VITOR MANUEL MARTINS CICOURO DE PERA

RECONHECIMENTO DE FALA CONTÍNUA COM PROCESSAMENTO SIMULTÂNEO DE DIFERENTES CARACTERÍSTICAS DO SINAL

FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO
2001

RECONHECIMENTO DE FALA CONTÍNUA COM PROCESSAMENTO SIMULTÂNEO DE DIFERENTES CARACTERÍSTICAS DO SINAL

VITOR MANUEL MARTINS CICOURO DE PERA

Dissertação apresentada na Faculdade de Engenharia da Universidade do Porto para obtenção do grau de Doutor em Engenharia Electrotécnica e de Computadores, sob a orientação do Professor Doutor Carlos Manuel Chaves Espain de Oliveira.

Junho 2001

004(043)/PERv/REC
UNIVERSIDADE DO PORTO
Faculdade de Engenharia
BIBLIOTECA
N.º 56105
CDU 004(043)
Data 23 / 10 / 20 01

Resumo

Nesta dissertação é apresentada uma abordagem ao problema da modelação acústica no reconhecimento automático da fala. A principal hipótese que se propôs investigar baseia-se no processamento simultâneo de múltiplas representações do sinal, contendo informação parcialmente diferente, com o objectivo de reduzir significativamente o erro de reconhecimento.

Com base num formalismo de descodificação que processa inicialmente de maneira independente e paralela as múltiplas sucessões de vectores de características extraídas do sinal, é desenvolvido um modelo estatístico original que estima a verosimilhança da observação conjunta dessas representações. De acordo com esse modelo, a influência local de cada uma das representações no processo global de decisão que conduz ao resultado do reconhecimento é dependente da classe acústica que está hipoteticamente associada ao respectivo segmento do sinal e é também dinamicamente adaptada em função de determinadas características extraídas do sinal. Os parâmetros do modelo são treinados segundo uma aproximação ao critério da maximização da verosimilhança da observação conjunta dos múltiplos vectores estendida ao conjunto de treino.

Um aspecto relevante nesta abordagem é o propósito de alargar a aplicação dos sistemas baseados em múltiplas representações do sinal a tarefas que se afastam daquelas que tipicamente lhes são atribuídas. Nesse sentido, este sistema é desenvolvido e avaliado numa tarefa de reconhecimento em que não se verificam problemas específicos para os quais a utilização de múltiplas representações é particularmente mais vantajosa.

O mérito das soluções propostas é avaliado experimentalmente numa tarefa de reconhecimento de fala contínua independente do falante e com um vocabulário de média dimensão. O desenvolvimento dos sistemas é suportado pela base de dados *Resource Managment* e a sua avaliação é feita usando o conjunto de teste *February'89*. Tomando como referência os resultados obtidos com um sistema baseado numa única representação acústica, a utilização de uma representação adicional conduz à diminuição da taxa de palavras erradas em aproximadamente $10\%_{rel}$, relativamente ao melhor dos sistemas convencionais. Sendo estes resultados obtidos com representações acústicas que apresentam uma dependência estatística significativa, é legítimo esperar que o modelo de recombinação das verosimilhanças proposto conduza a um desempenho ainda melhor se a combinação de características for menos desfavorável.

Esta dissertação está organizada em quatro capítulos. No primeiro são apresentadas algumas noções básicas na tecnologia do reconhecimento automático da fala. No segundo capítulo é descrito o desenvolvimento do reconhecedor que serve de referência a este trabalho, baseado numa única representação do sinal e na tecnologia dos modelos escondidos de Markov. No terceiro capítulo é apresentado o sistema baseado nas soluções propostas para a recombinação das múltiplas representações do sinal e são analisados os resultados do seu desempenho. Finalmente, são destacadas as principais conclusões deste trabalho e são discutidas algumas hipóteses de desenvolvimento.

Abstract

This thesis covers an approach to acoustic modelling for automatic speech recognition. The main hypothesis proposed for research is based on simultaneous processing of multiple signal representations containing partially different information to thereby significantly reduce recognition errors.

