Faculdade de Engenharia da Universidade do Porto


# Reconhecimento de Orador em Dois Segundos

Diana Rocha Mendes

Relatório Final de Preparação da Dissertação

Realizado no âmbito do

Mestrado Integrado em Engenharia Electrotécnica e de Computadores

Major Telecomunicações

Orientador e Co-orientador: Prof. Dr. Aníbal Ferreira

15 de Fevereiro de 2011

# Índice

# 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 [1]. 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 [2].

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 já 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 [3]. 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 penas em que é necessário confirmar a localização do indivíduo. Para tal, é feita automaticamente uma chamada para o local onde é previsto estar, e a identidade do orador é confirmada [3].

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 exigem utilização de equipamento proprietário ou equipamento adaptado à tecnologia em causa [3]. Esta característica traduz-se em vantagens a nível de variedade de áreas de implementação e também em autonomia dos sistemas implementados. Por exemplo, ao ser usado o telefone para fazer o reconhecimento de orador, esse reconhecimento pode ser automático, 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* (ou até a substituíram completamente) o reconhecimento por voz. É muito usado em *telephone banking*, 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 Glenview State Bank of Illinois implementou pela primeira vez esta tecnologia no seu serviço de *telephone banking* [3].

## 1.3 – Contextualização e Objectivos

O trabalho que se pretende desenvolver consiste numa solução de reconhecimento de orador, especificamente do âmbito de identificação de orador independente de texto. A implementação do sistema deve ter em vista a redução do tempo das amostras de voz necessárias para a identificação, mantendo níveis de robustez elevados, comparáveis a soluções do estado da arte. O objectivo é reduzir esse tempo para 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, com características que indiquem potencial capacidade de alcançar o objectivo pretendido.

No presente relatório será feita uma análise do estado da arte em reconhecimento de orador, com restrição ao âmbito em que se insere o trabalho a efectuar. Será 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 será feita uma breve caracterização dos métodos mais frequentemente utilizados e extensivamente analisados na literatura. Serão também apresentados alguns métodos actualmente em estudo que apresentam resultados promissores e podem vir a constituir a base para a solução a implementar no âmbito da Dissertação. Ainda compreendido no estado da arte, apresenta-se uma análise comparativa de alguns sistemas implementados.

Por fim é incluída a planificação do projecto e calendarização das etapas a completar, assim como as ferramentas e metodologias que serão utilizadas ou estudadas em maior profundidade no decorrer da elaboração da dissertação.

# 2 - Estado da Arte

Um sistema de reconhecimento de orador é geralmente constituído pelos seguintes componentes: extracção de características, *pattern matching* e decisão, como ilustrado na figura 1. Neste capítulo será feita uma explicação breve do funcionamento destes componentes e das principais técnicas que são utilizadas em cada um deles.

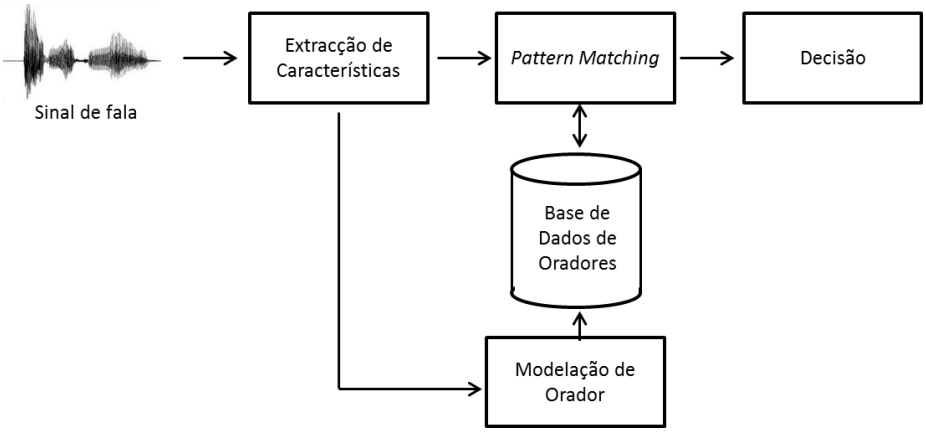


Figura 1: Estrutura genérica de um sistema de reconhecimento de orador.

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 [4].

## 2.1 - 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 [1][4]. 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 [4].

