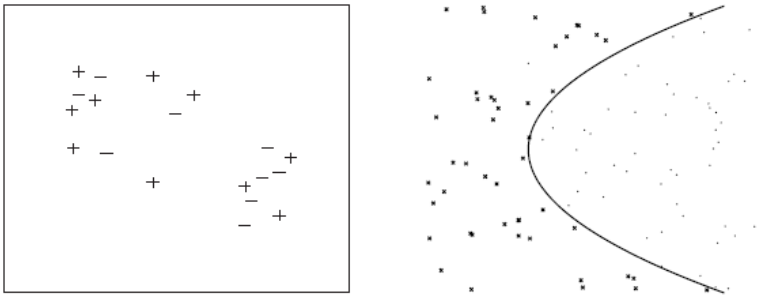

Figura 2.7: Duas classes de dados não linearmente separáveis no espaço original bidimensional (*esquerda*), são separáveis num espaço tridimensional (*direita*). Figura retirada de [4]

Na figura 2.7 o *kernel* utilizado para separar os dados é um *kernel* gaussiano ($K(x_i,x_j) = \exp^{-(x_i-x_j)^2/2\sigma^2}$). Várias funções de *kernel* podem ser utilizadas, sendo uma das mais comuns a função polinomial [4].

A classificação de um conjunto de dados que contém mais do que duas classes pode ser feita através de vários métodos. Os mais comuns baseiam-se na transformação do problema de classificação multi-classe em vários problemas de classificação binária.

2.4.4 Alternativas e Conclusões

GMMs e SVMs constituem actualmente os dois principais métodos de modelação de orador utilizados em reconhecimento de orador, havendo predominância do método GMM-UBM [8]. Nos últimos anos um novo método que incorpora GMM-UBM e SVMs tem ganho popularidade, pelo facto de combinar a robustez da modelação probabilística fornecida pelos *Gaussian Mixture Models* com o poder discrimatório de *Support Vector Machines* [8]. Este método, proposto em [28] [29], utiliza GMMs para treinar os dados e, através de adaptação *Maximum A Posteriori* (MAP), cria um *supervector* que contém as médias das componentes GMM. A partir destes *supervectors* é construido o *kernel* que será utilizado no classificador SVM.

Outros avanços na modelação e classificação dos dados, aplicados em conjunção com os métodos descritos, têm permitido melhorar significativamente os resultados obtidos pelos sistemas de reconhecimento de orador. Estes avanços têm-se focado na compensação da variabilidade entre as sessões de treino e de teste, a nível acústico. Alguns dos métodos mais frequentemente utilizados são *Joint Factor Analysis* (JFA) [30] e *Nuisance Attribute Projection* (NAP) [29].

A tabela 2.1 mostra resultados obtidos com sistemas avaliados na edição de 2006 do National Institute of Standards and Technology Speaker Recognition Evaluation (NIST SRE). Esta plataforma de avaliação é abordada no seguinte subcapítulo. As taxas de EER apresentadas foram

Tabela 2.1: EER obtido em sistemas GMM-UBM, sistemas GMM-UBM com compensação FA (GMM UBM+FA), sistemas SVM-UBM (GMM UBM+SVM), SVM-UBM com FA (GMM UBM+SVM+FA) e NAP (GMM UBM+SVM+NAP) [8].

Sistema	EER(%)
GMM UBM	8,47%
GMM UBM+SVM	6,88%
GMM UBM+FA	4,55%
GMM UBM+SVM+FA	4,48%
GMM UBM+SVM+NAP	5,28%

obtidas com bases de dados constituídas por vozes masculinas apenas e em inglês. Conclui-se que as técnicas de compensação têm, de facto, um grande impacto no desempenho dos métodos de modelação e consequentemente na tarefa de verificação de orador. Com o modelador GMM a redução da EER foi de 3,92% ao utilizar *factor analysis* (FA), comparativamente ao sistema sem compensação. Por outro lado, verifica-se que, usando FA, o sistema GMM-UBM e o sistemas GMM-SVM têm níveis de precisão muito próximos (EER de 4,55% e 4,48% respectivamente).

Embora os métodos discutidos obtenham os melhores resultados, outros métodos têm sido aplicados com sucesso à tarefa de reconhecimento de orador. Entre os mais populares encontram-se *Hidden Markov Models* (HMMs) e *Artificial Neural Networks* (ANNs). HMMs são baseados em máquinas de estados finitos, em que cada estado é descrito pela função densidade probabilidade de um vector de características, dado o estado actual. As transições entre estados são também definidas através de uma probabilidade. HMMs podem ser aplicados tanto em aplicações dependentes de texto como independente de texto, mas é na primeira que tem tido maior sucesso, pelo facto de a modelação HMM ter em conta sequenciamento temporal dos sons da fala - informação que é vantajosa apenas em aplicações dependentes de texto [31]. Como tal, em cenários independentes de texto este método tem desempenho comparável a métodos algo ultrapassados, como *Vector Quantization* (VQ), enquanto que em cenários dependentes de texto atinge melhores resultados que os métodos convencionais [11] [13].

Artificial Neural Networks foram dos primeiros métodos a serem aplicados com sucesso à tarefa de reconhecimento de orador [13]. Este método faz modelação da função de decisão que melhor discrimina oradores dentro de um conjunto conhecido, ao invés da abordagem tradicional de modelar os oradores [31]. Existem várias extenções e modificações deste método, como *neural tree network*, *modified neural tree network* e *multilayer perceptrons*. Geralmente estes métodos têm desempenho comparavél a VQ [32], e como tal têm sido ultrapassados pelas abordagens GMM e SVM.

2.5 Desempenho dos Sistemas Actuais

A avaliação comparativa do desempenho dos sistemas de reconhecimento de orador independente de texto tem sido conduzida em grande parte pelas avaliações do National Institute of Standards and Technology, como referido anteriormente. Este instituto tem efectuado anualmente,

desde 1996, a Speaker Recognition Evaluation: plataforma que permite aferir e comparar o desempenho das diferentes tecnologias, ao fornecer as mesmas condições de teste aos sistemas participantes. As avaliações focam-se principalmente na tarefa de verificação de orador em gravações telefónicas [10] [33]. A aderência a este evento tem aumentado significativamente desde o seu início, sendo que na última edição do NIST SRE, em 2010, participaram 58 entidades [10]. Para além dos participantes terem acesso livre a bases de dados actualizadas e de grandes dimensões, este tipo de avaliação permite promover as tecnologias mais avançadas e que atingem melhores resultados, guiando a investigação para os métodos mais promissores. Por outro lado, os patrocinadores têm a oportunidade de propor novos desafios e novas metodologias de teste que vão de acordo às necessidades existentes [8]. Este tipo de avaliação é assim em parte responsável pelo impulso à investigação e ao investimento nas áreas de reconhecimento de orador e de fala.

No sítio oficial do NIST SRE apenas se encontram disponíveis informações sobre os métodos de avaliação aplicados à edição de 2010 [34], sendo que os últimos resultados apresentados dizem respeito à edição de 2008 (disponíveis em [35]). Como tal, é sobre esta que incide a breve análise que se segue.

A informação disponível na referência indicada não explicita contudo, que tipo de tecnologias são aplicadas em cada sistema avaliado. A partir dos gráficos disponibilizados é apenas possível extrair informação sobre as condições de avaliação: duração dos segmentos de voz usados para treino e teste, número de segmentos, número de canais usados para gravação, tipo de discurso, entre outras variantes. De forma a analisar a evolução do desempenho em função das técnicas incorporadas, utilizam-se dados relativos aos sistemas submetidos pelo Laboratoire Informatique d'Avignon (LIA), Université d'Avignon et des Pays de Vaucluse. Todos estes sistemas e correspondentes metodologias são desenvolvidas através do *software* livre ALIZE/SpkDet [36], disponível em [37].

A nível de características os sistemas LIA utilizam características *Linear Frequency Cepstral Coefficientes* (LFCC). Na edição do NIST SRE de 2006 optimizaram esta componente ao estender o vector de características para dimensão 50, o que provocou uma diminuição de EER de 10% [8]. Assim, o vector de características é constituído por 19 coeficientes LFCC, 1 coeficiente de energia, 19 coeficientes LFCC derivados (ΔLFCC) e 11 coeficientes resultantes da segunda derivada ($\Delta\Delta$LFCC) [5].

Os esforços realizados com o objectivo de aumentar a precisão (reduzindo o EER, entre outras medidas), focaram-se no entanto nos métodos de modelação. Actualmente três subsistemas do LIA estão em desenvolvimento e avaliação nas últimas edições NIST e encontram-se descritos em [5]. O primeiro sistema é baseado no método GMM-SVM, com modelação da variabilidade de sessão *Latent Factor Analysis* (LFA). A partir dos resultados obtidos no NIST SRE'08, representados no gráfico 2.8, pode-se observar que este método é bastante mais eficaz quando se utilizam apenas vozes masculinas. O segundo e terceiro sistemas são baseados na abordagem clássica GMM-UBM, de novo usando LFA. A principal diferença entre os dois reside no número de componentes gaussianas utilizadas: 2048 no segundo subsistema e 512 no terceiro.

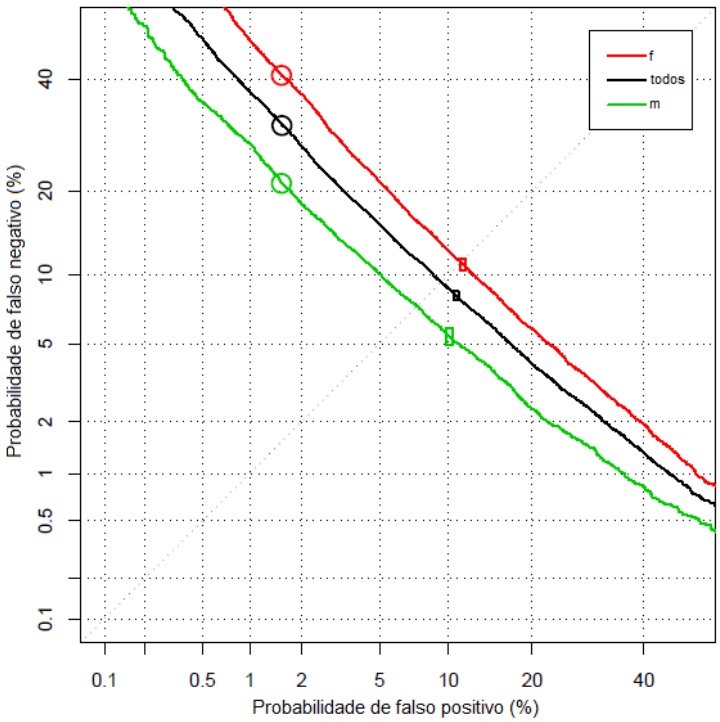

Figura 2.8: Sistema LIA SVM 512, NIST SRE '08. Figura adaptada de [5].

As percentagens de falsos negativos e de falsos positivos para os sistemas descritos são representados no gráfico 2.9. Para além dos três subsistemas descritos, estão presentes outros dois. O "GMM2048 reverse mode" é uma variante do segundo sistema em que se efectua a troca dos segmentos de treino pelos de teste e vice-versa. O subsistema "Fusion LIA2" representa um sistema obtido através da combinação dos subsistemas base, através do método *Linear Logistic Regression*, com auxílio da *toolkit* FoCal [38]. O EER do sistema com melhor desempenho nas condições apresentadas (Fusion LIA2) é de 7,66%.

Note-se que estes resultados são referentes a condições diferentes das presentes nos testes mencionados acima (tabela 2.1). Este teste foi realizado nas condições indicadas na plataforma NIST SRE como "short2-short3", em que [39]:

- Dados de treino ("short2"): excerto de conversa telefónica, com dois canais, com duração total de cerca de 5 minutos, com designação do canal referente ao orador em treino, ou excerto de conversa captada por microfone com duração de cerca de 3 minutos.

- Dados de teste ("short3"): excerto de conversa telefónica, com dois canais, com duração total de cerca de 5 minutos, com o canal referente ao orador em teste designado, ou conversa telefónica semelhante com o canal do orador em teste gravado por um microfone, ou uma conversa gravada por microfone com duração de cerca de 3 minutos.

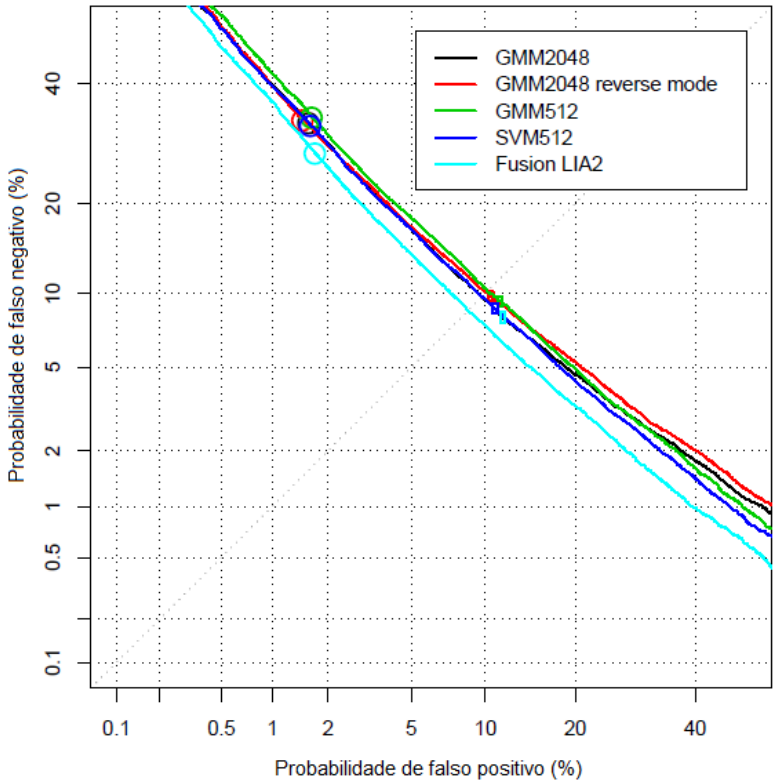

Figura 2.9: Sistemas LIA, NIST SRE '08. Figura adaptada de [5].

Em ambos os casos, quando o canal de cada orador não é designado e a conversa com inclusão dos dois oradores é fornecida, é utilizado um segmentador que através da energia do sinal estima quais os intervalos da conversa que contêm as falas do orador em causa. Os sistemas que participam na avaliação podem aplicar os sistemas de reconhecimento de orador a esses intervalos pré-determinados ou usar um sistema próprio de separação de oradores.

Outro sistema submetido por outra entidade - SRI International - obteve resultados semelhantes, com EER de 7,17% para as mesmas condições, com um método baseado em GMMs e características espectrais. Estes resultados podem ser consultados em [40].

Enquanto que a tarefa de reconhecimento de orador nas condições acima mencionadas tem sido alvo de investigação por várias entidades ao longo das últimas décadas, o reconhecimento de orador em situações de escassez de dados, tanto para treino como para teste, não tem sido tão intensivamente estudado. O desempenho dos sistemas nestas condições mais adversas não se encontra devidamente caracterizado. No entanto sabe-se que a redução da duração dos segmentos de voz utilizados provoca uma diminuição significativa da precisão e que nestas condições existe alguma dificuldade em obter identificação robusta de orador.

O facto de este não ser um problema muito estudado, comparativamente com a dimensão do universo de estudos em reconhecimento de orador, é ilustrado pelo facto de no NIST SRE'08 as únicas tarefas de verificação propostas com durações algo reduzidas proporem teste e treino

dos oradores com 10 segundos de voz. Para além de estas não serem tarefas obrigatórias para a participação dos sistemas no NIST SRE, 10 segundos, especialmente em duração de teste, representa duração demasiado extensa para algumas aplicações. A investigação nesta vertente do reconhecimento tem vindo, no entanto, a aumentar, e existem já alguns estudos realizados na área. Abordam-se de seguida alguns destes.

Utilizado um sistema GMM base, em [22] são indicados valores de percentagem de identificações correctas para as características MFCC. A redução do tempo de teste de 6 segundos para 2 segundos provoca a diminuição da taxa de identificação em 8.5%. Outro estudo mais recente que explora a questão da duração do treino e de teste encontra-se descrito em [9]. Neste trabalho são apresentados os resultados obtidos com diferentes sistemas de verificação de orador para duração de teste compreendida entre 2 segundos e 2,5 minutos, em combinação com duração de treino total (2,5 minutos) ou duração de treino igual à duração de teste.

Os resultados obtidos denotam que todos os sistemas sofrem uma quebra de desempenho para testes com duração inferior a 10 segundos. Para um sistema baseado em JFA, por exemplo, os EERs obtidos para treino máximo (2,5 minutos) evoluem, em função do tempo de teste, como indicado na tabela 2.2:

Tabela 2.2: EER obtido para sistema baseado em JFA, duração de treino de 2.5 minutos [9].

Tempo de teste (segundos)	EER(%)
2	22,48%
4	17,96%
8	13,43%
10	12,11%
20	7,67%
50	4,54%
150	3,37%

Como se pode observar por esta, há um decréscimo muito acentuado da precisão do sistema à medida que se consideram tempo de teste menores. Este cenário agrava-se quando o treino dos oradores não é feito com uma quantidade abundante de dados. Por exemplo, ao utilizar tempo de treino e tempo de teste de 10 segundos a EER desce para 21,17%.

Através destes dados pode-se concluir que esta área necessita de um estudo mais aprofundado e desenvolvimento acentuado de forma a poder ser realizado o reconhecimento de orador robusto em aplicações caracterizadas por escassez de dados.