Based on a decoding formalism that, in an independent and parallel manner, initially processes the multiple features streams extracted from the signal, we develop an original statistical model that estimates the joint observation likelihood of those representations. According to that model, the local influence of each representation on the overall decision process leading to the recognition result is dependant on the acoustic class hypothetically associated to the respective signal segment and is also dynamically adapted according to specific characteristics extracted from the signal. The model's parameters are trained according to an approximation to the criteria of maximising the likelihood of the joint observation of the multiple vectors extended to the training set.

A relevant goal of this approach is to extend the application of the multi-stream systems so that it covers tasks different from those which it is normally assigned. As such, this system is developed and assessed in a recognition task in which there are no specific problems for which the use of multiple representations is particularly more advantageous.

The merit of the proposed solutions is experimentally assessed in a continuous speech recognition task independently of the speaker and using a medium-range vocabulary. The development of the systems is supported by the Resource Management database and is assessed using the February'89 test set. Using the reference results of a system based on a single acoustic representation, the use of an additional stream decreases the word error rate by approximately $10\%_{rel}$, compared to the best conventional systems. Since these results are obtained with acoustic representations that have a significant statistical dependency, it is legitimate to expect that the proposed likelihood recombination model will provide even better performance if the combination of characteristics is not as unfavourable.

This dissertation has four sections. The first presents some basic notions about automatic speech recognition technology. The second section describes the development of the recogniser that is the reference for this work, based on a single representation of the signal and based on the technology of the hidden Markov models. The third section presents the system based on proposed solutions for recombining multiple streams and analyses the system's performance. Lastly, the main conclusions of this work are highlighted and some development hypotheses are discussed.

Résumé

Cette dissertation présente une approche au problème de la modélisation acoustique dans la reconnaissance automatique du langage. La principale hypothèse qui y est étudiée se base sur le traitement simultané de représentations multiples du signe, contenant une information partiellement différente afin de réduire substantiellement l'erreur de reconnaissance.

A partir d'un formalisme de décodage qui traite à l'origine d'une manière indépendante et parallèle les multiples successions de vecteurs de caractéristiques extraites du signe, se développe un modèle statistique original qui évalue la vraisemblance de l'observation conjointe de ces représentations. Selon ce modèle, l'influence locale de chacune des représentations dans le processus global de décision qui mène à la reconnaissance, dépend de la classe acoustique hypothétiquement associée au respectif segment du signe, étant également adaptée d'une façon dynamique en fonction de certaines caractéristiques extraites du signe. Les paramètres du modèle sont travaillés selon une approche du critère de maximisation de la vraisemblance de l'observation conjointe des divers vecteurs élargie à l'ensemble du travail.

Un aspect significatif de cette approche réside dans l'objectif d'élargir l'application des systèmes fondés sur des représentations multiples du signe à des tâches qui s'éloignent de celles qui lui sont typiquement attribuées. Dans ce cadre, ce système est développé et évalué selon une tâche de reconnaissance où l'on ne constate pas de problèmes spécifiques pour lesquels l'emploi de représentations multiples serait particulièrement plus avantageux.

Le mérite des solutions proposées est évalué expérimentalement dans le cadre de la reconnaissance du langage continu indépendant du parlant, ayant un vocabulaire de dimension moyenne. Le développement des systèmes est soutenu par la base de données *Resource Management* et l'ensemble du test *February '89* est utilisé pour l'évaluer. En prenant pour référence les résultats obtenus avec un système fondé sur une seule représentation acoustique, l'utilisation d'une représentation additionnelle provoque la diminution du taux de mots erronés dans environ 10% des cas par rapport à l'un des meilleurs systèmes conventionnels. Etant donné que ces résultats ont été obtenus avec des représentations acoustiques présentant une dépendance statistique substantielle, il est légitime d'espérer que le modèle de nouvelle combinaison des vraisemblances proposé provoque une performance encore plus élevée si la combinaison de caractéristiques n'est pas si défavorable.

Cette dissertation est présentée en quatre parties. La première présente quelques notions de base concernant la technologie de la reconnaissance automatique du

langage. La seconde partie décrit le développement du reconnaisseur qui sert de référence à ce travail, fondé sur une seule représentation du signe et de la technologie des modèles cachés de Markov. La troisième partie présente le système fondé sur les solutions proposées pour combiner à nouveau les multiples représentations du signe et analyse les résultats de sa performance. Enfin, les principales conclusions de ce travail sont mises en évidence et certaines possibilités de développement sont discutées.