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 [5][4].

### 2.1.1 - Codificação Linear Preditiva

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

$$s_n = -\sum_{k=1}^{p} a_k \cdot s_{n-k} + G \cdot u_n \quad (1)$$

Na expressão (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 fonatório, 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.

$$\hat{s}_n = -\sum_{k=1}^{p} a_k \cdot s_{n-k} \quad (2)$$

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

$$e_n = s_n + \sum_{k=1}^{p} a_k \cdot s_{n-k} \quad (3)$$

Deduz-se a partir de (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 perceptual no contexto de modelação do filtro, como Rácios *Log Area* [2]. 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 [4].

### 2.1.2 - *Mel-Frequency Cepstrum 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 [4].

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:

$$\text{Cepstrum}(\textit{frame}) = \text{IDFT}(\log(|\text{IDFT}(\textit{frame})|)) \quad (4)$$

Segue-se uma breve explicação da expressão (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 obtemos uma representação das duas componentes do sinal de voz em que estas se encontram perfeitamente distintas uma da outra.

Os coeficientes cepstrais *Mel* distinguem-se dos coeficientes cepstrais descritos pelo facto de a sua obtenção exigir um passo extra: a transformação das frequências segundo a escala *Mel* (daí existir também a designação *Mel-Warped Cesptrum*, pois faz-se *warp* no domínio das frequências). Estas transformação é feita de forma a dar menos ênfase às altas frequências, devido ao facto do sistema auditivo humano apresentar menor sensibilidade a estas.

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 modelizada 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 [2], daí serem um dos métodos mais utilizados actualmente [6].

## 2.2 - *Pattern Matching* e Modelação

*Pattern Matching,* consiste em gerar uma pontuação de correspondência (*match score*) entre o modelo da voz do orador de entrada e modelos previamente conhecidos [2]. 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 [4].

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 [2]. 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 apenas médias globais são usadas – denominam-se métodos independentes do tempo [2][4][5].

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 [5][4][2].

Apesar de os 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 GMMs. Métodos baseados nas médias das características geralmente apresentam resultados sub-óptimos e são particularmente sensíveis a variações do canal e a ruído de fundo [5][1]. 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* [2].

### 2.2.1 - *Vector Quantization*

*Vector Quantization* é um exemplo de um método *template.* Este baseia-se na construção de um *codebook* para cada orador no sistema, a partir de métodos de agrupamento (*clustering*), aplicados sobre os vectores de características. A correspondência é feita através da distância entre os vectores de características do orador em análise e os diferentes *codebooks*, sendo que o *codebook* que apresentar menor distância é seleccionado [2][6].

### 2.2.2 - *Gaussian Mixture Model*

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.

Uma densidade *Gaussian mixture* é a soma pesada das M densidades componentes:

$$p(\vec{x}|\lambda) = \sum_{i=1}^{M} p_i b_i(\vec{x}) \quad (5)$$

onde $\vec{x}$ é um vector de *D* dimensões, $b_i(\vec{x}), i = 1,\ldots,M$ são as densidades componentes e $p_i, i = 1,\ldots,M$ são os pesos atribuídos. Cada densidade componente é uma função gaussiana da forma:

$$b_i(\vec{x}) = \frac{1}{(2\pi)^{D/2}|\Sigma_i|^{1/2}} \exp\left\{-\frac{1}{2}(\vec{x} - \overrightarrow{\mu_i})'\Sigma_i^{-1}((\vec{x} - \overrightarrow{\mu_i})\right\} \quad (6)$$

com $\vec{\mu}_i$ como o vector de média e matriz de covariância $\Sigma_i$.

Uma densidade *Gaussian mixture* é parametrizada por vectores de média, matrizes de co-variância e pesos associados. Estes parâmetros são representados da seguinte forma:

$$\lambda = \{p_i, \overrightarrow{\mu_i}, \Sigma_i\}, i = 1, \ldots, M \quad (7)$$

Cada orador é representado por um modelo $\lambda$.

Existem diversos modelos de matrizes de covariância. Pode ser utilizada uma matriz por cada componente gaussiana (covariância nodal) ou uma matriz única para todas as componentes (covariância global). As matrizes utilizadas podem ter também dois formatos: completo ou diagonal [7]. Estes parâmetros afectam o desempenho global do modelo GMM usado, como será visto em 2.4.