2.6 Conclusões

Ao longo deste capítulo foram apresentados os principais métodos aplicados à área de reconhecimento de orador independente de texto, tanto na componente de extracção de características como na de algoritmos de *pattern matching* utilizados na modelação e classificação de orador.

É notório que as evoluções mais significativas nos últimos anos se têm dado na segunda área, pelo que os métodos de extracção mais utilizados têm-se mantido sem alterações significativas.

Estes métodos mais frequentes no estado da arte são LPCCs e MFCCs, havendo alguma predominância dos segundos. Como estes modelam o tracto vocal do orador, vários estudos foram já efectuados no sentido de incluir novas características relacionadas com a fonte glótica. No entanto não existe ainda destaque de um método singular.

Quanto aos métodos de *Pattern Matching*, dada à elevada precisão que conferem aos sistemas de reconhecimento, os GMMs são claramente os métodos mais utilizados. No entanto, vários métodos complementares têm sido estudados e aplicados com sucesso à tarefa de identificação e verificação de orador, como métodos de compensação de variabilidade e uso de SVMs.

Capítulo 3

Sistema de Reconhecimento de Orador

No presente capítulo é apresentado o sistema de reconhecimento de orador implementado no âmbito desta dissertação. Numa primeira instância são revistos os objectivos estabelecidos para a solução a implementar e é feita uma análise mais aprofundada dos algoritmos de extracção de características e de classificação seleccionados. Após a descrição teórica destes métodos é abordada a implementação do sistema em Matlab®.

3.1 Introdução

Como previamente mencionado na introdução do presente documento, o trabalho realizado ao longo da dissertação tem por objectivo principal o estudo da tarefa de reconhecimento de orador em situações de escassez de dados disponíveis para teste e treino dos algoritmos. Como ponto de referência relativo a esta duração estabeleceu-se, com base na capacidade actual dos sistemas de reconhecimento de orador e nas aplicações práticas para o sistema em causa, a duração média de dois segundos.

A implementação do sistema completo de reconhecimento de orador tem por objectivo constituir uma base para o estudo da problemática enunciada. O foco deste estudo será a investigação de características com elevado poder discriminatório de orador. Para este fim são analisadas as características do estado de arte e é traçado o perfil do desempenho destas em função da variação da duração dos segmentos utilizados, assim como variação de outros parâmetros pertinentes para o estudo em causa (como presença de vozeamento). Após este estudo inicial são estudadas características com potencial poder discriminatório que permitam aumentar o desempenho do sistema nas situações identificadas como "problemáticas" para os métodos actuais de extracção de características. As características seleccionadas para estudo denominam-se *Normalized Relative Delays* (NRDs) e são apresentadas em detalhe no subcapítulo 3.3.

Um segundo objectivo paralelo aos já enunciados está relacionado com o trabalho de investigação em curso no semestre de realização desta dissertação (Janeiro a Julho de 2011), resultante da colaboração entre a Universidade do Porto e a Polícia Judiciária, descrito no capítulo 1. Pretende-se obter uma solução adaptada às necessidades do trabalho de investigação, sobre a qual se conheçam dados do seu desempenho em condições independentes - isto é, usando como dados de entrada segmentos de uma base de dados extensamente utilizada na literatura como a base de dados TIMIT. Esta será descrita no próximo capítulo, na secção 4.2.1. No entanto, apenas uma parte do estudo conduzido terá resultados pertinentes ao trabalho de investigação colaborativo com a Polícia Judiciária, devido ao facto de as novas características de voz exploradas (NRDs) exigirem a presença da frequência fundamental da voz nos segmentos utilizados para treino e teste do sistema de reconhecimento. Dado que os segmentos disponíveis foram obtidos através de equipamentos GSM e a banda de frequências correspondente não incluir a frequência fundamental, a extracção de NRDs não pode ser efectuada sobre estes – como será evidente aquando da exposição dos fundamentos dos *Normalized Relative Delays*.

As partes do estudo realizado pertinentes para este trabalho paralelo são referentes às características *Mel-Frequency Cepstral Coefficients* e à modelação com *Gaussian Mixture Models*. É importante conhecer a viabilidade destes métodos e o desempenho que atingem antes de os aplicar num caso prático de identificação. É também através do estudo dos métodos com a base de dados TIMIT que será estudada a influência de certos parâmetros internos do sistema, como o número de gaussianas, de forma a conhecer qual a parametrização que permite desempenho óptimo, em diversas situações.

Em suma, a implementação do sistema de reconhecimento de orador foi feita com o trabalho de investigação descrito acima em mente. Devido a isto optou-se também por um cenário de identificação de orador *closed-set*, em que é considerado que o orador do segmento de teste se encontra obrigatoriamente entre os modelos dos oradores treinados. Não é tomada portanto, em caso algum, a decisão de que o orador de teste não corresponde a nenhum dos oradores conhecidos.

3.2 *Gaussian Mixture Models*

Tendo sido GMMs o método escolhido para modelação e classificação no sistema implementado (ver ponto 3.4.3), é importante aprofundar o funcionamento deste e dos algoritmos que o constituem - como o algoritmo *Expectation Maximization*. Para além disto, este está directamente relacionado com o valor de *likelihood*, medida de classificação utilizada neste trabalho.

3.2.1 O algoritmo *Expectation Maximization*

O algoritmo *Expectation Maximization* (EM) é um método frequentemente utilizado para simplificar problemas de maximização de *likelihood* [41]. Como enunciado no capítulo 2.4.2, a função de *log-likelihood* é dada por:

$$b_j(o_t) = \sum_{m=1}^{M} c_{jm} \frac{1}{\sqrt{2\pi|\Sigma_{jm}|}} \exp\left((o_t - \mu_{jm})^T \Sigma_{jm}^{-1} (o_t - \mu_{jm})\right). \tag{3.1}$$

Pretende-se encontrar os valores c_{jm}, Σ_{jm} e μ_{jm} que maximizam $b_j(o_t)$. Para isto são efectuados os seguintes passos: são escolhidos valores de inicialização do modelo, calcula-se a probabilidade *a posteriori* através dos valores actuais das variáveis, tal como expresso na equação 3.2 (passo E) e de seguida usa-se a probabilidade calculada para actualizar os valores das variáveis (passo M), de acordo com as equações 3.3, 3.4 e 3.5 [41]. A inicialização do modelo pode ser feita de forma aleatória ou através de um algoritmo que otenha uma primeira estimativa dos parâmetros das gaussianas, como o *K-means clustering*.

$$\gamma(z_j) = \frac{c_j N(o_t|\mu_j \Sigma_j)}{\sum_{m=1}^{M} c_{jm} N(o_t|\mu_{jm}, \Sigma_{jm})} \tag{3.2}$$

$$\mu_j = \frac{1}{N_j} \sum_{d=1}^{D} \gamma(z_{jd}) o_{td}, \tag{3.3}$$

$$\Sigma_j = \frac{1}{N_j} \sum_{d=1}^{D} \gamma(z_{jd})(o_{td} - \mu_j)(o_{td} - \mu_j)^T, \tag{3.4}$$

$$c_j = \frac{N_j}{N}, \tag{3.5}$$

com

$$N_j = \sum_{d=1}^{D} \gamma(z_{jd}). \tag{3.6}$$

O *log-likelihood* está relacionado com o *fitting* do modelo GMM aos dados de treino. Ao fazer a maximização do valor de *log-likelihood* faz-se uma aproximação das gaussianas constituintes do modelo aos dados. Este significado subjacente torna-se evidente ao reformular a expressão do *likelihood*. De forma a tornar os cálculos mais perceptíveis, parte-se da expressão de *likelihood* para matrizes de covariância diagonais:

$$b_j(o_t) = \prod_{d=1}^{D} \frac{1}{\sqrt{2\pi\sigma_{jd}^2}} \exp\left(-\frac{1}{2} \frac{(o_{td} - \mu_{jd})^2}{\sigma_{jd}^2}\right). \tag{3.7}$$

Calculando o logaritmo de 3.7, obtém-se a seguinte expressão:

$$\begin{aligned} \log b_j(o_t) &= -\frac{1}{2} \sum_{d=1}^{D} \left[\log\left(2\pi + \sigma_{jd}^2\right) + \frac{(o_{td} - \mu_{jd})^2}{\sigma_{jd}^2}\right] \\ &= -\frac{1}{2} \sum_{d=1}^{D} \log\left(2\pi + \sigma_{jd}^2\right) - \frac{1}{2} \sum_{d=1}^{D} \frac{(o_{td} - \mu_{jd})^2}{\sigma_{jd}^2} \end{aligned} \tag{3.8}$$

em que D é a dimensão do conjunto de dados. A segunda parcela da equação 3.8 está relacionada

com uma medida de distância frequentemente utilizada. Na realidade, o *log-likelihood* é calculado a partir da distância Mahalanobis dos pontos pertencentes aos dados ao centro das várias gaussianas do modelo GMM. Dai o *log-likelihood* ser uma medida do *fitting* dos modelos GMM aos dados. No cenário de reconhecimento de orador, o maior *log-likelihood* de um segmento relativamente a um modelo GMM, comparativamente a outro, significa que o orador do segmento exibe maior proximidade ao orador treinado nesse modelo.

Este valor de *log-likelihood* é usado no sistema implementado, à semelhança de vários outros sistemas baseados em GMMs, como a medida de classificação dos oradores. Dado que *log-likelihood* traduz o *fitting* do modelo aos dados, o modelo seleccionado como o mais provável de corresponder ao segmento de teste é o modelo que apresenta maior *log-likelihood*. Desta forma, no sistema desenvolvido no âmbito deste trabalho, a decisão é feita directamente a partir do máximo *log-likelihood* calculado.

3.3 *Normalized Relative Delays*

Como foi já analisado no capítulo 2.3.3 do Estado da Arte, as técnicas mais comuns de caracterização do sinal de fala baseiam-se no modelo fonte-filtro, e representam tipicamente as ressonâncias do tracto vocal. Apesar de a fonte glótica conter informação especifica de orador, esta não é geralmente considerada.

Os coeficientes NRDs representam uma solução para o problema enunciado, pois fazem a modelação dos atrasos relativos entre os diferentes harmónicos, reflectindo a contribuição da fonte glótica e do atraso de grupo do filtro formado pelo tracto vocal. Vários estudos têm comprovado que esta informação de fase é bastante importante para a percepção humana de um som periódico [6].

Para além dos estudos já realizados com os *Normalized Relative Delays*, que indicam que a informação de fase tem capacidade discriminatória no tocante à distinção de cantores e de vogais cantadas (como será visto no ponto seguinte), outros estudos que têm sido efectuados sobre a informação de fase reforçam também a motivação para a investigação neste tipo de características pelos resultados auspiciosos que apresentam.

Um desses estudos [42], publicado já após a terminação dos testes realizados com os NRDs, demonstra que através de um sistema baseado em GMMs foi possível atingir níveis de precisão na identificação de orador idênticos aos atingidos com os MFCCs.

3.3.1 *Normalized Relative Delays*

Nesta secção é descrito o conceito de *Normalized Relative Delays* e o algoritmo de extracção dos mesmos, segundo explicitado em [7].

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

$$
\begin{aligned}
s[n] &= A_0 \sin(n\omega_0 + \varphi_0) + \sum_{i=1}^{M-1} A_i \sin(n\omega_i + \varphi_i) \\
&= A_0 \sin \omega_0 (n + n_0) + \sum_{i=1}^{M-1} A_i \sin(\omega_i (n + n_i)) \qquad (3.9)
\end{aligned}
$$

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 efectuar uma transformação para o domínio das frequências, através de uma transformada como a DFT, obtêm-se os coeficientes espectrais $X[k] = DFTx[n]$, 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 $h[n]$. Como esta janela apresenta simetria par, o atraso de grupo é constante e independente da frequência; de facto, o atraso de grupo corresponde ao centro da janela utilizada. Assim, sendo N o tamanho da janela e da transformada DFT, o atraso de grupo é dado por $(N-1)/2$. Desta forma pode ser omitida da expressão 3.9 a variável n e a expressão pode ser reformulada de modo a destacar o conceito de *Normalized Relative Delays* (NRDs):

$$
\begin{aligned}
s &= A_0 \sin(n_0 \omega_0) + \sum_{i=1}^{M-1} A_i \sin(n_i \omega_i) \\
&= A_0 \sin\left(2\pi \frac{n_0}{P_0}\right) + \sum_{i=1}^{M-1} A_i \sin\left(2\pi \frac{n_i}{P_i}\right) \\
&= A_0 \sin\left(2\pi \frac{n_0}{P_0}\right) + \sum_{i=1}^{M-1} A_i \sin\left(2\pi \frac{n_0 + n_i - n_0}{P_i}\right) \\
&= A_0 \sin\left(2\pi \frac{n_0}{P_0}\right) + \sum_{i=1}^{M-1} A_i \sin\left[2\pi \left(\frac{n_0}{P_i} + \frac{n_i - n_0}{P_i}\right)\right] \\
&= A_0 \sin\left(2\pi \frac{n_0}{P_0}\right) + \sum_{i=1}^{M-1} A_i \sin\left[2\pi \left(\frac{n_0}{P_i} + NRD_i\right)\right]. \qquad (3.10)
\end{aligned}
$$

em que P_i representa o período da sinusóide i.

Como pode ser observado em 3.10, NRD_i representa o atraso relativo normalizado (*normalized relative delay*) do harmónico de índice i. Estes coeficientes NRD caracterizam a forma de onda do sinal, independentemente do deslocamento no tempo e da frequência fundamental do sinal.

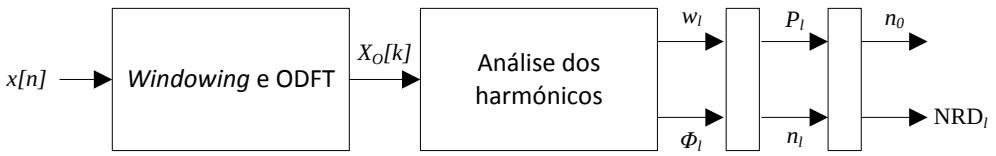

Figura 3.1: Algoritmo de estimação dos NRDs e do deslocamento temporal n_0. Figura adaptada de [6].

A estimação dos NRDs é feita da seguinte forma: 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 com o intuito de 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 ($n_i = \varphi_i/\omega_i$), que por sua vez permitem calcular os NRDs, como explicitado anteriormente.

3.3.2 Estudos preliminares

Vários testes foram já realizados de forma a estudar aprofundadamente os coeficientes NRDs e o seu poder discriminatório relativo a diversas características. Em [7] foram utilizados os NRDs para classificação de dois tipos diferentes de características: classificação de vogal e classificação de cantor; enquanto que em [6] foram usados na tarefa de e classificação de tipo de fonação. Os principais resultados obtidos através dos estudos mencionados são apresentados de seguida.

A classificação de vogal foi feita sobre vogais cantadas por diferentes oradores (referidos como cantores neste caso). A base de dados utilizada nestes testes é descrita no capítulo 4.2.2, e denomina-se "Vogais Cantadas". O classificador escolhido, pelo desempenho superior que demonstrou em testes preliminares, foi o *Nearest Neighbour*, em combinação com o método de *cross-validation 10-fold*, ambos utilizados no ambiente de análise estatística WEKA. Como medida de desempenho recorreu-se à *F-measure*:

$$\text{F-measure} = \frac{2 \times \text{TPR} \times \text{Precisão}}{\text{TPR} + \text{Precisão}} \tag{3.11}$$

em que a Precisão representa o rácio das instâncias classificadas correctamente como X pelas instâncias classificadas como X, e TPR (*True Positive Rate*) o rácio entre as instâncias correctamente classificadas como X e o número de instâncias do tipo X.

Os valores de *F-measure* obtidos encontram-se resumidos na figura 3.2:

O desempenho dos NRDs foi comparado com o de MFCCs e com características relacionadas com a estimativa do impulso glótico. Como se pode observar, os NRDs possuem poder discriminatório bastante competitivo quando comparado ao dos outros conjuntos de características considerados.

Na tarefa de classificação de cantor os NRDs apresentam comportamento semelhante, como se pode observar na figura 3.3

3.4 Implementação

No presente subcapítulo será descrito o sistema de reconhecimento de orador implementado no âmbito da dissertação e do estudo realizados resultante da colaboração com a Polícia Judiciária (descrito em 3.1). Este sistema é baseado na abordagem mais frequente no estado da arte: modelação de características espectrais através de *Gaussian Mixture Models*. Numa primeira parte são apresentadas as principais ferramentas utilizadas no estudo e desenvolvimento dos algoritmos.

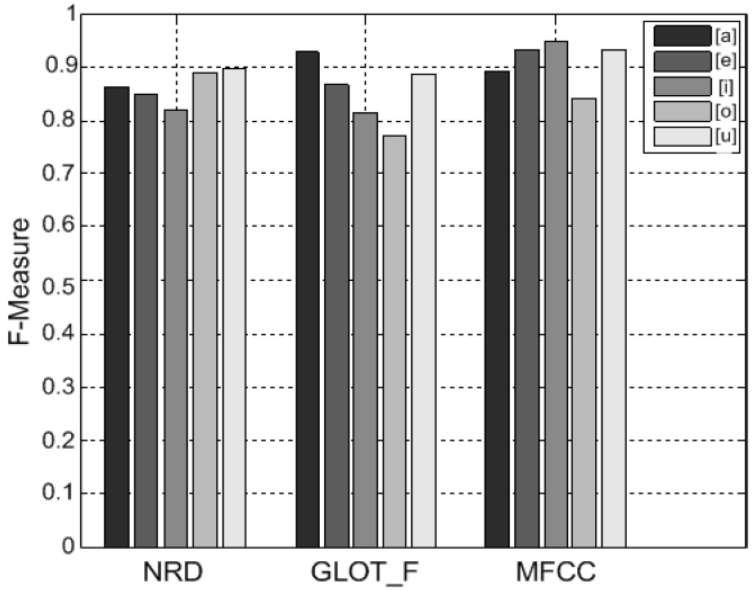

Figura 3.2: Classificação de vogais cantadas através das características: NRD, características do fluxo glótico (GLOT_F) e MFCC. Figura adaptada de [7].