Agradecimentos

Desejo expressar o meu agradecimento ao Professor Doutor Carlos Espain e ao Professor Doutor Jean-Pierre Martens por me terem orientado neste trabalho e por terem sempre tentado ajudar-me a vencer dificuldades várias.

Agradeço a todos os que me apoiaram na realização deste trabalho.

Este trabalho beneficiou de uma bolsa de estudo concedida pela Fundação Ciência e Tecnologia.

Índice

1 Introdução **1**
1.1 A fala 1
1.2 O reconhecimento automático da fala 2
1.2.1 Conceito 2
1.2.2 Dificuldades 3
1.2.3 Aplicações 4
1.3 Sistemas de reconhecimento automático da fala 7
1.3.1 Função 7
1.3.2 Estrutura base 7
1.3.3 Conhecimento apriorístico e treino 9
1.3.4 Níveis de conhecimento 10
1.3.5 Modelos estatísticos e regras sintácticas 12
1.3.6 Sistema estatístico integrado para reconhecer fala contínua 14
1.3.6.1 Perspectiva histórica 14
1.3.6.2 Reconhecimento integrado de padrões 16
1.3.6.3 Reconhecimento estatístico de padrões 17
1.3.6.4 Gramática 19
1.3.6.5 Unidades acústicas 20
1.3.6.6 Léxico 21
1.3.6.7 Modelos acústicos 22
1.3.6.8 Treino 23
1.3.6.9 Reconhecimento 24
1.3.6.10 O *estado da arte* 25

2 Sistema básico de reconhecimento de fala contínua **26**
2.1 Introdução 26
2.2 Tarefa de reconhecimento 27
2.3 Bases de dados 28
2.4 Estrutura do reconhecedor 30
2.5 Análise do sinal 30
2.5.1 Introdução 30
2.5.2 O módulo extractor de características 33
2.5.2.1 Estrutura e resultado 33
2.5.2.2 Pré-ênfase do sinal 34
2.5.2.3 Formação de sectores e aplicação da janela Hamming 37
2.5.2.4 Cálculo dos coeficientes Mel-cepstrais 40
2.5.2.5 Cálculo da energia 47
2.5.2.6 Análise dinâmica 48
2.6 Modelação acústica 50
2.6.1 Introdução 50
2.6.2 Unidades fonéticas 51
2.6.2.1 Fundamentos 51

2.6.2.2 Descrição 55
2.6.3 Modelos escondidos de Markov 61
2.6.3.1 Introdução 61
2.6.3.2 Formalização dos modelos 64
2.6.3.3 Aplicação dos modelos 68
2.7 Modelação linguística 75
2.7.1 Introdução 75
2.7.2 Gramática 76
2.7.3 Léxico 78
2.8 Descodificação 79
2.8.1 Introdução 79
2.8.2 Estrutura integrada de informação 79
2.8.3 Algoritmo de descodificação 83
2.9 Treino do sistema 89
2.9.1 Introdução 89
2.9.2 Critério de treino 89
2.9.3 Processo de treino 92
2.9.3.1 Estratégia 92
2.9.3.2 Iniciação 95
2.9.3.3 Iteração nuclear de treino 97
2.9.3.4 Resultados 100

3 Sistema baseado em múltiplas representações do sinal 112
3.1 Introdução 112
3.2 Estrutura do reconhecedor 114
3.3 Análise do sinal 115
3.3.1 Introdução 115
3.3.2 Conceitos de Teoria da Informação 117
3.3.3 Características seleccionadas 118
3.3.4 Extracção das características 122
3.3.4.1 Características baseadas nos coeficientes cepstrais 122
3.3.4.2 Frequência fundamental 123
3.4 Classificação do sinal 126
3.4.1 Introdução 126
3.4.2 Critério de classificação 127
3.4.3 Formalização do modelo acústico *multi-stream* 128
3.4.4 Modelos acústicos compostos 130
3.4.5 Formalização do modelo de recombinação 135
3.4.6 Treino dos modelos acústicos compostos 141
3.4.6.1 Introdução 141
3.4.6.2 Treino dos parâmetros dos modelos *single-stream* 142
3.4.6.3 Treino do modelo de recombinação 147
3.4.7 Descodificação 151
3.5 Resultados 154
3.5.1 Comparação dos sistemas *single-stream* 154
3.5.2 Recombinação básica dos *streams* 156
3.5.3 Recombinação através do modelo definido 158