## 2.3 – Análise comparativa de sistemas implementados

Nos subcapítulos anteriores foram apresentados alguns dos métodos mais frequentemente utilizados em sistemas de reconhecimento de orador, tanto a nível de características do sinal de voz seleccionadas e extraídas, como a nível de construção dos modelos de orador, medidas de correspondência entre esses modelos e de classificação.

Perante estes métodos alternativos, é pertinente questionar sobre o seu desempenho relativo entre estes. No entanto, a comparação directa entre os diferentes métodos é difícil, por diversos factores. Por um lado, existem demasiadas alternativas de implementação dos diferentes módulos de um sistema de reconhecimento de orador. Por exemplo, no âmbito das GMMs podem ser escolhidas técnicas de modelação da variabilidade entre sessões como *latent factor analysis* (FA) e *nuisance atribute projection* (NAP) [8], podem ser utilizados diferentes métodos para estimação dos parâmetros da GMM como *expectation-maximization* (EM), entre muitas outras possibilidades. Por outro lado, os sistemas podem ser testados em diferentes condições, com amostras de diferentes durações, usando diferentes bases de dados, o que limita a comparação entre sistemas que não foram alvo da mesma metodologia de teste.

Neste capítulo são apresentados alguns sistemas de reconhecimento de orador publicados em artigos científicos. Em cada um destes foram estudados os efeitos de uma ou várias evoluções em relação a sistemas de estado da arte, na data de realização do estudo, e apresentam-se resultados comparativos entre a solução proposta e soluções anteriores. Essa evolução é restrita geralmente a uma parte do sistema de reconhecimento de orador: na parte de extracção de características ou nos métodos de *matching*, classificação ou decisão. Assim, a comparação é feita entre sistemas que diferem apenas numa dessas partes, e utilizam o(s) mesmo(s) método(s) nas outras. Com base nestes resultados pretende-se obter algumas

indicações sobre o desempenho relativo de alguns métodos e em que situações podem apresentar vantagens ou desvantagens.

## 2.3.1 - Métodos de Modelação e *Pattern Matching*

### 2.3.1.1 - GMM-UBM e GMM/SVM

GMM-UBM (Gaussian Mixture Model – Universal Background Model) é uma variante do método GMM, em que em vez de se fazer treino dos modelos através do critério de Verossimilhança Máxima, cada modelo é criado a partir de uma adaptação Bayesiana do UBM [9].

Em [8] é proposta uma abordagem que combina GMM-UBM com *Support Vector Machines* (SVMs), com o intuito de obter a robustez da modelação estatística que os sistemas GMM-UBM alcançam com o poder discriminatório dos SVMs. Relativamente a sistemas que utilizam apenas GMM-UBM, este método apresenta um aumento de 18% em termos de EER (Equal Error Rate). EER representa o ponto em que a probabilidade de uma falsa aceitação é igual à probabilidade de uma falsa rejeição [8].

Foi também concluído que este método apresenta resultados ainda melhores para testes de duração elevada, isto é, quando se usam amostras de voz para teste mais longas. Ao serem usados na sessão de treino três ficheiros de voz, ao invés de um, os valores de EER apresentam uma melhoria significativa (de 2,96% de EER para 1,04%).

Outro estudo de comparação entre GMM-UBM e GMM-SVM é feito em [10], e neste o último método também teve resultados relativos superiores. É também proposto um método de "fusão" entre as diferentes técnicas. Para mais detalhe ver [10].

### 2.3.1.2 – VQ e GMM

Analisam-se de seguida dois artigos que fazem a comparação entre GMMs e o método *template* VQ. Enquanto que o primeiro ([7]) afirma a superioridade dos Gaussian Mixture Models pela elevada taxa de correctas identificações que consegue atingir, o segundo ([6]) apresenta um sistema em que os métodos VQ e GMM apresentam taxas de erros similares, mas em que o VQ demonstra superior eficiência computacional.