Nos seguintes subcapítulos aborda-se a implementação do sistema dividida em duas componentes principais: a extracção de características *Mel-Frequency Cepstral Coefficients* e a modelação GMM. Com referência a cada uma destas são descritos os algoritmos utilizados e os principais parâmetros de funcionamento destes.

3.4.1 Ferramentas utilizadas

A escolha das ferramentas utilizadas para a implementação do sistema de reconhecimento de orador baseou-se essencialmente nos seguintes factores: facilidade de utilização; número de algoritmos e *toolboxes* disponíveis compatíveis com a ferramenta seleccionada e aplicáveis a um sistema de reconhecimento e número de estudos realizados com base na mesma ferramenta, isto é, nível de robustez dos algoritmos comprovada por várias fontes. Com base neste factores foram seleccionadas as ferramentas de seguida apresentadas.

3.4.1.1 Matlab

Devido à grande aceitação do Matlab tanto em ambiente académico como empresarial para estudo e desenvolvimento de algoritmos, o desenvolvimento da aplicação foi baseado nesta plataforma. O facto de ser extensivamente utilizado leva a que se encontrem disponíveis vários algoritmos de livre acesso úteis para o desenvolvimento da aplicação em causa, e garante também que vários destes algoritmos tenham sido utilizados em vários estudos, tendo sido já corrigidos, estendidos e melhorados desde a sua disponibilização. Uma *toolbox* que inclui vários algoritmos

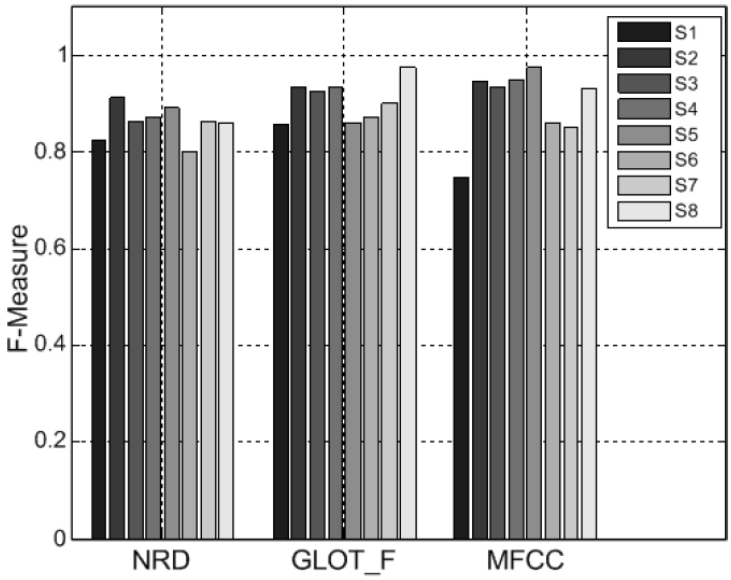

Figura 3.3: Classificação de cantor através das características: NRD, características do fluxo glótico (GLOT_F) e MFCC. Figura adaptada de [7].

utilizados na área de reconhecimento de orador que satisfaz estes critérios é a Voicebox, a ser descrita seguidamente.

3.4.1.2 Voicebox

A Voicebox [43] foi criada por Mike Bookes em 1998 e é alvo ainda hoje de expansão e optimização. Esta *toolbox* inclui vários métodos relacionados com processamento de sinal, processamento da voz e de fala. Para além de ser bastante abrangente em termos de algoritmos implementados, estes geralmente apresentam várias possibilidades de configuração através da escolha de parâmetros de funcionamento. Para além disto é uma *toolbox* frequentemente referida na literatura, tendo sido utilizada em diversos estudos na área.

3.4.1.3 Weka

O *software open-source* Weka [44] (Waikato Environment for Knowledge Analysis) é uma ferramenta de análise estatística que disponibiliza vários algoritmos de *machine learning* e reconhecimento de padrões. O *software* permite importar uma base de dados e a partir desta fazer pré-processamento dos dados, classificação, *clustering*, selecção de atributos mais significativos, entre outras funcionalidades. Esta ferramenta, cuja primeira versão se encontra disponível desde 1992, tem vindo a expandir-se desde o seu lançamento, tanto a nível de algoritmos integrados como a nível de número de utilizadores. É actualmente uma ferramenta de referência em *data mining* em ambiente académico e empresarial [45].

Embora o Weka não tenha sido directamente utilizado na implementação do sistema baseado em GMMs, este *software* foi aplicado no estudo das características de voz extraídas (análise estatística das mesmas) e como tal auxiliou no desenvolvimento da solução implementada. Para além disto, o Weka foi também utilizado para modelação e classificação no caso prático estudado, como é descrito no capítulo 4.5.

3.4.2 Extracção de características MFCC

Como referido no capítulo 2, as características mais frequentemente utilizadas no estado da arte são MFCCs, pelo facto de obterem os níveis de precisão mais elevados. Como tal, é com base nestes que se realiza parte do estudo descrito neste documento. O sistema base GMM+MFCC é utilizado para comparação dos novos métodos estudados, com o intuito de aferir se proporcionam melhor desempenho no reconhecimento com segmentos de reduzida duração.

3.4.2.1 Código

Foram utilizadas funções incluídas na Voicebox para extracção dos coeficientes MFCC. Testes preliminares foram realizados com a Auditory Toolbox [46], e foi concluído que, em integração no sistema de reconhecimento de orador, os resultados obtidos são bastante comparáveis, embora ligeiramente inferiores para o caso da Auditory Toolbox. Por outro lado, a função da Voicebox permite a selecção de um conjunto alargado de parâmetros, tornando a implementação de várias variantes de extracção de MFCCs mais imediata.

3.4.2.2 Parametrização dos algoritmos de extracção de características

O número de coeficientes extraído (excluindo coeficientes derivados) foi de 13, excluindo o de ordem zero. Este número foi seleccionado com base na média escolhida nos sistemas descritos na literatura (alguns exemplos: 13 coeficientes em [16] e 12 em [22]). O tamanho da janela de extracção é, excepto quando indicado o contrário, calculado a partir da expressão $2^{\lfloor \log_2(0{,}03 f_s) \rfloor}$ (em que f_s é a frequência de amostragem do segmento), de acordo com a parametrização por omissão da função Voicebox. Por fim, o avanço da janela de extracção é de metade do tamanho da janela, ou seja, aplica-se um *overlap* de 50% de forma a garantir uma caracterização mais detalhada do sinal.

Todos os outros parâmetros de extracção dos coeficientes MFCC foram os definidos por omissão na função utilizada da Voicebox.

3.4.3 Modelação e classificação GMM

Seguindo o mesmo critério de selecção aplicado na escolha das características MFCC, o método de modelação e classificação baseado em GMM foi aplicado no sistema desenvolvido pelo facto de ser actualmente o sistema mais frequentemente utilizado e que atinge níveis de precisão mais elevados. Para além disto, apesar de no estado da arte se conjugar geralmente GMMs com

outros métodos de forma a melhorar o desempenho (como combinação com SVMs, métodos de normalização, entre outros), os modelos GMM em funcionamento separado resultam em níveis de *performance* elevados, comparativamente a outros métodos isolados.

3.4.3.1 Código

Novamente, a Voicebox foi utilizada para criação dos modelos GMM e cálculo dos valores de *log-likelihood* utilizados para classificação. De forma a aferir a robustez dos algoritmos foram realizados testes com as funções incluídas na Statistics Toolbox [47] - *toolbox* geralmente distribuída com o Matlab. Realizaram-se vários testes de classificação (com segmentos da base de dados TIMIT - ver capítulo 4.2.1) e através dos valores de *log-likelihood* obtidos concluiu-se que ambas as *toolboxes* devolviam os mesmos resultados. Concluiu-se assim que as duas ferramentas são equivalentes para a finalidade em questão, e foi possível consequentemente validar os métodos incluídos na Voicebox.

3.4.3.2 Parametrização dos algoritmos de modelação

Na construção de um modelo GMM o parâmetro com maior impacto na modelação adequada dos dados é o número de gaussianas que constituem o modelo. A utilização de um número demasiado reduzido de componentes gaussianas pode conduzir à construção de um modelo que não caracteriza os dados introduzidos para treino, enquanto que a utilização de um número demasiado grande de gaussianas provoca *overfitting*: o modelo irá traduzir variações demasiado pequenas dos dados, incluindo ruído, fazendo com que o modelo não caracterize devidamente o orador. Por outro lado, o número de gaussianas deve ser escolhido em função da dimensão dos dados, entre outros factores. Assim, o número de gaussianas é também alvo de estudo na realização dos testes, descritos no próximo capítulo.

Para todos os modelos GMMs criados foram utilizadas matrizes de covariância diagonais, pelo facto de na literatura esta ser a opção mais comum. A utilização de matrizes diagonais reduz significativamente o tempo de processamento e a complexidade computacional, permitindo alcançar no entanto o mesmo desempenho que as matrizes completas.

Para melhor desempenho, o método de inicialização do algoritmo *Expectation Maximization*, com os valores iniciais de média, variância e pesos associados a cada componente gaussiana, é feito através do algoritmo de *clustering K-harmonic means*. É aplicado este método antes de treinar o modelo gaussiano, e em função dos *clusters* formados são dados ao algoritmo EM os valores de inicialização.

3.5 Conclusão

Neste capítulo foram aprofundados dois conceitos fundamentais no desenvolvimento do sistema de reconhecimento realizado nesta dissertação. O primeiro é referente ao algoritmo de estimação dos parâmetros dos modelos GMM actualmente mais utilizado, o algoritmo *Expectation*

Maximization. No mesmo ponto foi feita uma análise do cálculo do valor de *likelihood*, o qual o algoritmo EM procura maximizar. Dado que a classificação dos segmentos de voz no presente trabalho é feita através deste valor, é importante compreender o seu cálculo e o seu significado no contexto presente.

Seguidamente foram descritas as novas características de voz que são exploradas nesta dissertação, os *Normalized Relative Delays*. Para além de ser explorado o algoritmo que leva à sua extracção, são apresentados estudos publicados sobre o desempenho que estes coeficientes atingem em diferentes tarefas de classificação, em diferentes condições. Estabeleceu-se que os NRDs, na identificação de vogal e na identificação de cantor, obtêm níveis de precisão bastante elevados e comparáveis aos obtidos com as características do estado da arte, MFCCs. Estes resultados motivam o estudo do uso de NRDs em segmentos de voz falada, num sistema de reconhecimento de orador. No capítulo seguinte são apresentados os testes realizados com este fim.

Por fim, foram listadas as principais ferramentas utilizadas no desenvolvimento da solução implementada, assim como quais os principais parâmetros de funcionamento dos algoritmos utilizados. Pretende-se assim dar a conhecer todas as condições que conduziram às conclusões retiradas. Assim, aquando da apresentação dos resultados no capítulo 4, apenas os parâmetros que foram alvo de variação são mencionados.

Capítulo 4

Testes e Resultados

Neste capítulo são descritos os procedimentos realizados com o intuito de medir o desempenho do sistema de reconhecimento de orador apresentado no capítulo anterior. A partir dos resultados obtidos é possível inferir sobre a influência dos parâmetros internos do sistema no desempenho deste (como o número de componentes Gaussianas utilizadas) e sobre a influência das condições proporcionadas pelos dados de entrada (como duração do treino e de teste).

Na primeira parte do capítulo é apresentada a estrutura subjacente aos testes realizados e as bases de dados utilizadas; na segunda parte são descritos os testes realizados ao sistema tendo como base as características *Mel-Frequency Cepstral Coefficients* e as características *Normalized Relative Delays*. Por último é apresentado o caso prático em que um dos métodos desenvolvido ao longo da dissertação é aplicado, e são analisados os resultados obtidos.

4.1 Apresentação dos testes realizados

Neste subcapítulo é feita uma descrição sumária dos testes realizados ao sistema apresentado no capítulo 3. Numa primeira parte aborda-se a organização subjacente aos testes e são resumidas as principais variantes entre cada conjunto de testes. Nesta apresentação não é justificada a escolha dos parâmetros ou ferramentas utilizados; a motivação para essas opções encontra-se explicitada ao longo dos capítulos 4.3 e 4.4.

4.1.1 Organização dos testes

A execução dos testes foi realizada em duas fases principais. Numa primeira fase foi estudado o desempenho do sistema implementado, em que as características da voz utilizadas são MFCCs. Todos os testes realizados incorporam segmentos de voz retirados da base de dados TIMIT, a ser descrita no próximo subcapítulo. No entanto, diferentes partes do sinal de fala foram consideradas para testar a solução desenvolvida. Inicialmente foi usado todo o segmento de voz, excluindo intervalos de silêncio do sinal; o teste seguinte foi realizado apenas com as vogais presentes nos

Tabela 4.1: Tabela resumo dos testes realizados ao sistema de reconhecimento de orador implementado.

Características	**Base de Dados**	**Tipo de Voz**
MFCCs	TIMIT	Completa
		Vogais
		Vozeada
		Não Vozeada
MFCCs e NRDs	TIMIT	Vogais
	Vozes Cantadas	
	Vogais Faladas	

segmentos; no último teste a segmentação foi feita de forma a separar as partes vozeadas das partes não vozeadas da fala, e realizaram-se dois estudos correspondentes a estes dois tipos de voz.

Numa segunda fase de testes as características em foco são NRDs. No entanto, dado que em alguns casos as condições de teste não são equiparáveis aos testes anteriores, são utilizadas novamente as características MFCCs e extraídos novos resultados, de forma a serem comparados directamente com os resultados obtidos com NRDs. Por último, combinam-se MFCCs e NRDs e o conjunto das duas características é dado ao sistema para modelar os oradores. Várias bases de dados foram utilizadas para obtenção de resultados – a TIMIT, à semelhança da primeira fase de testes, uma base de dados constituída por vogais cantadas e uma base de dados constituída por vogais faladas. Mais detalhes sobre a escolha destas bases de dados encontram-se mais adiante na dissertação, no entanto indica-se desde já que a base de dados de vogais cantadas foi utilizada com o intuito de recriar as condições de teste dos estudos preliminares a que os NRDs foram submetidos (secção 3.3.2). Para estes *corpus* mencionados, os segmentos seleccionados incluem apenas vogais.

4.2 Bases de Dados

4.2.1 TIMIT

A base de dados fonética acústica de fala contínua TIMIT foi desenvolvida pelo Massachusetts Institute of Technology (MIT) conjuntamente com a Texas Instruments, Inc. e a SRI International com o objectivo de fornecer um conjunto de dados que pudesse ser utilizado em estudos acústico-fonéticos e no desenvolvimento de sistemas de reconhecimento de fala. Este *corpus* encontra-se amplamente divulgado e, pelo facto de compreender um universo bastante alargado de oradores e por ter documentação de apoio bastante extensiva, é uma das bases de dados mais frequentemente utilizadas no teste de sistemas de reconhecimento de orador.

A TIMIT contém 630 oradores, 438 masculinos e 192 femininos, distribuídos por 10 dialectos diferentes de Inglês Americano, de acordo com a região dos Estados Unidos da América em que

o orador passou a sua infância. Cada orador lê 10 frases, sendo que duas delas são repetidas por todos os oradores, de forma a evidenciar as diferenças entre dialectos.

Para cada frase existem quatro ficheiros disponíveis – um ficheiro áudio de extensão ".wav", um ficheiro de texto com a transcrição da frase proferida no ficheiro áudio, um ficheiro de extensão ".wrd" com a delimitação temporal de cada palavra da frase (expressa em amostras) e um ficheiro de extensão ".phn" com a transcrição fonética da frase e a delimitação de cada fonema, novamente em amostras. Os ficheiros áudio foram gravados à frequência de amostragem de 16 kHz, no formato PCM mono de 16 bit.

4.2.2 Vogais Cantadas

Esta base de dados inclui ficheiros áudio de oito cantores diferentes, tanto masculinos como femininos. Para cada um dos cantores estão disponíveis as cinco vogais cantadas, com duração média de 200 milissegundos, sendo que cada segmento de voz foi previamente seleccionado de forma a incluir apenas o intervalo de tempo em que a forma de onda se encontra mais "estável" – geralmente a parte central do fonema. Nesta parte central existe variação mínima entre *frames*. Os ficheiros áudio encontram-se em formato PCM, com um canal de 16 bit e foram gravados com a frequência de amostragem de 44100 Hz.

4.2.3 Vogais Faladas

Por último, recorreu-se a outra base de dados constituída apenas por vogais [48]. São ditas as cinco vogais por um conjunto alargado de oradores – 27 crianças e 17 adultos, 11 dos quais femininos e 6 masculinos. Em semelhança à base de dados de vogais cantadas, os segmentos presentes nesta base de dados incluem apenas o intervalo estável do sinal, sendo excluídos os intervalos iniciais e finais em que podem haver regiões de transição entre fonemas, entre outros factores. Cada vogal tem duração exacta de 100 milissegundos.

4.3 Desempenho das características MFCC

Sendo os MFCCs considerados o estado da arte no que toca à extracção de características de voz em sistemas de reconhecimento de orador, o teste da solução desenvolvida no âmbito desta dissertação passa naturalmente por determinar o nível de desempenho que esta atinge utilizando este tipo de características. Devido à vasta literatura disponível sobre reconhecimento de orador utilizando modelação GMM e características MFCC, é possível comparar os resultados obtidos com resultados publicados por vários autores, e assim validar a solução implementada.

Tendo concluído sobre a capacidade dos métodos utilizados baseados em características MFCC, é pertinente estudar duas adições feitas frequentemente a estas características: normalização *cepstral mean subtraction* (CMS) e inclusão de características MFCC diferenciais - ΔMFCC e ΔΔ MFCC. Desta forma pretende-se determinar o nível de desempenho máximo que é possível atingir, em diversas condições de teste, através dos métodos que constituem o estado da arte. Este

estudo tem por objectivo identificar os cenários de funcionamento do sistema em que se observam níveis de desempenho satisfatórios e em que cenários ocorre, pelo contrário, uma diminuição significativa da taxa de identificação correcta de orador. Espera-se que através das características *Normalized Relative Delays* seja possível melhorar o desempenho geral do sistema de reconhecimento, mas em especial melhorar o desempenho nas condições de funcionamento que revelem ser um desafio para o método baseado em MFCC apenas. É com o mesmo intuito que é também estudada a influência da segmentação selectiva de partes da fala – vogais, partes vozeadas e partes não vozeadas.

4.3.1 Normalização *Cepstral Mean Subtraction* e MFCCs diferenciais

O estudo da influência de Δ MFCC e $\Delta\Delta$MFCC na *performance* do sistema implementado não pode ser feito independentemente de outros parâmetros de funcionamento do sistema, nomeadamente dos parâmetros de modelação GMM dos oradores. Ao duplicar (no caso de inclusão de ΔMFCC) ou triplicar (no caso de inclusão de ΔMFCC e de $\Delta\Delta$MFCC) a dimensão dos vectores de características dados para treino dos modelos GMM, torna-se necessário averiguar novamente o número de componentes que serão necessárias para uma correcta modelação dos dados. Este número de componentes deve ser elevado o suficiente de forma ao modelo traduzir a individualidade de cada orador, mas não tão elevado que ocorra *overfitting*. Por este motivo, para cada configuração de características (combinação entre MFCCs, MFCCs diferenciais e CMS), foi variado o número de componentes gaussianas utilizadas para construção do modelo GMM do orador entre 8 e 100.

Os dados utilizados para este teste provêm da base de dados TIMIT, e pertencem ao conjunto dos dialectos D1 e D2 (dialecto de New England e dialecto *Northern*, respectivamente), dado que foi seleccionada uma porção da base de dados para simplificar e agilizar os testes realizados. Os oradores seleccionados são apenas oradores masculinos, pelo facto de a inclusão de oradores femininos tornar a tarefa em causa mais complexa, como visto no estado da arte. Vários sistemas actualmente utilizam estratégias específicas para lidar com reconhecimento de oradores masculinos e de oradores femininos simultaneamente. Dado que este não é o foco do presente trabalho, foi decidido excluir oradores femininos.

Por fim, foram realizados testes em universos de oradores de dimensão diferente – 8, 20 e 40 oradores. Esta variação foi incluída neste teste pelo facto do número de oradores influenciar também o número óptimo de gaussianas a utilizar no modelo GMM – um universo maior exige um modelo mais complexo.

Para cada orador foram seleccionados sete segmentos para treino e deixados para teste os restantes três. Cada segmento de teste foi testado contra o modelo de cada orador e foi feita uma decisão baseada no valor *log-likelihood* resultante de cada teste. Assim, para um universo de N oradores são realizadas $3N$ testes.

Os resultados representados na figura 4.1 dizem respeito ao teste com 8 oradores.

Como se pode concluir pela análise da figura, a extensão do vector de características com os coeficientes ΔMFCC e $\Delta\Delta$MFCC tende a provocar uma diminuição do desempenho do sistema.

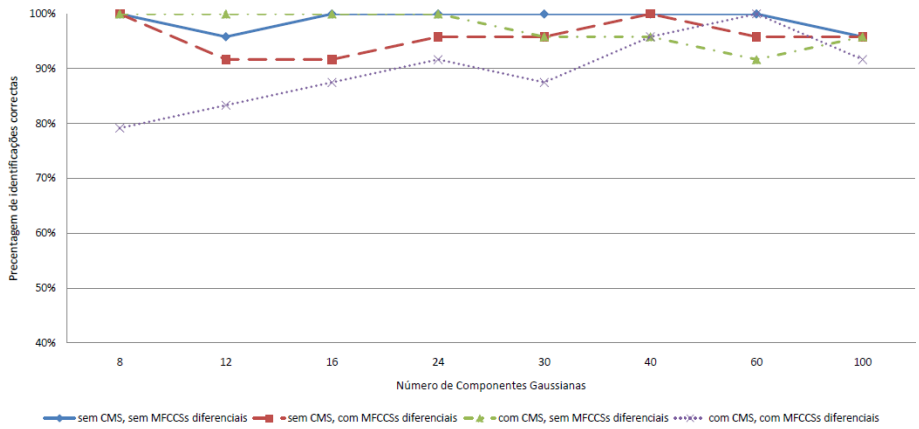

Figura 4.1: Percentagem de identificações correctas num cenário de 8 oradores, com variação de diversos parâmetros – número de componentes gaussianas, uso de normalização CMS e inclusão de MFCCs diferenciais.

As duas curvas que se encontram geralmente mais abaixo no gráfico são as que representam os resultados obtidos com MFCCs diferenciais. Por outro lado, a curva referente aos resultados obtidos sem normalização e sem acréscimo de características (linha contínua do gráfico 4.1) apresenta na maioria dos casos o desempenho máximo.

No entanto, a posição relativa dos métodos em termos de desempenho não se mantém constante para qualquer número de gaussianas usadas. Por este facto, partindo apenas destes resultados, é difícil concluir sobre a contribuição dos métodos em estudo para a tarefa de identificação de orador. Assim, é feita seguidamente a análise dos mesmos testes realizados em cenários de identificação mais complexos – com 20 e 40 oradores.

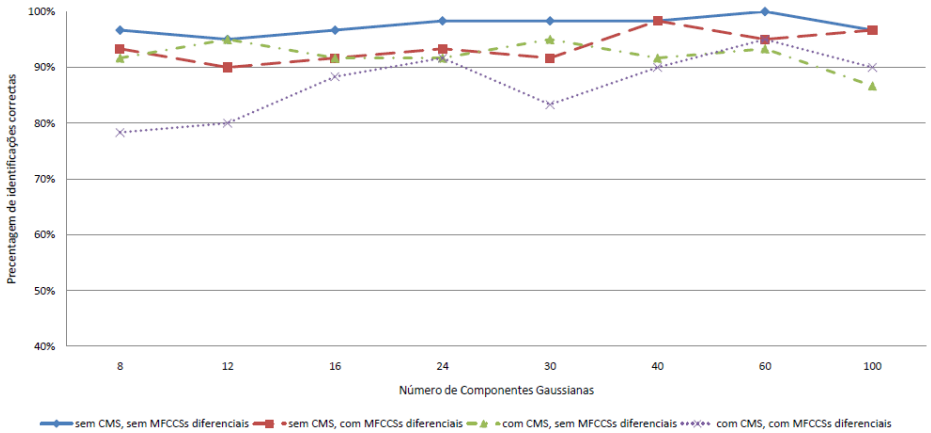

Figura 4.2: Percentagem de identificações correctas num cenário de 20 oradores, com variação de diversos parâmetros – número de componentes gaussianas, uso de normalização CMS e inclusão de MFCCs diferenciais.

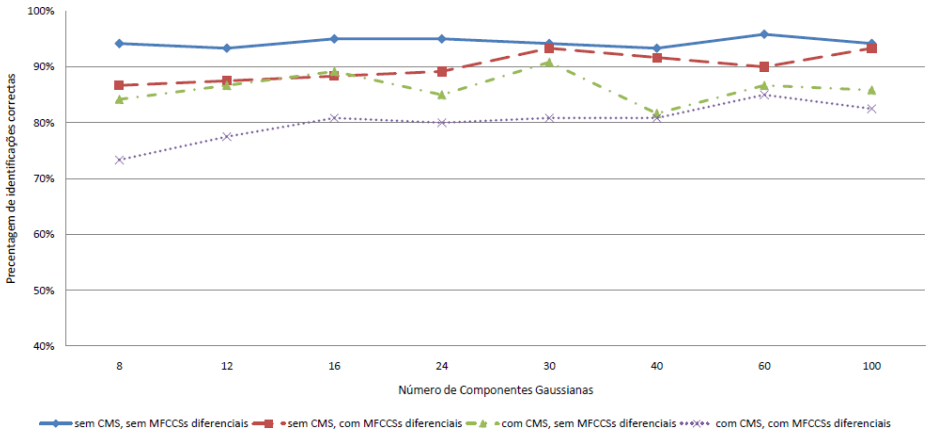

Figura 4.3: Percentagem de identificações correctas num cenário de 40 oradores, com variação de diversos parâmetros – número de componentes gaussianas, uso de normalização CMS e inclusão de MFCCs diferenciais.

Os resultados obtidos com universos de oradores de maior dimensão – 20 e 40 oradores – exibem um grau de disparidade bastante maior entre os diferentes cenários de teste em causa. Dois destes cenários destacam-se claramente em relação aos restantes: o de uso simultâneo de CMS e de MFCCs diferenciais e o caso oposto, ou seja, sem normalização e usando apenas os 13 coeficientes MFCC. O primeiro cenário resulta em percentagens de identificação correcta bastante abaixo da média dos restantes resultados, enquanto que o segundo obteve as maiores taxas de identificação para qualquer número de componentes.

A melhor separação entre as diferentes curvas nos gráficos referentes à identificação entre 20 e 40 oradores, relativamente ao gráfico referente à identificação com 8 oradores, deve-se ao facto de este último cenário ser muito simples e levar facilmente a níveis de precisão elevados, independentemente da eficácia do método utilizado. Por outro lado, dado que o número de testes é proporcional ao número de oradores considerados, o cálculo da percentagem de identificações correctas torna-se mais fiável à medida que a dimensão do universo de oradores aumenta.

Em suma, através da observação dos gráficos apresentados é possível concluir que tanto a normalização CMS como as características ΔMFCC e ΔΔMFCC prejudicam a identificação de orador nas condições enunciadas. O efeito da normalização pode ser explicado pelo facto de esta reduzir a influência de variação de canal e de ruído convolutivo nos segmentos áudio utilizados, mas também eliminar paralelamente alguma informação referente ao orador. Visto que a base de dados TIMIT apresenta níveis de ruído e quantidade de artefactos reduzidos, pensa-se que neste caso a normalização CMS tem maior peso na eliminação de informação discriminatória de orador do que na eliminação de efeitos de ruído.

Os resultados obtidos com a inclusão das características MFCC diferenciais são, no entanto, algo surpreendentes. Devido à natureza dos coeficientes diferenciais foi considerada a hipótese de estes serem mais adequados à tarefa de reconhecimento de orador dependente de texto, pois traduzem características relacionadas com a fonética. Para testar a hipótese enunciada, dado que

nenhuma base de dados disponível contém a mesma frase dita múltiplas vezes pelo mesmo orador, foram gravados alguns segmentos de voz. Três oradores leram a mesma frase, resultando em segmentos com cerca de oito segundos. Após ter sido aplicado o mesmo sistema usado anteriormente, a percentagem de identificações correctas obtida foi de 100% em todos os testes, com e sem inclusão de MFCCs diferencias. Este valor deve-se à dimensão reduzida do universo de oradores, e não permite assim tirar conclusões sobre o tópico.

Face aos resultados obtidos, os testes posteriores foram realizados com os 13 coeficientes MFCC considerados inicialmente, e foi excluído o método de normalização.

4.3.2 Desempenho do sistema utilizando voz completa

Neste ponto é feita a continuação dos testes anteriores, através da variação de um parâmetro não estudado anteriormente: a duração dos segmentos usados para teste.

Na secção anterior cada teste foi realizado com um único segmento de voz da base de dados. Cada segmento, após remoção dos silêncios, tem duração média de dois segundos. Visto que este valor representa já uma duração bastante reduzida e que na maioria das aplicações práticas de um sistema de reconhecimento de orador os segmentos de voz disponíveis têm maior duração, estabeleceu-se a duração indicada como limite mínimo. Assim, de forma a estudar a influência da duração de teste, optou-se por comparar o caso anterior com um caso mais favorável: testes com duração média de 6 segundos. Para este fim foram concatenados os três segmentos de cada orador reservados para teste, havendo agora apenas um segmento de teste disponível para cada orador.

É importante referir que as condições de treino dos modelos dos oradores se mantêm idênticas às do estudo anterior: são extraídas características MFCC de sete segmentos por orador, sendo a matriz resultante os dados de treino. A duração total média dos sete segmentos, após remoção dos silêncios, é de cerca de 20 segundos. Assim, não são realizados $3N$ testes (N sendo o número de oradores), mas sim N testes.

Tabela 4.2: Percentagem de identificações correctas para cenário de 8, 20 e 40 oradores e diferente número de componentes gaussianas.

Nº Componentes	8 oradores	20 oradores	40 oradores
8	100%	100%	100%
12	100%	100%	100%
16	100%	100%	100%
24	100%	100%	100%
30	100%	100%	100%
40	100%	100%	98%
60	100%	100%	100%
100	100%	100%	100%

Através dos resultados apresentados na tabela 4.2 e na figura 4.4 podemos concluir que o desempenho obtido com testes de duração de 6 segundos é "óptimo" em quase todos os cenários apresentados. Apenas um caso apresenta taxa de identificação correcta menor que 100% (98%), e foi obtido para o cenário em que são considerados mais oradores, e para número de componentes

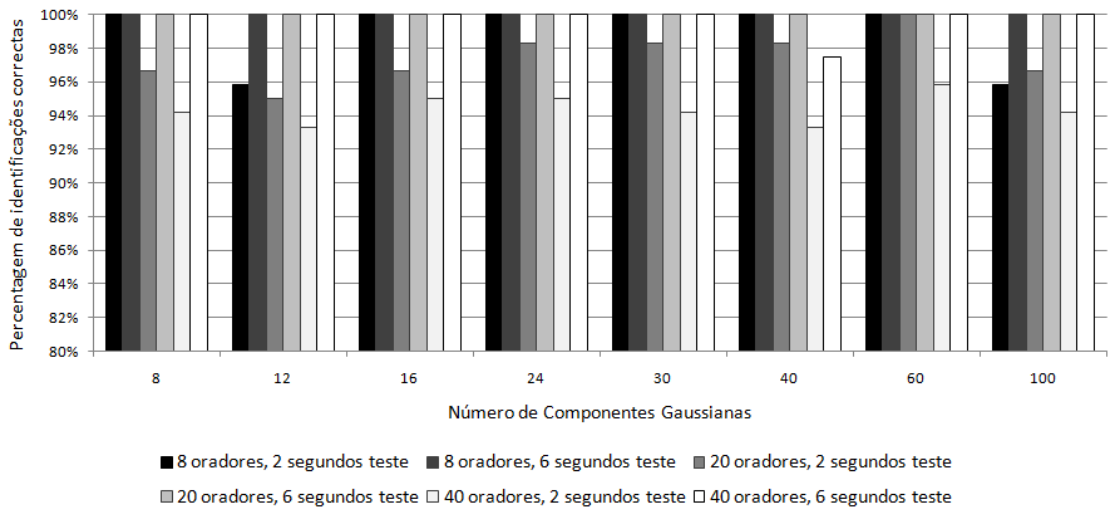

Figura 4.4: Comparação entre desempenho do sistema usando tempo de teste de 2 segundos e 6 segundos. Resultados apresentados para diferente número de gaussianas utilizadas e diferente número de oradores considerados.

gaussianas bastante elevado – 40 componentes. Uma pequena variação local como esta verificada não deve, no entanto, ser tomada como base para conclusões sobre a parametrização ideal do sistema. O treino dos modelos GMM não produz modelos de oradores exactamente idênticos entre diferentes simulações do sistema – isto devido ao facto de a optimização do *fitting* das componentes gaussianas não partir do mesmo ponto inicial.

Verifica-se através deste estudo que para as condições de treino enunciadas o incremento do tempo de teste de 2 para 6 segundos é muito significativo, e que, pelo menos para um universo de oradores de dimensão próxima de 40, 6 segundos são suficientes para se efectuar identificação de orador com níveis de desempenho muito elevados. O cenário de identificação com 2 segundos é portanto o cenário que merece um estudo mais aprofundado e onde a aplicação de diferentes métodos poderá trazer melhorias significativas.

4.3.3 Desempenho do sistema utilizando segmentação da voz de acordo com vozeamento

A fala vozeada e a fala não-vozeada apresentam características demarcadamente distintas: enquanto que a fala vozeada constitui um sinal periódico e estável, a fala não vozeada é de natureza aleatória. Pelas suas características estacionárias, a utilização de apenas as partes vozeadas da voz para a tarefa de reconhecimento de orador pode contribuir para melhor desempenho dos sistemas e abrir novas possibilidades em termos de possíveis características extraídas.