Em [7] foram testadas duas alternativas para ambas as técnicas apresentadas: um sistema em que os modelos VQ utilizavam 50 vectores por *codebook* (VQ-50), outro que utilizavam 100 vectores (VQ-100), uma variação do sistema com GMMs em que estas apresentavam 50 componentes gaussianas com covariância nodal (GMM-nv) e por fim outra variação em que apresentavam também 50 componentes mas com uma única variância global por modelo (GMM-gv). Para testes de 5 segundos, a percentagem de identificações correctas foi superior para o modelo GMM-nv, com o sistema VQ-100 em segundo lugar, por uma diferença de 1,6%.

| MODELO | IDENTIFICAÇÃO CORRECTA (%) |
|---|---|
| GMM-nv | 94,5 |
| VQ-100 | 92,9 |
| VQ-50 | 90,7 |
| GMM-gv | 89,5 |

Tabela 1: Percentagem de identificações correctas atingida pelos sistemas GMM-nv, GMM-gv, VQ-100 e VQ-50.

Já no sistema descrito em [6], através de um método de pré-quantização, em que se reduz o débito binário dos vectores de características extraídos, as taxas de erro na identificação são bastante próximas, e até menores para o método VQ, dependendo do tamanho do *codebook*. O sistema apresentado foca-se em reduzir o tempo necessário para fazer a identificação, e enquanto que a pré-quantização faz com que o método GMM seja bastante mais rápido (nalguns casos, causa uma diminuição de 37,93 segundos para 1,11), o método VQ apresenta consistentemente menor complexidade computacional, e atinge assim valores de tempo de 0,64 segundos. Neste cenário, em que a agilidade do sistema é priorizada, este pode ser considerado o método mais vantajoso.

## 2.3.2 - Extracção de Características

### 2.3.2.1 - LPCCs e MFCCs

Usando GMMs como método de modelação dos oradores, foi feita em [11] uma comparação entre os dois tipos de características mais usados – LPCCs e MFCCs. Para o cálculo dos LPCCs foi usada ordem de predição 14, e o número de vectores extraídos para ambos os métodos foi de 10.

Realizaram-se testes com diferentes parâmetros dos GMMs usados e em duas bases de dados diferentes: a TIMIT, para excertos de fala sem ruído, e a NTIMIT, que contém excertos de fala por telefone, afectados naturalmente por ruído. Os resultados, em percentagem de identificações correctas, apresentam-se de seguida:

| PARÂMETROS GMM | LPCC (%) | MFCC (%) |
|---|---|---|
| 4/completa | 92,8 | 96,3 |
| 8/completa | 95,6 | 96,6 |
| 32/diagonal | 93,3 | 95,9 |
| 64/diagonal | 91,4 | 95,5 |

Tabela 2: Percentagem de identificações correctas, utilizando a base de dados TIMIT.

| PARÂMETROS GMM | LPCC (%) | MFCC (%) |
|---|---|---|
| 4/completa | 49,7 | 50,0 |
| 8/completa | 54,5 | 52,4 |
| 32/diagonal | 50,3 | 49,7 |
| 64/diagonal | 49,4 | 47,9 |

Tabela 3: Percentagem de identificações correctas, utilizando a base de dados NTIMIT.

Os parâmetros dos modelos GMMs que foram variados neste estudo foram o número de *mixtures* utilizadas (4, 8, 32 e 64) e o tipo de matrizes utilizadas (completa e diagonal), como indicado nas tableas.

Pode-se concluir que MFCCs têm melhor desempenho para amostras sem ruído. Em [11] atribui-se essa superioridade ao facto dos MFCCs incluírem alguma informação da fonte glótica, enquanto que os LPCCs descartam-na completamente (fica contida no erro de predição). Para amostras mais ruidosas o desempenho dos dois tipos de coeficientes é similar.

#### 2.3.2.2 - Frequência Fundamental e MFCCs

Apesar de ter havido uma grande evolução nos sistemas de reconhecimento de orador ao longo das últimas décadas, a extracção de características tem permanecido relativamente inerte, pelo que a maioria dos sistemas têm-se baseado na modelação do tracto vocal e ignorado a informação contida nas características da fonte glótica.