Em [22] e [49], por exemplo, são seleccionadas apenas as partes vozeadas para extracção de MFCCs e do *pitch*. Em [22] é feito o estudo da influência da utilização dos segmentos vozeados em comparação com todo o sinal de fala (excluindo os silêncios) e obtiveram-se melhorias da taxa de identificações correctas até 4,5% e até 6% quando se inclui informação relativa ao *pitch*, num

Símbolo utilizado na TIMIT	Símbolo IPA equivalente	Palavra exemplo	Presença de vozeamento
b	b	bee	sim
d	d	day	sim
g	g	gay	sim
p	p	pea	não
t	t	tea	não
k	k	key	não
q	ʔ	bat	não
jh	dʒ	joke	sim
ch	tʃ	choke	não
s	s	sea	não
sh	ʃ	she	não
z	z	zone	sim
zh	ʒ	azure	sim
f	f	fin	não
th	θ	thin	não
v	v	van	sim
dh	ð	then	sim
m	m	mom	sim
n	n	noon	sim
ng	ŋ	sing	sim
em	əm	bottom	sim
en	ən	button	sim
eng	-	washington	sim
nx	-	winner	sim
l	l	lay	sim
r	r	ray	sim
w	/w/	way	sim
y	j	yacht	sim
hh	h	hay	não
hv	ɦ	ahead	sim
el	əl	bottle	sim
iy	/iː/	beet	sim
ih	ɪ	bit	sim
eh	ɛ	bet	sim
ey	eɪ	bait	sim
ae	æ	bat	sim
aa	ɒ	bott	sim
aw	aʊ	bout	sim
ay	aɪ	bite	sim
ah	ʌ	but	sim
ao	ɔ	bought	sim
oy	ɔɪ	boy	sim
ow	oʊ	boat	sim
uh	ʊ	book	sim
uw	uː	boot	sim
ux	ʉ	toot	sim
er	/ər/	bird	sim
ax	aʊ	about	sim
ix	.ɪ	debit	sim
axr	ɚ	butter	sim
ax-h	ə	suspect	não

Figura 4.5: Fonemas incluídos nas anotações que acompanham a TIMIT.

sistema baseado em GMMs. Para além disto, a duração dos segmentos de teste utilizados varia entre 0,5 e 6 segundos, reforçando a motivação para a segmentação com base no vozeamento em aplicações baseadas em intervalos de voz muito curtos.

De forma a realizar a separação das partes vozeadas e não vozeadas dos segmentos da base de dados TIMIT, foram utilizados os ficheiros ".phn" disponibilizados com a mesma. Estes contêm anotações dos fonemas presentes em cada ficheiro áudio e o instante de início e de fim respectivos. Na figura 4.5 apresentam-se os fonemas presentes nas transcrições utilizadas e os símbolos correspondentes segundo o International Phonetic Alphabet (IPA), assim como a classificação quanto ao vozeamento de cada um destes fonemas, segundo as fontes [50] e [51].

4.3.3.1 Vogais

Após a segmentação descrita acima, foi feita uma selecção das vogais adequadas para uso no sistema implementado. Dos fonemas acima indicados, foram consideradas vogais os fonemas com os seguintes símbolos (conforme a TIMIT):

- iy
- ih
- eh
- ey
- ae
- aa
- aw
- ay
- ah
- ao
- oy
- ow
- vh
- uw
- ux
- er
- ax
- ix
- axr

Estabeleceu-se que as vogais incluídas deveriam ser suficientemente longas de forma a ser possível extrair, no mínimo, 3 vectores de coeficientes MFCC. Dado que a janela de extracção dos MFCCs em uso tem dimensão de 128 amostras (dado que os dados foram subamostrados para 8 kHz de forma a poder ser feito um paralelo mais próximo para o caso prático estudado no ponto 4.5) e o passo de avanço desta janela é de 64 amostras, cada vogal seleccionada tem pelo menos 256 amostras de duração. Como esta selecção é feita antes da subamostragem para 8 kHz, ou seja, quando os segmentos ainda se encontram a 16 kHz, são seleccionadas as vogais com 512 amostras. No final, foram rejeitadas cerca de 4% das vogais disponíveis, e a média das vogais aceites é de 1512 amostras (quando a frequência de amostragem é 16kHz), o que equivale a cerca de 95 milissegundos.

Na base de dados TIMIT, fazendo a selecção de acordo com o descrito no ponto anterior, as vogais constituem cerca de 45% da fala. Como tal, há uma redução acentuada da quantidade de dados disponível, tanto para treino como para teste. No conjunto de testes que se segue, o tempo de treino foi fixado no número mínimo de amostras disponíveis. Por número de amostras disponíveis para treino entende-se a soma do número de amostras de cada vogal aceite contida

nos sete segmentos seleccionados para treino anteriormente. Foi adoptado este procedimento pelo facto de, ao lidar com uma quantidade reduzida de dados, pequenas variações nesta quantidade podem ser já bastante significativas no desempenho do sistema. Assim, a quantidade de dados de treino é constante para os testes que se seguem. Na prática esta fixação é feita através da selecção de uma porção da matriz MFCC obtida através da junção dos vectores MFCC retirados para cada vogal. Note-se que não é feita a concatenação das vogais, pelo que cada janela MFCC abrange apenas uma vogal, e não duas vogais concatenadas. Desta forma cada coeficiente MFCC contém apenas informação relativa a uma única vogal. São seleccionados 578 vectores MFCC para treino, e é este valor que será referido na apresentação dos resultados deste ponto em diante.

O mesmo método é aplicado no processamento das vogais de teste, e a variação é feita em intervalos de 50 vectores MFCC.

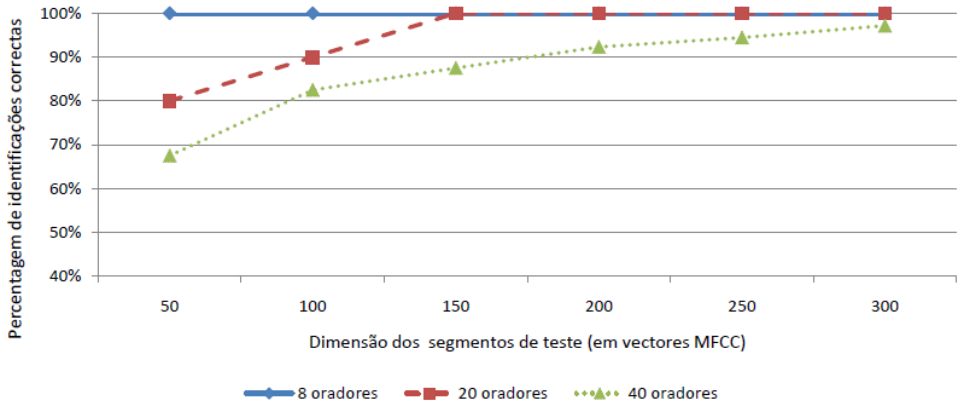

Figura 4.6: Percentagem de identificações correctas utilizando apenas vogais, para universo de oradores de diferente dimensão (8, 20 e 40 oradores) e teste de diferente duração.

Os resultados obtidos através dos procedimentos descritos anteriormente encontram-se sumariados no gráfico acima. Continuaram a ser testados diferentes números de componentes gaussianas, e foi marcado no gráfico o desempenho máximo obtido, independentemente no número de gaussianas utilizado. O número de gaussianas mais adequado em cada situação de teste será abordado mais adiante na dissertação.

Numa primeira instância podemos observar que, como seria esperado, à medida que a dimensão do teste aumenta, a *performance* do sistema melhora. Identifica-se a dimensão de 150 vectores como a dimensão a partir da qual a diminuição da quantidade de dados de teste provoca uma queda de desempenho mais acentuada do sistema. Confirma-se também que o cenário de identificação em que o universo de oradores tem dimensão de apenas 8, um sistema baseado em GMMs atinge percentagem de identificações correctas de 100% mesmo para condições de treino e teste bastante hostis em termos de duração – isto utilizando apenas as vogais. Uma análise mais aprofundada destes resultados é feita mais à frente na dissertação, aquando da comparação entre os diversos tipos de vozes utilizados.

4.3.3.2 Parte Vozeada

Considerando os fonemas que se encontram assinalados na figura 4.5 com presença de vozeamento como a parte vozeada da fala, verificou-se que esta constitui cerca de 70% da fala nos segmentos da base de dados TIMIT. De novo, foi feita a selecção descrita baseada na duração mínima do fonema, o que levou à exclusão de cerca de 9% dos fonemas disponíveis. De facto, a totalidade dos fonemas vozeados apresenta duração média inferior às vogais isoladas – cerca de 80 milissegundos, e não 95 milissegundos. Os resultados obtidos apresentam-se na figura 4.7.

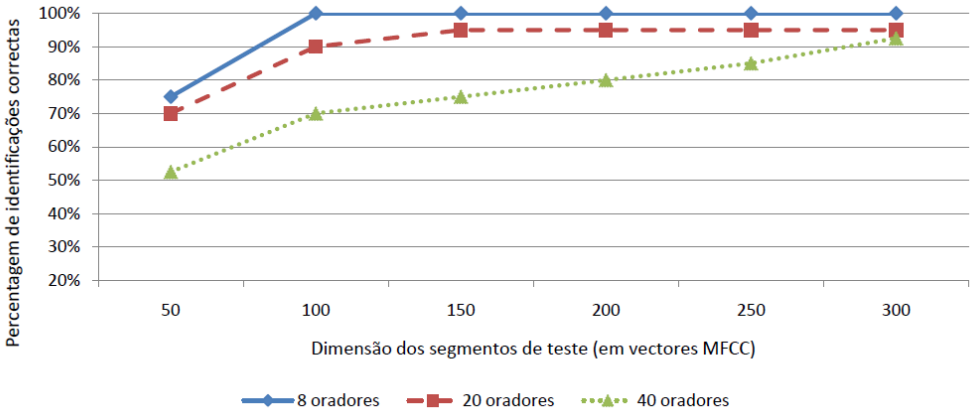

Figura 4.7: Percentagem de identificações correctas utilizando partes vozeadas, para universo de oradores de diferente dimensão (8, 20 e 40 oradores) e teste de diferente duração.

No tocante às conclusões referentes à dimensão de teste a partir da qual há uma queda mais acentuada do desempenho do sistema, verifica-se que, utilizando apenas as partes vozeadas dos segmentos originais, esse valor mantém-se entre 100 e 150 vectores MFCC. A partir destes gráficos pode verificar-se que a utilização de apenas as vogais provoca um melhor desempenho, para qualquer número de oradores considerados. A comparação directa dos dois "tipos" de voz é feita no ponto "Comparação" apresentado no final deste subcapítulo.

4.3.3.3 Parte Não Vozeada

Como mencionado anteriormente, a parte não vozeada da fala compreende os fonemas não incluídos na parte vozeada, constituindo assim a parte não vozeada cerca de 30% dos segmentos da base de dados TIMIT. Dada esta percentagem, a quantidade mínima de dados de treino (anteriormente 578 vectores MFCC) é forçosamente menor para os testes que se seguem. Foi apurado que o número mínimo de vectores disponível é 301, sendo este o valor fixado para o treino nos testes que se seguem.

Quanto à duração de teste, apenas uma pequena percentagem dos oradores considerados apresenta quantidade suficiente de dados para serem utilizadas as durações de 250 e 300 vectores MFCC. Por este motivo, a variação da duração do teste foi feita entre 50 e 200 vectores MFCC.

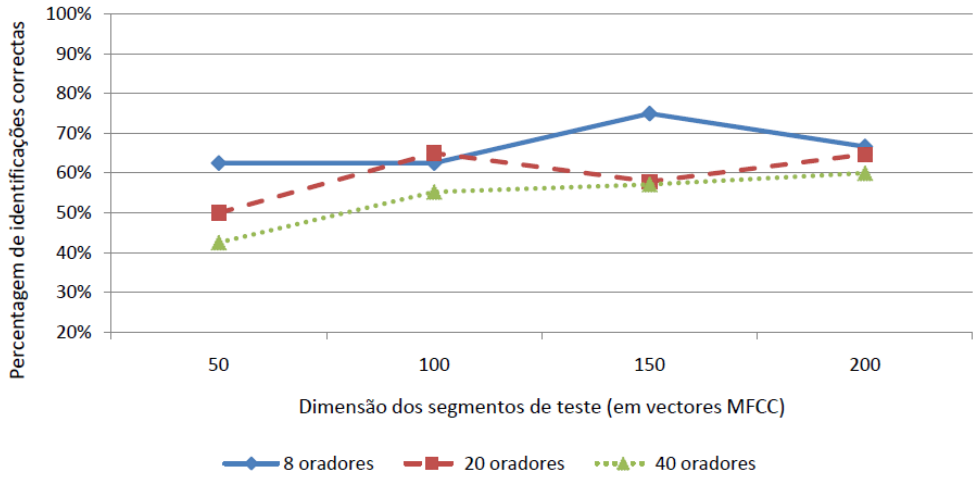

Figura 4.8: Percentagem de identificações correctas utilizando partes não vozeadas da fala, para universo de oradores de diferente dimensão (8, 20 e 40 oradores) e teste de diferente duração.

A análise comparativa do desempenho do sistema usando a parte não vozeada da voz não pode ser feita com os dados apresentados até este ponto na dissertação, pela diferença do tempo de treino considerado. A diminuição na performance observada na figura 4.8 é devida, pelo menos em parte, ao menor tempo de treino considerado nestes testes, e consequentemente não é possível inferir sobre a capacidade discriminativa das partes não vozeadas da voz. Esta análise requer a repetição dos testes com a voz completa e com as partes vozeadas da voz, considerando o mesmo tempo de teste em todas as situações avaliadas. Este estudo é apresentado seguidamente.

4.3.3.4 Comparação

Nas figuras 4.9 e 4.10 apresentam-se os gráficos com a comparação dos tipos de voz estudados, para quantidade de dados de treino diferente - 301 vectores MFCC e 578, respectivamente. Como referido anteriormente, ao utilizar as partes não vozeadas da voz não é possível extrair 578 vectores MFCC para todos os oradores considerados. Como tal, no gráfico 4.10 apenas figuram os resultados obtidos para voz completa, partes vozeadas e vogais. Ambos os gráficos são referentes aos cenários de identificação com 40 oradores.

Por observação da figura 4.9, é possível concluir numa primeira instância que o sistema baseado em vogais é o que obtém maior percentagem de identificações correctas, claramente acima dos outros tipos de voz testados. A voz completa, por outro lado, proporcionou o pior desempenho. As partes vozeadas (selecção mais abrangente do que nas vogais) e as não-vozeadas têm, nas condições de treino e teste apresentadas, desempenho idêntico. A comparação destes dois tipos de voz encontra-se dificultado pelo facto de não ser possível testar a fala não-vozeada em todos os cenários de teste presentes, devido a estarem disponíveis apenas dados suficientes para os testes com dimensão até 200 vectores MFCC.

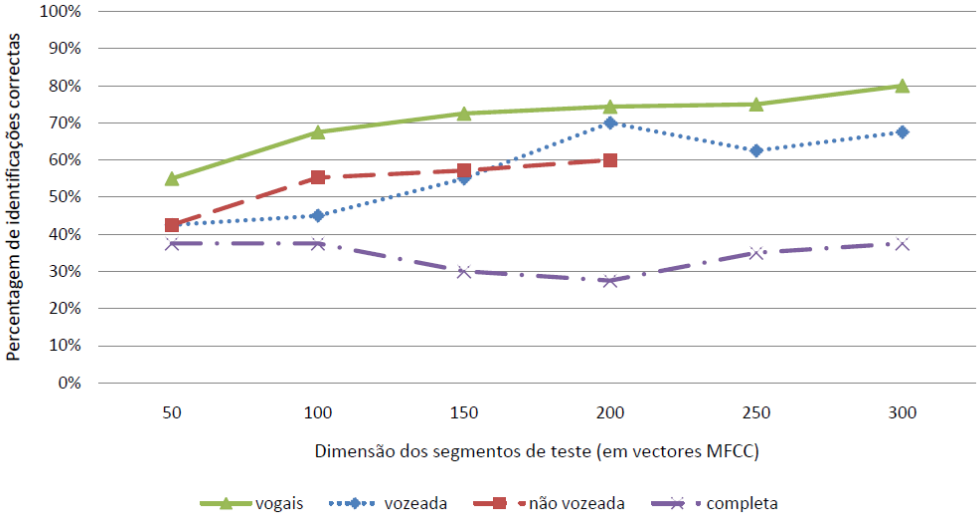

Figura 4.9: Comparação do desempenho do sistema considerando diferentes partes da voz, para 40 oradores e tempo de treino de 301 vectores MFCC.

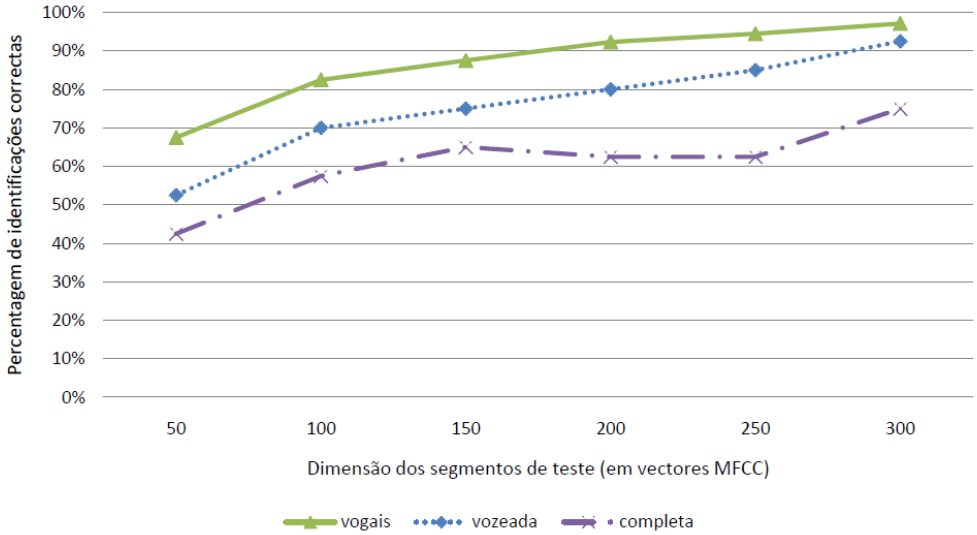

Figura 4.10: Comparação do desempenho do sistema considerando diferentes partes da voz, para 40 oradores e tempo de treino de 578 vectores MFCC.

No gráfico 4.10 o desempenho relativo entre as vogais, partes vozeadas e voz completa mantém-se: as vogais apresentam melhor desempenho, com taxas de identificação superiores às partes vozeadas da voz em mais de 10%, e entre 20% e 30% superiores comparativamente à voz completa.

A partir destes resultados pode ser concluído que as vogais captam regiões de sinal de fala com características mais estacionárias e mais facilmente conotadas com dado orador, o que beneficia a tarefa de reconhecimento. As partes vozeadas, por não serem tão estacionárias têm pior desempenho. Possivelmente o resultado mais surpreendente é o relativo às partes não vozeadas: verifica-se

que apesar de serem por natureza não estacionárias, captam também movimentos articulatórios específicos de cada orador. Por último, o sinal completo de voz, por ser mais indiferenciado, provoca maior dispersão na modelação e como tal conduz ao pior desempenho.

Estes resultados serviram também de base para extrapolação do número adequado de gaussianas consoante a quantidade de dados e o tamanho do universo de oradores. Utilizando os valores obtidos para o desempenho do uso das vogais, foi possível identificar a gama de número de componentes que alcançou os melhores resultados. Note-se que os melhores resultados obtidos, independentemente do número de gaussianas utilizado, foram os retratados em todos gráficos anteriores - para simplificação designa-se aqui resultados "máximos".

No caso em que os dados de treino foram limitados a 301 vectores MFCC, o número de componentes mais adequado oscilou maioritariamente entre as 8 e as 12 gaussianas, tanto nas situações em que se consideraram 8 oradores como 20 e 40. Já no caso em que se treinaram 578 vectores, o número de gaussianas oscilou já entre 12 e 16 na maioria dos testes. Apenas para o cenário de 8 oradores é que o número de gaussianas "óptimo" foi de 12 em todos os testes.

Se fosse necessário fixar o número de componentes, 12 seria, à semelhança do caso com 301 vectores, o valor que atinge resultados mais próximos dos resultados "máximos". Possivelmente a quantidade de dados de 578 vectores é ainda demasiado reduzida para serem consideradas 16 gaussianas, no entanto verifica-se que, como esperado, em média é necessário aumentar o número de gaussianas à medida que aumenta a quantidade de dados. Da mesma forma, também um número maior de oradores considerados necessita também de modelos mais complexos - com mais gaussianas.

4.4 Desempenho das características NRD

4.4.1 Classificador *Nearest Neighbour*, ambiente Weka

Dado que os testes preliminares efectuados com as características NRD foram realizados, como mencionado no capítulo 3.3.2, no ambiente de análise estatística Weka [44], a primeira fase de testes no trabalho presente foi também efectuada com auxílio a esta ferramenta. Foi utilizado o classificador "NNge", disponibilizado nesta, de acordo com os testes preliminares. Este método é derivado do *Nearest Neighbour* (mais detalhe em 4.5.2.1).

4.4.1.1 Vogais Cantadas

De forma a confirmar o correcto uso do algoritmo de extracção das características e da ferramenta de classificação, foram repetidos os testes preliminares assentes na base de dados de vogais cantadas, descrita em 4.2, para a de identificação de cantor. Estes testes foram apresentados em 3.3.2. A única modificação ao sistema consistiu no uso de um diferente algoritmo de extracção de MFCCs – foi utilizado o algoritmo disponível na Voicebox, enquanto que nos estudos preliminares foi usado um algoritmo adaptado pelos autores dos artigos [7] e [6].

Foram variados diversos parâmetros de forma a apurar o melhor desempenho possível para a tarefa em questão. Estes parâmetros incluíram o tamanho da janela de extracção de MFCCs e NRDs, a frequência de amostragem final após re-amostragem dos segmentos de voz e o número de coeficientes NRD utilizados. Para a obtenção deste último valor foi utilizada a ferramenta de selecção de características do Weka, que analisa as características introduzidas e devolve as características mais significativas para discriminação do cantor.

Os valores de desempenho obtidos, após optimização dos parâmetros indicados, encontram-se resumidos na tabela abaixo.

Tabela 4.3: Percentagem de identificações correctas obtidas através de características MFCC e NRD, em ambiente Weka, com segmentos de entrada pertencentes à base de dados Vogais Cantadas.

Características	% de identificações correctas
MFCCs	96,875%
NRDs	94,06%
MFCCs+NRDs	98,44%

Os resultados apresentados foram obtidos para as condições: tamanho da janela de extracção de MFCCs e NRDs de 1024 amostras, frequência de amostragem original (44100 Hz) e uso de 5 coeficientes NRD. Estes foram seleccionados a partir de 15 NRDs originais, e os que se revelaram mais significativos foram os de ordem 2 (sendo os de ordem 1 os NRDs de valor nulo, relativos à fase tomada como referência), ordem 4, 5, 9 e 11.

As percentagens de identificações correctas obtidas são bastante satisfatórias dado que, apesar do desempenho obtido com MFCCs ser muito elevado, os NRDs usados conjuntamente com os MFCCs permitem elevar a percentagem de identificações correctas em aproximadamente 2%, valor muito significativo em taxas de desempenho desta ordem.

Face a estes resultados, foi feito um estudo com o objectivo de averiguar se a melhoria de desempenho obtida se deve ao aumento do tamanho do vector de características. Foram então consideradas apenas características MFCC, e foi variado o número de coeficientes incluído. Concluiu-se que a extensão do tamanho do vector não produziu aumento na percentagem de identificações correctas, o que comprova que a melhoria do desempenho ilustrada na tabela 4.3 é devida à inclusão das características de fase NRDs.

4.4.1.2 TIMIT

Após se ter confirmado a abordagem utilizada através da base de dados de Vogais Cantadas, foi aplicada a mesma metodologia à base de dados em estudo, a TIMIT. Os segmentos seleccionados foram as vogais, resultantes da segmentação descrita em 4.3.3

Os resultados não correspondem aos resultados obtidos com a base de dados de Vogais Cantadas. O nível geral de desempenho está muito abaixo do obtido anteriormente e os NRDs não contribuíram favoravelmente à identificação de orador. Vários factores podem constituir o motivo

Tabela 4.4: Percentagem de identificações correctas obtidas através de características MFCC e NRD, em ambiente Weka, com segmentos de entrada pertencentes à base de dados TIMIT.

Características	% de identificações correctas
MFCCs	62%
NRDs	46%
MFCCs+NRDs	58%

subjacente a este comportamento. Por um lado, a voz cantada é caracterizada por possuir harmónicos melhor definidos que a voz falada, o que é essencial para a extracção dos coeficientes NRD. Por outro lado, a frequência de amostragem desta base de dados é bastante inferior à considerada anteriormente: 16000 e não 44100 Hz. Por fim, as vogais segmentadas a partir das anotações que acompanha a base de dados TIMIT não apresentam as mesmas características a nível de estabilidade da vogal que a base de dados Vogais Cantadas. De facto, breves regiões de silêncio e outras regiões que introduzem variabilidade ao sinal são incluídas frequentemente nos intervalos de tempo assinalados. Para além disso, as vogais utilizadas não são pronunciadas de uma forma constante para todos os oradores (são utilizadas várias formas de pronunciação da mesma vogal) e não são vogais sustentadas.

A base de dados que foi testada após a TIMIT apresenta algumas das características da base de dados de vogais cantadas que são desejáveis para o reconhecimento de orador, mas compreende no entanto vogais faladas e não cantadas como anteriormente. O teste do sistema com esta base de dados permite despistar algumas das razões possíveis para o fraco desempenho obtido no teste anterior. Assim, se os resultados forem positivos pode-se afirmar que a voz falada possui harmónicos com definição suficiente para serem extraídos com sucesso os NRDs e obter resultados satisfatórios a partir destes. Se, pelo contrário, isso não se verificar, elimina-se como possível causa a inclusão de partes não estáveis da vogal e os outros factores relacionados com a selecção das vogais.

4.4.1.3 Vogais Faladas

Vogais Faladas

Tabela 4.5: Percentagem de identificações correctas obtidas através de características MFCC e NRD, em ambiente Weka, com segmentos de entrada pertencentes à base de dados Vogais Faladas.

Características	% de identificações correctas
MFCCs	82%
NRDs	76%
MFCCs+NRDs	89%

Os resultados obtidos permitem confirmar que os NRDs têm capacidade de discriminação sobre o orador, e que possuem informação complementar à informação descrita pelos MFCCs, mesmo em situações de voz falada.

4.4.2 Classificador *Gaussian Mixture Model*

Após os testes preliminares efectuados em ambiente Weka, com classificador *Nearest Neighbor*, foi repetido o teste com vogais faladas utilizando o sistema de reconhecimento de orador implementado e descrito no capítulo 3. Apenas esta base de dados foi utilizada devido aos resultados prévios com a TIMIT terem apontado que a segmentação das vogais não foi feita adequadamente, tendo em conta as necessidades do sistema.

Os resultados demonstraram que o classificador GMM não é apropriado para a dimensão dos dados em causa. Mesmo na classificação baseada em MFCCs apenas, como nos testes apresentados no capítulo 4.3, a percentagem de identificações correctas é de apenas 32%. Esta diminuição drástica do desempenho deve-se à quantidade diminuta dos dados de teste e treino, incomparável à quantidade de dados considerada em testes prévios e em testes descritos na literatura. Por cada segmento de teste é possível apenas extrair 3 vectores de coeficientes MFCC, e dos segmentos de treino apenas 25. O mínimo considerado ao longo deste trabalho foi de cerca de 300 vectores para treino e 50 para teste, e mesmo estas condições produzem o desempenho insatisfatório de cerca de 55%, usando apenas vogais.

Conclui-se assim que o teste dos coeficientes NRD com um sistema de reconhecimento de orador baseado em *Gaussian Mixture Models* não é possível através das bases de dados a que houve acesso ao longo desta dissertação.

4.5 Aplicação num caso prático

Apresenta-se neste subcapítulo o trabalho realizado no âmbito do projecto de colaboração entre a Universidade do Porto e a Polícia Judiciária, o qual se intersectou com o estudo realizado ao longo da presente dissertação.

O trabalho proposto tem como objectivo estabelecer correspondência entre o orador presente numa gravação e um dos oradores presentes num conjunto de gravações disponibilizado separadamente. Todas as gravações correspondem a chamadas telefónicas, realizadas a partir de diferentes dispositivos, não havendo qualquer controlo ou informação sobre os mesmos. A identidade dos oradores presentes no conjunto de gravações mencionado é também desconhecida, sendo que parte da tarefa a realizar prevê o estabelecimento de correspondências entre os oradores das diferentes gravações, pois vários oradores estão presentes em mais do que uma chamada.

O sistema de reconhecimento de orador baseado em características MFCC e modelação GMM descrito nos capítulos anteriores foi aplicado a este caso de aplicação, assim como outros métodos de classificação baseados na ferramenta já descrita anteriormente, Weka. Os resultados serão apresentados e analisados neste subcapítulo.

4.5.1 Base de dados e pré-processamento

As gravações telefónicas disponibilizadas são constituídas originalmente por 23 ficheiros áudio, correspondentes a 23 conversas telefónicas. Como trabalho preparatório foram separados

manualmente os intervalos de tempo em que falam os dois oradores distintos presentes em cada gravação. Todos os segmentos resultantes foram concatenados novamente de forma a resultarem dois segmentos por conversa telefónica: segmento do orador A e segmento do orador B.

Como referido anteriormente, pretende-se identificar o orador de uma conversa telefónica específica. A gravação correspondente foi dividida, à semelhança das restantes, em dois segmentos para cada orador. Dado que a identidade de um dos oradores é conhecida e não é relevante para o trabalho, esse segmento foi excluído. O outro segmento, cujo orador se pretende classificar, será denominado neste documento como segmento A, de forma a facilitar a exposição dos métodos e dos resultados obtidos.

Exceptuando o segmento A, encontram-se disponíveis no total 44 segmentos. Foram excluídos 3 destes segmentos pelo facto de se ter apurado, por audição dos mesmos, que correspondem a crianças e oradores femininos (que não correspondem à partida ao orador do segmento A). Foram também retirados os silêncios em todos os segmentos. Utilizou-se um algoritmo Matlab para excluir os intervalos de forma automática, excepto para o segmento A, em que a remoção foi manual. Isto deve-se ao facto de este segmento ter sido mais afectado por ruído que a maioria dos outros segmentos, e pelo facto de ser necessário preservar a maior quantidade de sinal possível, visto que se trata do orador que se pretende classificar.

Excluindo os silêncios, os segmentos têm durações compreendidas entre cerca de 4 segundos e 177 segundos, sendo a média 43 segundos. O segmento A tem duração de 17 segundos. A frequência de amostragem original dos ficheiros é 8 kHz. No entanto, de forma a os segmentos poderem ser utilizados noutros estudos realizados no âmbito do projecto (em que foi feita a extracção de parâmetros acústicos como *jitter* e *Harmonics-to-Noise Ratio*), foi feito um *upsample* para 22050 Hz.

É de salientar que as gravações disponíveis apresentam condições bastante adversas do ponto de vista de reconhecimento automático de orador. A nível acústico as chamadas são afectadas por bastante ruído, artefactos e distorção introduzidos pela transmissão, codificação GSM, entre outros factores. Os ruídos com maior impacto no sinal foram eliminados em pré-processamento das gravações, no entanto permanecem ao longo da maioria dos segmentos várias fontes de perturbação. Por outro lado, o discurso presente apresenta características que dificultam a tarefa em questão: em alguns segmentos as vozes variam entre sussurros e elevação exagerada da voz, são feitas várias interrupções que quebram a cadência normal da fala, há variação do estado emocional dos oradores, entre outros. Em suma, estes factores formam um cenário de identificação bastante diferente dos cenários geralmente considerados nas bases de dados comummente utilizadas.

O pré-processamento dos dados inclui também a segmentação manual das vogais. Esta segmentação foi necessária para a extracção dos parâmetros acústicos no âmbito do estudo paralelo realizado, e também foi motivada pelos resultados obtidos no estudo com a base de dados TIMIT apresentado no ponto 4.3. Visto que o sistema GMM+MFCC obteve resultados claramente superiores, em termos de percentagem de identificações correctas, quando se utilizaram apenas as vogais, foi adoptada esta abordagem no projecto descrito.

Por fim refere-se que foi realizada previamente a correspondência entre os oradores dos segmentos disponíveis. Esta correspondência foi obtida através do consenso entre as correspondências feitas subjectivamente, através da audição dos segmentos, por diferentes participantes. Esta classificação é considerada, como tal, o *ground truth* - assume-se como sendo a classificação "verdadeira". De acordo com esta, nos 41 segmentos finais (excluindo o segmento A) estão presentes 12 oradores diferentes.

4.5.2 Métodos Weka

Neste ponto são descritos os métodos de classificação utilizados baseados no ambiente Weka, assim como os resultados obtidos.

A primeira fase de testes tem por objectivo estabelecer correspondência entre os oradores dos segmentos disponibilizados - excepto o segmento A, como explicitado acima. Esta correspondência é seguidamente confrontada com a correspondência feita através da classificação subjectiva dos segmentos, o *ground truth*. Desta forma, ao confrontar a classificação obtida pelo método automático de reconhecimento com a classificação assumida como verdadeira, é possível avaliar o desempenho do método utilizado. Após ser conhecida a eficiência dos algoritmos realiza-se o segundo teste, de forma a realizar a tarefa de classificação proposta - classificação do segmento A.

De modo a aferir quais os segmentos que pertencem ao mesmo orador, nos testes realizados foi feita a comparação directa entre pares de segmentos. Entende-se por segmento o conjunto das vogais retiradas de cada registo telefónico, para cada orador. Foi também utilizado em cada teste um modelo representativo dos outros oradores. Esta abordagem integra assim o conceito de *background model* (BM), descrito em 2.4.2. Como referido na análise do estado da arte, a selecção destes oradores alternativos que constituem o *background model* não se encontra bem definida na literatura pois é por vezes difícil definir o espaço de oradores alternativos. Neste trabalho esta tarefa encontra-se facilitada, no entanto, pelo facto de possuirmos um conjunto fechado de oradores considerados para teste e treino (os 12 oradores únicos, identificados na avaliação perceptiva que conduziu ao *ground truth*).

Foi adoptada a abordagem de construção de um *background model* específico para cada teste realizado. Assim, a definição deste foi feita da seguinte forma:

- Tendo dois segmentos de teste, foram identificados os seus oradores correspondentes através da classificação subjectiva mencionada anteriormente.
- Recorrendo novamente às classificações subjectivas, consideradas como *ground truth*, foram eliminados todos os segmentos pertencentes aos oradores identificados no passo anterior.
- Os segmentos que não foram eliminados na fase anterior são os que constituem *background model* para o teste em questão.

Quanto à quantidade de dados incluída para cada orador, esta não foi, como seria o caso ideal, a mesma. Isto deve-se à grande disparidade entre a quantidade disponível entre cada orador: alguns

oradores estão presentes em apenas uma gravação, enquanto outros estão presentes em mais de dez gravações. Considerar a mínima quantidade de dados disponível para qualquer orador resultaria num *background model* demasiado rarefeito, e por isto não foi feita equalização da quantidade de dados.