No entanto, nos últimos anos têm-se investigado métodos de extracção de características que preservem não só informação relativa à fonte glótica como também levem em consideração a dependência entre a fonte e o tracto vocal. Uma metodologia que incorpora estas características, através da geração de modelos de vectores de características para cada intervalo de frequências fundamentais da voz, é proposta em [12]. Na prática, o que é feito é integrar uma probabilidade *a posteriori* de observar um vector de MFCCs, dada uma determinada frequência fundamental. Este método é comparado com um que utiliza extracção dos MFCCs usuais (sistema *baseline*), ambos com modelação por GMMs. Os resultados apresentam-se na seguinte tabela:

| TEMPO (SEGUNDOS) | SISTEMA *BASELINE* (%) | MFCCS+F.FUNDAMENTAL(%) |
|---|---|---|
| 0,1 | 36,8 | 40,5 |
| 0,5 | 63,4 | 69,4 |
| 1 | 75,4 | 80,8 |
| 2 | 84,2 | 88,0 |
| 3 | 88,0 | 90,5 |
| 4 | 90,0 | 93,3 |
| 5 | 91,4 | 94,7 |
| 6 | 92,7 | 95,2 |

Tabela 4: Comparação entre as percentagens de identificações correctas atingido pelo sistema *baseline* e pelo sistema que utiliza MFCCs e Frequência Fundamental.

O tempo refere-se ao tempo das amostras utilizadas para teste. Como se pode observar, a maior melhoria em termos de desempenho oferecida por esta abordagem verifica-se para testes de duração menor que três segundos, atingindo até um aumento de identificações correctas de 4,3% para amostras de 500 milissegundos.

### 2.3.2.3 - Fonte e MFCCs

Partindo da mesma motivação mencionada em 2.3.1.1, outro estudo foi realizado com objectivo de explorar novas características para além de LPCCs e MFCCs clássicos. Em [13] são apresentadas sete novas características: *spectral centroid* (SC), *spectral bandwidth* (SBW), *spectral band energy* (SBE), *spectral crest factor* (SCF), *spectral flatness measure* (SFM), *Shannon entropy* (SE) e *Renvi entropy* (RE). Os testes foram feitos utilizando sistemas com base em GMMs, e a base de dados escolhida foi a TIMIT. De seguida apresentam-se as percentagens de identificações correctas que foi possível atingir ao utilizar estas características em conjugação com MFCCs e ΔMFCCs (derivada de primeira ordem dos MFCCs):

| CARACTERÍSTICAS | IDENTIFICAÇÃO CORRECTA (%) |
|---|---|
| MFCC & ΔMFCCs | 95,30 |
| MFCC & ΔMFCCs & SC | 97,32 |
| MFCC & ΔMFCCs & SBE | 97,32 |
| MFCC & ΔMFCCs & SBW | 96,98 |
| MFCC & ΔMFCCs & SCF | 96,31 |
| MFCC & ΔMFCCs & SFM | 81,55 |
| MFCC & ΔMFCCs & SE | 90,27 |
| MFCC & ΔMFCCs & RE | 98,32 |
| MFCC & ΔMFCCs & SBE & SC | 96,98 |
| MFCC & ΔMFCCs & SBE & RE | 96,98 |
| MFCC & ΔMFCCs & SC & RE | 99,33 |

Tabela 5: Percentagem de identificações correctas para diferentes características de fonte extraídas.

Estes resultados comprovam que o uso de características relacionadas com a fonte glótica traz ao sistema informação discriminatória importante sobre o orador. Consegue-se obter uma melhoria em termos de percentagem de identificações correctas de 4,03%, usando *Renvi entropy*. Seria interessante futuramente utilizar estas características juntamente com métodos mais sofisticados de modelação, como GMM-UBM, e inferir se seriam atingidos valores de robustez elevados para testes com amostras de duração menor, já que esse parâmetro não foi estudado em [13].

# 3 - Plano de Trabalhos

O plano de trabalhos previsto para o desenvolvimento do projecto inserido no âmbito da dissertação está representado no diagrama 1. Foram identificadas as principais etapas necessárias para o cumprimento dos objectivos propostos e foi feita uma estimativa do tempo que deverá ser despendido em cada uma delas.