De cada segmento incluído no *background model* foi retirada a mesma quantidade de vectores MFCCs – 10. Este valor foi escolhido pelo facto de se ter verificado em testes preliminares que proporciona o melhor desempenho do algoritmo de classificação. Pretende-se que o BM inclua dados suficientes para haver correcta modelação deste, mas que simultaneamente não apresente uma quantidade de dados demasiado desproporcional em relação aos dados de teste.

Para a tarefa de reconhecimento em questão, o desempenho da classificação encontra-se dependente do nível de separação dos dados (características MFCC) entre dois oradores diferentes, assim como da separação destes em relação ao *background model*. De forma a eliminar zonas de sobreposição entre os *clusters* dos oradores e do BM, este último foi delimitado à área de maior concentração dos dados, eliminado assim *outliers* que possam dificultar a identificação clara dos *clusters* individuais. Assim, cada coeficiente MFCC retirado dos segmentos que constituem o BM foi reduzido a:

$$x > \mu - \sigma \wedge x < \mu + \sigma \tag{4.1}$$

sendo μ a média de cada coeficiente e σ o desvio padrão.

4.5.2.1 *Nearest Neighbour*

O primeiro método de classificação aplicado encontra-se disponível no Weka sob a designação "NNge". Trata-se de uma variante do método *Nearest Neighbour* denominada *Nearest Neighbour generalised*. Neste algoritmo são construídos "hiperrectângulos" que delimitam cada classe presente no conjunto de dados. Estes hiperrectângulos podem ser vistos como um conjunto de regras para a classificação de cada ponto [52]. Um exemplo destas regras pode ser observado na tabela abaixo:

Tabela 4.6: Exemplo de regras que definem um hiperrectângulo.

Classe A se	característica1 = 1 ou 2 e característica2 > 20 e característica2 $\leq$ 42 e característica3 = 12,5

Para cada classe são especificadas regras que definem intersecção de espaços. Existem tantos espaços quanto características.

Em [52] é mostrado que este método tem melhor desempenho que o *Nearest Neighbor* original em diversas situações. É nomeadamente mais robusto em casos em que existem características

irrelevantes no conjunto de dados fornecido. Estudos anteriores realizados pelos autores do presente relatório levam a crer que o NNge é dos algoritmos com melhor desempenho disponíveis no *software* Weka.

MESMO ORADOR

Oradores:
a. Orador6seg1
b. Orador6seg2

a	b	c	classif.
62%	**4%**	35%	**a**
2%	**47%**	51%	**b**
11%	6%	83%	**c**

Oradores:
a. Orador15seg2
b. Orador15seg3

a	b	c	classif.
45%	7%	47%	**a**
4%	53%	44%	**b**
3%	5%	92%	**c**

Oradores:
a. Orador6seg3
b. Orador6seg4

a	b	c	classif.
64%	2%	35%	**a**
5%	35%	60%	**b**
10%	5%	85%	**c**

ORADORES DIFERENTES

Oradores:
a. Orador6seg1
b. Orador15seg1

a	b	c	classif.
69%	5%	25%	**a**
11%	51%	38%	**b**
15%	19%	65%	**c**

Oradores:
a. Orador6seg1
b. Orador15seg4

a	b	c	classif.
60%	11%	29%	**a**
11%	60%	29%	**b**
13%	5%	82%	**c**

Oradores:
a. Orador6seg1
b. Orador1seg1

a	b	c	classif.
44%	4%	53%	**a**
4%	75%	22%	**b**
9%	4%	88%	**c**

Figura 4.11: Matrizes de confusão obtidas para os testes realizados com o algoritmo NNge.

Apresenta-se na figura 4.11 algumas das matrizes de confusão devolvidas pelo *software* Weka resultantes dos testes efectuados. Na coluna da esquerda apresentam-se testes realizados entre segmentos de voz do mesmo orador (aferido através de avaliação perceptiva), enquanto que na coluna da direita figuram testes entre segmentos de voz de oradores diferentes. Os segmentos testados e o respectivo orador são listados no início de cada tabela, e indicados como a e b na matriz de confusão. A referência c corresponde ao *background model* em todos os quadros apresentados.

Em cada quadro as linhas correspondem a um teste. Tomando como exemplo o primeiro quadro da coluna da esquerda, a primeira linha pode ser interpretada como: 62% das amostras do segmento a foram classificadas como o orador do segmento a, 4% foram classificadas como

pertencentes ao orador do segmento *b* e 35% foram classificadas como pertencentes ao *background model*.

Num teste entre dois segmentos do mesmo orador era esperado observar um equilíbrio entre as percentagens atribuídas ao *cluster a* e ao *cluster b*, pelo facto das características MFCC formarem, idealmente, dois *clusters* sobrepostos. Desta forma o classificador deveria atribuir percentagens idênticas aos oradores dos dois segmentos (que correspondem na realidade ao mesmo). Estas percentagens encontram-se assinaladas a fundo cinzento e negrito na primeira tabela da figura 4.11. Já no caso de teste entre dois segmentos diferentes o oposto deve ser verificado – a percentagem do segmento atribuída à classe *a* deve ser próxima de 100% e a percentagem atribuída à classe *b* próxima de 0%.

Por observação das tabelas incluídas na figura 4.11 observa-se que os cenários esperados, descritos acima, não se verificam. Numa primeira instância pode verificar-se que as tabelas da coluna da esquerda e da direita apresentam configurações bastante semelhantes – o que denuncia à partida a incapacidade do método de distinguir as situações em que os segmentos pertencem ao mesmo orador e em que pertencem a oradores diferentes. De facto, todas as tabelas indicam que o classificador faz a distinção entre dois segmentos diferentes, mesmo quando estes pertencem ao mesmo orador.

Pode observar-se, por outro lado, que o classificador não consegue isolar devidamente o *background model*, classificando uma grande percentagem das amostras dos segmentos como *background model*. Tendo em conta que este é sempre construído com segmentos de oradores diferentes dos que se encontram em teste, estes resultados revelam ineficiência do classificador ou até das próprias características MFCC na tarefa de reconhecimento em causa.

4.5.2.2 *Support Vector Machines*

Pelo facto de os resultados obtidos através do algoritmo NNge indicarem que este método é incapaz, nas condições indicadas, de fazer a classificação das vozes com um nível razoável de precisão, foram repetidos os testes com outro algoritmo disponível no Weka. Devido à grande utilização de *Support Vector Machines* na área de reconhecimento de orador, foi utilizado o algoritmo "SMO" do Weka. Este treina uma SVM através do método *Sequential Minimal Optimization* (SMO), descrito em [53].

Procedeu-se de forma semelhante à descrita no ponto anterior para obter as matrizes de confusão.

Por observação da figura 4.12 conclui-se que o método SMO utilizado não permitiu obter uma distinção mais demarcada entre a situação de teste entre dois oradores iguais e dois oradores diferentes. Para além disto, neste conjunto de resultados verifica-se que o algoritmo obteve uma pior separação dos *clusters* dos oradores e do *background model*. O que se pode concluir através da análise das matrizes de confusão apresentadas é que o SMO disponível no Weka é um classificador menos eficaz que o NNge.

MESMO ORADOR

Oradores:
a. Orador6seg1
b. Orador6seg2

a	b	c	classif.
0%	0%	100%	**a**
0%	29%	71%	**b**
0%	6%	94%	**c**

Oradores:
a. Orador15seg2
b. Orador15seg3

a	b	c	classif.
35%	4%	62%	**a**
5%	45%	49%	**b**
2%	3%	95%	**c**

Oradores:
a. Orador6seg3
b. Orador6seg4

a	b	c	classif.
0%	0%	100%	**a**
0%	0%	100%	**b**
0%	0%	100%	**c**

ORADORES DIFERENTES

Oradores:
a. Orador6seg1
b. Orador15seg1

a	b	c	classif.
16%	5%	78%	**a**
2%	27%	71%	**b**
2%	14%	84%	**c**

Oradores:
a. Orador6seg1
b. Orador15seg4

a	b	c	classif.
13%	9%	78%	**a**
7%	42%	51%	**b**
3%	11%	86%	**c**

Oradores:
a. Orador6seg1
b. Orador1seg1

a	b	c	classif.
0%	2%	98%	**a**
4%	60%	36%	**b**
0%	2%	98%	**c**

Figura 4.12: Matrizes de confusão obtidas para os testes realizados com o algoritmo SMO (SVMs).

4.5.2.3 Análise e Conclusões

Dada a ineficiência de ambos os métodos testados, não foi realizada a classificação do segmento A, como era o objectivo da tarefa proposta.

Visto que foram utilizados dois métodos que são descritos frequentemente na literatura como tendo níveis de precisão elevados, foram analisadas as características extraídas, com o intuito de compreender os resultados obtidos.

Na figura 4.13 mostra-se a projecção das matrizes MFCC para dois segmentos pertencentes ao mesmo orador. As projecções de qualquer par de características podem ser visualizadas no Weka, através do menu "Visualize". Esta ferramenta permitiu observar a localização dos *clusters* dos diferentes segmentos e do BM nas diferentes projecções.

Na grande maioria das projecções a distribuição dos dados apresenta a configuração demonstrada na figura 4.13: as nuvens de pontos relativas aos dois segmentos em teste encontram-se

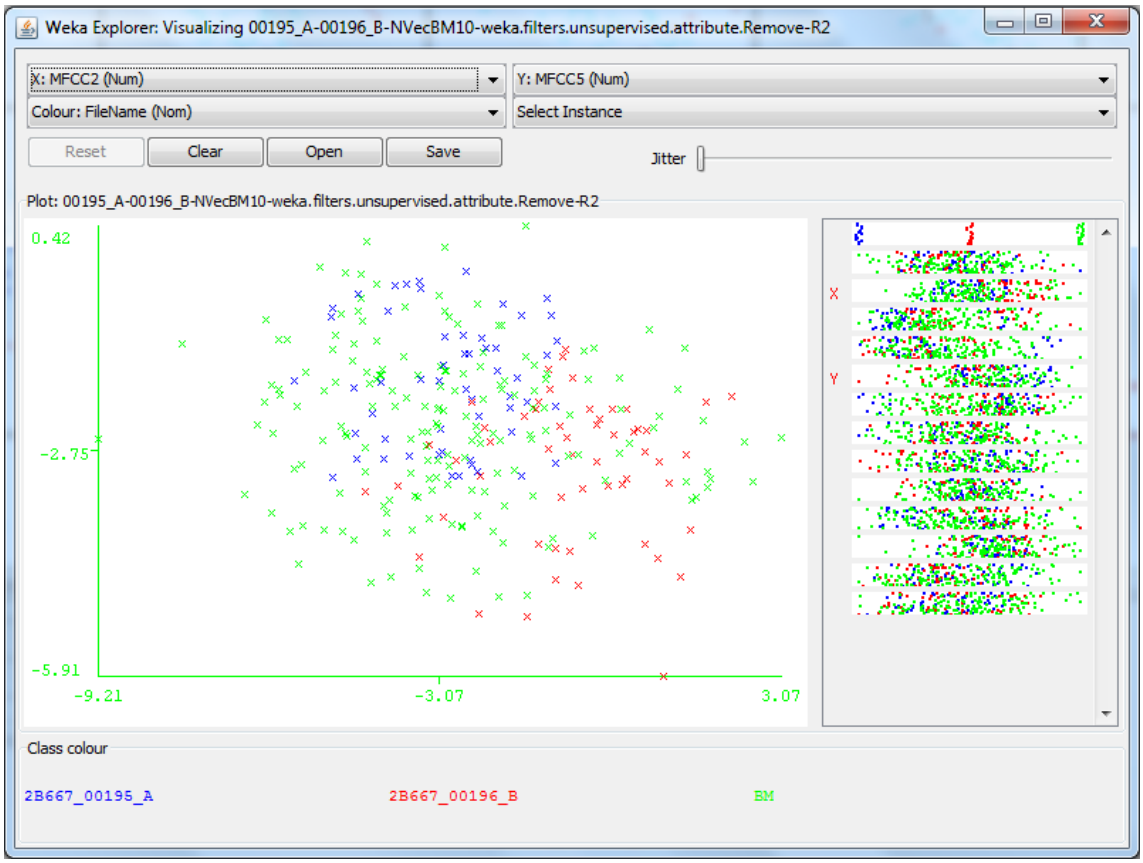

Figura 4.13: Visualização da projecção dos coeficientes MFCC de dois segmentos do mesmo orador. A projecção mostra os coeficientes de ordem 2 e de ordem 5.

bastante separadas (nuvens vermelha e azul), enquanto que o BM (pontos a verde) encontra-se sobreposto com os outros dois *clusters*. Esta situação é bastante problemática dado que os dois segmentos pertencem ao mesmo orador. Estes resultados indicam que as características MFCC extraídas contêm informação relativa ao próprio segmento: ao ruído que afecta a chamada, ao canal utilizado e até ao próprio discurso e atitude fonatória contidos no segmento (dado que os MFCCs são também bastante eficazes na tarefa de reconhecimento de fala). Os resultados apontam assim para a conclusão que estas fontes de perturbação mascaram a informação relativa ao orador, o que dificulta a tarefa do classificador de associar dois segmentos diferentes ao mesmo orador.

4.5.3 Método GMM

O sistema baseado em características MFCC e modelação GMM descrito em 3.4 e cujos resultados são apresentados em 4.3 foi aplicado ao presente caso de estudo. A escolha entre as diferentes variantes e parametrizações de funcionamento foi baseada nos resultados retirados através de testes com a base de dados TIMIT, de forma a obter a maior precisão de identificação possível.

A primeira fase de testes tem por objectivo, como referido para os testes em ambiente Weka, estabelecer correspondências entre os oradores dos segmentos e aferir, a partir do *ground truth* apurado anteriormente, o nível de desempenho do sistema. Com este fim foram seleccionados alguns segmentos para teste contra os modelos dos 12 oradores únicos. Cada teste é configurado da seguinte forma: é identificado o orador pertencente ao segmento de teste; na construção do modelo desse mesmo orador são extraídos MFCCs de todos os segmentos disponíveis para o mesmo orador (considerando as correspondências decorrentes do *ground truth*) excepto o próprio segmento de teste; por fim todos os outros modelos considerados no teste são treinados com todos os dados disponíveis para os respectivos oradores.

São retirados novamente 55 vectores MFCC de cada segmento de teste. Para a construção de cada modelo GMM são também considerados apenas 55 vectores. Isto deve-se ao facto de alguns modelos terem de ser treinados apenas com um segmento, e para correcta classificação todos os modelos devem ser treinados com a mesma quantidade de dados. Com base nesta quantidade de dados utilizada para construção de cada modelo e nas conclusões retiradas pelos estudos estudos anteriores, com a base de dados TIMIT, foi estabelecido o número de gaussianas de 8.

4.5.3.1 Resultados

Foram feitos 17 testes, com 17 segmentos diferentes. A percentagem de identificações correctas obtida nestes testes foi de apenas 12%. Para além disto verifica-se que o classificador atribui todos os segmentos de testes apenas às classes de dois oradores presentes (orador 3 e 4). Há assim um claro enviesamento do classificador em relação a estes dois oradores. O motivo para esta tendência não é claro, visto que os modelos são treinados com a mesma quantidade de dados. Por outro lado, são usados vários segmentos para construir cada modelo GMM (quando existe mais do que um segmento disponível), o que atenua efeitos de ruído ou artefactos pontuais a um segmento que possam enviesar os resultados.

Apesar de estes resultados revelarem a inadequação do método de classificação nas circunstâncias presentes neste trabalho, foi feita a classificação do segmento A, como era o objectivo final deste estudo. Este teste confirmou, no entanto, o comportamento tendencioso do classificador, pois estabeleceu novamente correspondência entre o segmento A e o orador 4.

4.5.3.2 Conclusão

Com base dos resultados apresentados é possível concluir que o método de modelação e classificação baseado em MFCCs e GMMs não é adequado ao tipo de registos telefónicos fornecidos. Por um lado, a quantidade dos dados extraídos dos registos é demasiado escassa para uma abordagem deste género (abordagem estatística): os modelos de alguns oradores tiveram de ser treinados apenas com um segmento, e a quantidade de vogais que é possível extrair nas condições acústicas presentes é por vezes igualmente escassa.

Por outro lado, os resultados indicam que os coeficientes MFCC extraídos não traduzem devidamente as características individuais dos oradores, não permitindo ao classificador isolar cada

um destes. Isto deve-se ao facto de os segmentos de sinal de voz utilizados apresentarem características pouco favoráveis, devido à incidência de ruído e alterações de sinal devido às fases de captação, codificação GSM e transmissão de sinal. Estas alterações, nomeadamente as que afectam o segmento que se pretende classificar (segmento A), são mais severas do que as que são tipicamente consideradas em estudos da área (incluindo o estudo feito com a base de dados TIMIT).

4.6 Conclusões

Neste capítulo descreveram-se os testes realizados à solução de reconhecimento de orador implementada, apresentada no capítulo 3. Numa primeira instância o estudo incidiu sobre o desempenho do sistema base, constituído pelos métodos mais destacados do estado da arte, GMMs e MFCCs.

Vários objectivos motivaram este primeiro estudo: por um lado pretendia-se aferir que níveis de precisão era possível atingir com o sistema implementado, de forma a validar os métodos utilizados. Por outro, de forma a maximizar o desempenho, foram estudados simultaneamente vários parâmetros de funcionamento dos algoritmos, com o intuito de aferir que parametrização é mais adequada em função das condições de teste presentes. O principal parâmetro em estudo foi o número de componentes gaussianas, e foi variado em função da quantidade de dados de treino utilizada, da quantidade de dados de teste e dimensão do universo de oradores considerados na tarefa de identificação. Por último, o teste do sistema *baseline* GMM+MFCC permitiu caracterizar o desempenho (em termos de percentagem de identificações correctas) em função do tempo de teste, o que permitiu também identificar que durações de teste constituem um desafio para o sistema. Neste aspecto foi visto que a duração de 6 segundos de teste permite já ordem de desempenho próxima de 100%, e que a duração de 2 segundos provoca já uma diminuição significativa: entre redução 4% a 7% para a situação mais complexa considerada, 40 oradores.

Seguidamente foi estudado o efeito da segmentação da voz consoante a presença de vozeamento, motivado pela necessidade de segmentação das partes vozeadas para estudo das características também em estudo neste trabalho, NRDs, e pelo facto de as características específicas das partes vozeada da voz poderem trazer vantagens ao reconhecimento de orador nos cenários em estudo.

Foi concluído, no tocante à segmentação, que o uso de apenas vogais permite desempenho até 30% superior comparativamente ao atingido através da voz completa. Esta última produz, de facto, o desempenho mais baixo de entre os diferentes tipos de voz analisados: voz completa, vozeada, não vozeada e vogais. Conclui-se portanto que a inclusão de toda a variedade de fonemas existente provoca a construção de modelos de oradores menos capazes de traduzir as características únicas de cada orador.

Por fim, no ponto 4.5 foi descrito um trabalho que decorreu paralelamente a esta dissertação, no qual foi proposta uma tarefa de reconhecimento de orador. Devido às gravações fornecidas, apenas o sistema baseado em MFCCs foi utilizado no caso em questão. No tocante a métodos

de modelação e classificação, foram utilizados: *Gaussian Mixture Models* (através do sistema implementado ao longo da dissertação), *Nearest Neighour* e *Support Vector Machines*, ambos os últimos através dos algoritmos disponibilizados no Weka. Foi concluído que nenhum dos sistemas permitiu estabelecer correspondência entre os oradores das gravações disponibilizadas. O fraco desempenho obtido para todos estes métodos é devido em grande parte às condições acústicas que afectam os segmentos fornecidos. Para além disto, após segmentação das vogais, a quantidade de dados que foi possível utilizar é bastante reduzida, o que indica que métodos como GMMs não são indicados para reconhecimento de orador neste caso de estudo. No entanto, os resultados apontam que os próprios coeficientes MFCC retirados, nos quais assentam todos os modelos de modelação e classificação, traduzem mais facilmente o ruído que afecta cada gravação, o canal utilizado e factores relacionados com o discurso e atitude fonatória presentes nos segmentos do que as características específicas de orador.

Capítulo 5

Conclusão

O trabalho realizado ao longo da presente dissertação teve por objectivo o estudo da tarefa de identificação de orador num cenário independente de texto, *close-set* e com testes de duração reduzida (como valor de referência, dois segundos). Para além do aumento da robustez através da utilização de novas técnicas, pretendia-se também obter metodologias aplicáveis a um caso real de reconhecimento de orador que foi desenvolvido simultaneamente com esta dissertação.

A implementação do sistema base e subsequentes testes teve por objectivo inicial estudar a precisão, a nível de percentagem de identificações correctas, que é atingida com um sistema baseado nos algoritmos mais utilizados no estado da arte (características MFCC e modelação GMM). Após a aferição do desempenho atingido, foram estudados novos métodos a integrar nos sistemas com capacidade de melhorar a precisão da identificação de orador. Este estudo foi concentrado nos dados de entrada do sistema e no módulo de extracção de características, e concentrou-se em duas técnicas principais: segmentação das gravações consoante a presença de vozeamento e introdução de novas características de voz baseadas na fase, *Normalized Relative Delays*.

Relativamente à segmentação de acordo com presença de vozeamento, foi possível mapear os diferentes tipos de voz em termos de precisão relativa: os segmentos que contêm vogais permitiram maior percentagem de identificações correctas, seguidos pelas partes vozeadas (que contêm as vogais e consoantes vozeadas) e pelas partes não vozeadas, encontrando-se por último a voz completa. O ganho em termos de precisão obtido ao utilizar apenas as vogais é superior a utilizar a voz vozeada até 10%, e 30% relativamente à voz completa.

Este estudo salienta a importância de uma segmentação rigorosa das gravações: verifica-se uma diferença muito significativa entre o uso de vogais e de partes vozeadas da voz. O estudo seguinte, relativo aos NRDs, veio comprovar de outra forma esta conclusão. Foi visto que a melhoria do desempenho com junção de NRDs aos vectores de MFCCs verificou-se apenas no uso de uma base de dados de vogais faladas, e não na TIMIT. Por outro lado, o desempenho obtido com a TIMIT (no classificador NNge do Weka) mesmo utilizando apenas MFCCs, foi bastante menor que o obtido para a base de dados de vogais faladas estudada: verificou-se uma

diferença de percentagem de identificações correctas entre 20% a 30%. A principal diferença entre os segmentos utilizados das duas bases de dados prende-se na segmentação efectuada: enquanto que nas Vogais Faladas as vogais incluídas compreendem apenas zonas perfeitamente estáveis do sinal, as vogais segmentadas da TIMIT, a partir das anotações que a acompanham, incluem frequentemente zonas de silêncio e de transição entre fonemas.

Conclui-se que, conjuntamente com uma segmentação rigorosa das vogais, os *Normalized Relative Delays* têm potencial para melhorar significativamente o desempenho dos actuais sistemas de reconhecimento de orador, e o seu uso pode ser de especial interesse nos casos de identificação com escassez de dados - casos em que as características tradicionais, MFCCs, não são suficientes para a identificação robusta de orador.

Quanto ao caso prático de identificação de orador, a incapacidade dos métodos utilizados na classificação dos oradores dos segmentos fornecidos vem apontar algumas das limitações dos métodos actuais. De facto, apesar de haver um crescente interesse em soluções robustas a ruído e condições acústicas desfavoráveis, uma grande maioria dos sistemas de reconhecimento em estudo ainda tem por aplicação um ambiente bastante controlado de gravação de vozes. Por outro lado, uma grande parte dos sistemas é inserido em ambiente em que há cooperação entre os oradores e o sistema, e como tal a quantidade de dados disponível não impõe uma restrição forte. Numa aplicação forense como a apresentada são, no entanto, comuns os cenários em que existe escassez de dados disponíveis e condições acústicas para as quais os sistemas de reconhecimento de orador não estão presentemente preparados.

5.1 Trabalho Futuro

Dado os resultados promissores obtidos com a extracção de NRDs, em combinação com MFCCs, do sistema de reconhecimento aplicado à base de dados Vogais Faladas, seria importante futuramente aprofundar o estudo destas características. Numa primeira fase poderia ser feita a segmentação adequada das vogais presentes na base de dados TIMIT, ou outras bases de dados extensas - desta forma seria possível obter uma quantidade de dados suficiente para aplicar o sistema de reconhecimento baseado em GMMs e rever os resultados obtidos anteriormente.

Por outro lado, de forma a alargar as aplicações de um sistema baseado na extracção de NRDs, que necessita de vogais bem segmentadas para obter melhorias de desempenho, seria importante estudar uma solução de segmentação automática da voz. Alguns reconhecedores automáticos de presença de vozeamento são já utilizados, como tal iniciaria-se o estudo com a determinação do desempenho atingido com os segmentos obtidos de forma automática, e posteriormente podem ser estudados melhoramentos aos algoritmos.

O estudo mais aprofundado do sistema desenvolvido passaria de igual forma pelo alargamento da base de dados a oradores femininos, visto que não foram incluídos neste estudo. Por último, o teste em diferentes base de dados para além da TIMIT, em especial base de dados com condições menos favoráveis como a NTIMIT e outras comummente utilizadas, permitia determinar se haveria melhoramento de desempenho através dos NRDs. Dado que na aplicação forense considerada

a largura de banda das gravações disponibilizadas não permitiu o uso de NRDs, seria importante verificar futuramente se estas características de fase trazem valor acrescentado a estes cenários de reconhecimento.

Referências

[1] Biometria - Impressão vocal, disponível em: http://www.gta.ufrj.br/grad/09_1/versao-final/impvocal/propdosinal.html, Junho 2011.

[2] E. Karpov. Real-time speaker identification. Tese de Mestrado, University of Joensuu, 2003.

[3] F. Zheng, G. Zhang, e Z. Song. Comparison of different implementations of MFCC. *Journal of Computer Science & Technology*, páginas 582–589, Setembro 2001.

[4] C. Campbell e Y. Ying. *Learning with Support Vector Machines*. Morgan ClayPool, segunda edição, 2011.

[5] A. Larcher, C. Lévy, D. Matrouf, e J.-F. Bonastre. LIA NIST-SRE'10 systems. *Proceedings of the NIST-SRE Workshop, Brno, República Checa*, Junho 2010.

[6] R. Sousa e A. Ferreira. Importance of the relative delay of glottal source harmonics. *39th AES Conference*, 2010.

[7] R. Sousa e A. Ferreira. Singing voice analysis using relative harmonic delays. *12th Annual Conference of the International Speech Communication Association*, 2011.

[8] J. P. Campbell, W. Shen, W. M. Campbell, R. Schwartz, J.-F. Bonastre, e D. Matrouf. Forensic speaker recognition: A need for caution. *IEEE Signal Processing Magazine*, páginas 95–103, Março 2009.

[9] A. Kanagasundaram, R. Vogt, D. Dean, S. Sridharan, e M. Mason. i-vector based speaker recognition on short utterances. *Interspeech 2011*, páginas 2340–2344, 2011.

[10] H. Li e B. Ma. Best of the Web. *IEEE Signal Processing Magazine*, páginas 139–142, Novembro 2010.

[11] J. P. Campbell. Speaker recognition: A tutorial. *Proceedings of the IEEE*, páginas 1437–1462, Setembro 1997.

[12] J. Markowitz. Voice authentication in the real world. *National Center for Biometric Studies Conference, Voice Authentication for Identity Management*, 2006.

[13] D. E. Sturim, W. M. Campbell, e D. A. Reynolds. Speaker classification I. chapter Classification Methods for Speaker Recognition, páginas 278–297. Springer-Verlag Berlin Heidelberg, 2007.

[14] M. A. Hossan, S. Memon, e M. A. Gregory. A novel approach for MFCC feature extraction. *Signal Processing and Communication Systems (ICSPCS)*, páginas 1–5, 2010.

[15] T. Kinnunen, P. Fränti, e E. Karpov. Real-time speaker identification. *IEEE Transactions on Audio, Speech and Language Processing, 14(1)*, páginas 277–288, Janeiro 2006.

[16] D. Hosseinzadeh e S. Krishnan. Combining vocal source e MFCC features for enhanced speaker recognition performance using GMMs. *Proceedings of the 9th IEEE International Workshop on Multimedia Signal Processing*, Outubro 2007.

[17] N. Kleynhans e E. Barnard. A channel normalization technique for speech recognition in mismatched conditions. *Proceedings of the 19th Annual Symposium of the Pattern Recognition Association of South Africa*, 2008.

[18] A. A. Garcia e R. J. Mammone. Channel-robust speaker identification using modified-mean cepstral mean normalization with frequency warping. *IEEE International Conference on Acoustics, Speech, and Signal Processing, 1999 (ICASSP '99)*, páginas 325–328, Março 1999.

[19] D. Chow e W. H. Abdulla. Robust speaker identification based on perceptual log area ratio and Gaussian mixture models. *INTERSPEECH 2004 – ICSLP*, 2004.

[20] H. Gish e H. Schmidt. Text independent speaker identification. *IEEE Signal Processing Magazine*, páginas 18–32, Outubro 1994.

[21] K. P. Markov e S. Nakagawa. Comparison between LPC cepstrum and MFCC for speaker recognition using clean and telephone speech. *Toyohashi University of Technology*, 1999.

[22] H. Ezzaidi, J. Rouat, e D. O'Shaughnessy. Towards combining pitch and MFCC for speaker recognition systems. *EUROSPEECH 2001*, páginas 2825–2828, 2001.

[23] M. D. Plumpe, T. F. Quatieri, e D. A. Reynolds. Modeling of the glottal flow derivative waveform with application to speaker identification. *IEEE Transactions on Speech and Audio Processing, vol. 7, n° 5*, páginas 569–586, 1999.

[24] R. Short e K. Fukunaga. The optimal distance measure for nearest neighbor classification. *IEEE Transactions of Information Theory, vol. IT-27, n°. 5*, página 622–627, Maio 1981.

[25] D. Jurafsky e J. H. Martin. *Speech and Language Processing: An Introdution to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall, 2009.

[26] D. A. Reynolds. Speaker identification and verification using Gaussian mixture speaker models. *Speech Communication 17*, páginas 91–108, 1995.

[27] D. A. Reynolds, T. F. Quatieri, e R. B. Dunn. Speaker verification using adapted Gaussian mixture models. *Digital Signal Processing Vol. 10, n° 1-3*, páginas 19–41, 2000.

[28] W. M. Campbell, D. E. Sturim, e D. A. Reynolds. Support vector machines using GMM supervectors for speaker verification. *Signal Processing Letters, IEEE, Volume 13, n° 5*, páginas 308–311, 2006.

[29] W. M. Campbell, D. E. Sturim, D. A. Reynolds, e A. Solomonoff. SVM based speaker verification using a GMM supervector kernel and NAP variability compensation. *IEEE International Conference on Acoustics, Speech and Signal Processing*, páginas 97–100, 2006.

[30] P. Kenny, G. Boulianne, P. Ouellet, e P. Dumouchel. Joint factor analysis versus eigenchannels in speaker recognition. *IEEE Transactions on Audio, Speech, and Language Processing, Vol. 15, n.º 4*, páginas 1435–1447, 2007.

[31] D. Reynolds e R. Rose. Robust text-independent speaker identification using Gaussian mixture speaker models. *IEEE Transactions on Speech and Audio Processing*, páginas 72–83, Janeiro 1997.

[32] K. R. Farrell, R. J. Mammone, e K. T. Assaleh. Speaker recognition using neural networks and conventional classifiers. *IEEE Transactions on Speech and Audio Processing, Vol. 2, nº 1*, páginas 194–205, 1994.

[33] M. Przybocki e A. Martin. NIST speaker recognition evaluation chronicles. *Proc. Odyssey '04, Toledo, Espanha*, 2004.

[34] NIST SRE 2010, plano de avaliação. http://www.itl.nist.gov/iad/mig/tests/sre/2010/index.html, Junho 2011.

[35] NIST SRE 2008. http://www.itl.nist.gov/iad/mig/tests/sre/2008/official_results/index.html, Junho 2011.

[36] J.-F. Bonastre, N. Scheffer, D. Matrouf, C. Fredouille, A. Larcher, A. Preti, G. Pouchoulin, N. Evans, B. Fauve, e J. Mason. ALIZE/SpkDet: a state-of-the-art open software for speaker recognition. *Proc. IEEE Odyssey, ISCA Speaker Recognition Workshop*, 2008.

[37] ALIZE/SpkDet. http://mistral.univ-avignon.fr/mistral_dev/?page_id=6, Junho 2011.

[38] FoCal. http://www.dsp.sun.ac.za/~nbrummer/focal/index.htm, Junho 2011.

[39] The NIST year 2008 speaker recognition evaluation plan, http://www.itl.nist.gov/iad/mig/tests/sre/2008/sre08_evalplan_release4.pdf, Junho 2011.

[40] M. Graciarena, S. Kajarekar, N. Scheffer, E. Shriberg, A. Stolcke, L. Ferrer, e T. Bocklet. The SRI NIST SRE08 speaker verification system. *NIST SRE Workshop 2008, Montreal*, 2008.

[41] T. Hastie, R. Tibshirani, e J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2009.

[42] I. Hernáez, I. Saratxaga, J. Sanchez, E. Navas, e I. Luengo. Use of the harmonic phase in speaker recognition. *Interspeech 2011*, páginas 2757–2760, 2011.

[43] Voicebox, disponível em http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html, Junho 2011.

[44] Weka, disponível em: http://www.cs.waikato.ac.nz/ml/weka/, Junho 2011.

[45] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, e I. Witten. The WEKA data mining software: An update. *Special Interest Group on Knowledge Discovery and Data Mining, Vol. 11*, páginas 10–18, 2009.

[46] Auditory Toolbox, disponível em http://cobweb.ecn.purdue.edu/~malcolm/interval/1998-010/, Junho 2011.

[47] Statistics Toolbox, http://www.mathworks.com/products/statistics/index.html, Junho 2011.

[48] A. Ferreira. Static features in real-time recognition of isolated vowels at high pitch. *Journal of the Acoustical Society of America, Vol. 122, nº 4*, páginas 2389–2404, 2007.

[49] K. P. Markov e S. Nakagawa. Text-independent speaker recognition using multiple information sources. *Proc. ICSLP-98*, páginas 173–176, Dezembro 1998.

[50] P. Birjandi e M. A. Salmani-Nodoushan. *An Introduction to Phonetics*. Zabankadeh Publications, 2005.

[51] C. McCully. *The sound strucuture of English: An Introduction*. Cambridge University Press, 2009.

[52] B. Martin. Instance-based learning: Nearest neighbour with generalisation. Tese de Mestrado, University of Waikato, 1995.

[53] J. Platt. Sequential minimal optimization: A fast algorithm for training support vector machines. *Advances in Kernel Methods - Support Vector Learning, MIT Press*, páginas 185–208, 1999.