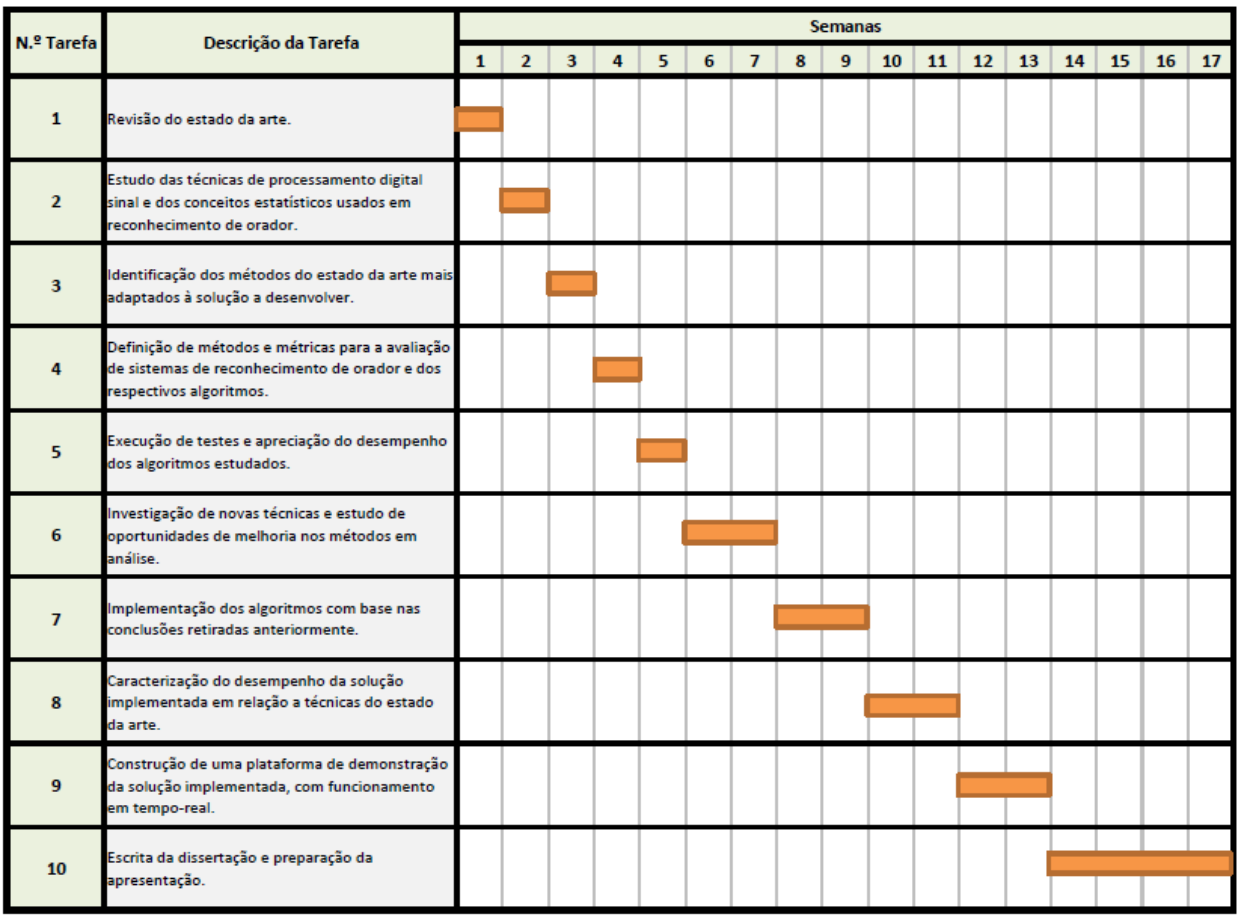


Diagrama 1: Plano de Trabalhos para o projecto a desenvolver, duração (em semanas) das tarefas previstas.

# 4 - Ferramentas

As ferramentas que poderão ser utilizadas no âmbito do desenvolvimento da solução a implementar serão mais profundamente estudadas no início do próximo semestre, pelo facto de ser necessária uma definição mais aprofundada do sistema para fazer a escolha do *software* mais indicado. No entanto, algumas possibilidades já foram exploradas no âmbito da disciplina Preparação da Dissertação. Houve já uma familiarização com as ferramentas Adobe Audition e Cool Edit, indicadas para edição e manipulação de ficheiros áudio. Outra ferramenta que poderá ser útil nesta área é o Praat, pelo que é especializado para sinais de fala.

Para desenvolvimento dos algoritmos e estudo dos diferentes métodos a principal ferramenta será o Matlab e a *toolbox* deste para processamento de fala, VoiceBox. Será estudado também o *software* de tratamento e análise estatística de dados, Weka.

Por último, uma ferramenta que tem vindo a ser utilizada no desenvolvimento de muitos sistemas de reconhecimento de orador recentes é o ALIZE. Este *software open-source* é orientado à implementação de soluções de reconhecimento de orador, embora possa também ser utilizado para implementação de outros sistemas como reconhecimento facial e de impressões digitais. São incluídas nesta *toolkit* os métodos mais actuais de modelação e classificação estatística, compensação de canal, entre outras ferramentas, todas estas optimizadas para identificação de orador independente de texto. Para além de desenvolvimento de soluções nas áreas indicadas, o ALIZE permite a avaliação do desempenho dos sistemas através do uso de bases de dados e de protocolos padronizados e utilizados a nível mundial, facilitando assim a comparação entre sistemas [14]. Pelas potencialidades que oferece e pelo nível de adaptação ao tipo de solução que se pretende desenvolver, esta ferramenta constituirá provavelmente a base para o trabalho.

# 5 - Conclusão

No presente relatório foi apresentada uma visão geral sobre o estado da arte em sistemas de reconhecimento de orador. Foi feita uma análise crítica de alguns destes sistemas, com o intuito de comparar em termos de desempenho as técnicas mais usadas. Este levantamento de tecnologias tinha como objectivo inferir sobre quais os métodos mais promissores, isto é, quais os métodos que indicam ter, com a exploração devida, capacidade para produzir os resultados pretendidos para a solução a implementar. Assim, as conclusões retiradas nesta fase irão orientar uma futura pesquisa mais aprofundada, que incidirá nas técnicas identificadas como as de maior interesse para o projecto de dissertação.

Atendendo ao facto de que a solução a implementar tem por objectivo reduzir o tempo das amostras necessário para identificação robusta, uma das áreas a explorar é o da extracção de características da fonte glótica, como se pode concluir pelos resultados em 2.3.2. No que toca a técnicas de *pattern matching* entre os modelos de vários oradores, foi visto que GMMs representam a base para a grande maioria dos sistemas actuais, pelos valores elevados de robustez que oferecem. Outros métodos alternativos foram estudados e foi concluído que apesar de apresentarem outras vantagens, nomeadamente em termos de eficiência computacional, não indicam ser os mais adequados para o sistema a implementar.

# 6 - Referências


[1] Li, H. e Ma, B., “Best of the Web”, *IEEE Signal Processing Magazine*, Novembro 2010, pp. 139-142

[2] Campbell, J. P., “Speaker Recognition: A Tutorial”, *Proceedings of the IEEE*, Setembro 1997, pp. 1437-1462

[3] Markowitz, J., “Voice Authentication in the Real World”, *National Center for Biometric StudiesConference, Voice Authentication for Identity Management*, 2006

[4] Karpov, E., “Real-Time Speaker Identification”. Tese de mestrado, Universidade de Joensuu, 2003

[5] Gish, H. e Schmidt M., “Text Independent Speaker Identification”, *IEEE Signal Processing Magazine*, Outubro 1994, pp. 18-32

[6] Kinnunen, T., Karpov, E. e Fränti, P., “Real-Time Speaker Identification”, 2002

[7] Reynolds, D. e Rose, R., “Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models”, *IEEE Transactions on Speech and Audio Processing*, Janeiro 1997, pp. 72-83

[8] Campbell, J. P. et al, “Forensic Speaker Recognition”, *IEEE Signal Processing Magazine*, Março 2009, pp. 95-103

[9] Zheng, R. et al, “Text-Independent Speaker Identification Using GMM-UBM and Frame Level Likelihood Normalization”, 2004

[10] Larcher, A., Lévy, C. e Matrouf, D., “LIA NIST-SRE’10 Systems”, 2010

[11] Markov, K. P. e Nakagawa, S., “Comparison between LPC Cepstrum and MFCC for Speaker Recognition using Clean and Telephone Speech”, 1999

[12] Ezzaidi, H., Rouat, J. e O’Shaughnessy, D., “Towards Combining Pitch and MFCC for Speaker Identification Systems”, 2001

[13] Hosseinzadeh, D. e Krishnan, S., “Combining Vocal Source e MFCC Features for Enhanced Speaker Recognition Performance Using GMMs”, 2007

[14] Bonastre, J.-F. et al, “ALIZE/SpkDet: a State-of-the-art Open Source Software for Speaker Recognition”, *Proc. Odyssey: the Speaker and Language Recognition Workshop*, 2008