4 Conclusão **170**
4.1 Conclusões 170
4.2 Desenvolvimento do trabalho 172

Bibliografia **175**

Anexo: Léxico **190**

Capítulo 1

Introdução

Neste capítulo são apresentadas algumas questões básicas relacionadas com o reconhecimento automático da fala. Inicialmente são referidas algumas noções acerca dos processos de produção, transmissão e recepção da fala. Em seguida, é feita uma breve reflexão sobre a natureza do problema do reconhecimento da fala, a sua complexidade e dificuldade. São então analisados diferentes aspectos essenciais relacionados com os sistemas automáticos de reconhecimento da fala, sendo prestada uma particular atenção aos sistemas desenvolvidos para aplicações de fala contínua.

1.1 A fala

Embora o senso comum permita estabelecer com relativa facilidade uma noção empírica sobre o que é a fala, para compreender melhor a sua natureza é importante enquadrar o fenómeno acústico que lhe está associado no processo mais abrangente que é a comunicação oral. Um modelo de comunicação geralmente aceite, designado *cadeia da fala*, considera três fases fundamentais: produção, transmissão e recepção.

Na fase de produção podem ser distinguidos, segundo alguns autores [6], dois processos realizados pelo falante. São eles os processos de geração simbólica e de produção acústica. O processo de geração simbólica consiste na transformação da construção mental original, associada à mensagem que o falante pretende transmitir ao ouvinte, numa estrutura linguística. Assim, para que a comunicação se possa estabelecer sem ambiguidade e de maneira eficiente é necessário que o falante conheça o vocabulário e as regras gramaticais da língua que está a utilizar. Ao processo de produção acústica, que materializa esta representação simbólica, podem ser associadas determinadas características prosódicas que modificam o significado da mensagem. De acordo com as instruções determinadas e veiculadas pelo sistema nervoso, são accionados os músculos necessários para gerar o fluxo de ar proveniente dos pulmões e modelá-lo através das vibrações das cordas vocais e da geometria do tracto vocal. É assim criada uma onda de pressão acústica. O sinal acústico, que se propaga através do ar e também do corpo do falante, é captado pelo próprio aparelho auditivo do falante e

injectado num sistema de realimentação, permitindo a avaliação e o controlo do processo de produção.

Durante a transmissão, o sinal da fala é com frequência substancialmente afectado por diferentes tipos de ruído e distorção que ocorrem no canal.

Na recepção, o ouvinte capta o sinal acústico e tenta extrair a informação que nele está codificada. A onda acústica é captada pelo ouvido externo e é transformada, pelas secções média e interna do ouvido, em sinais neurológicos que são então interpretados no cérebro. Neste processo de descodificação, para além da informação estritamente contida no sinal de fala, geralmente o ouvinte socorre-se de outras fontes que de alguma maneira o possam ajudar como, por exemplo, a expressão facial ou os gestos do falante e os diferentes níveis de contexto que vão sendo definidos ao longo da conversação.

Pode então afirmar-se que a fala é um sinal, transmitido através dum campo acústico, cujo fim é a comunicação entre pessoas. Essa informação encontra-se codificada através duma sucessão de padrões acústicos que constituem uma representação abstracta da mensagem transmitida. Para além da informação linguística, evidenciada no modelo de comunicação *cadeia da fala*, geralmente a comunicação através do canal acústico envolve também, entre outras, informação física, social e psicológica sobre o falante[1].

1.2 O reconhecimento automático da fala

1.2.1 Conceito

O reconhecimento da fala consiste na percepção do conteúdo linguístico associado ao sinal acústico. Esta actividade de percepção pode ser entendida, no essencial, como uma operação de reconhecimento de padrões acústicos. Efectuar esta tarefa de maneira automática implica, portanto, o desenvolvimento de máquinas capazes de classificar os sinais acústicos da fala em determinadas representações simbólicas ou conduzir à execução de acções.

Neste último caso, é necessário dotar a máquina com alguma capacidade de entendimento, no sentido de habilitá-la a transformar os eventos acústicos nas instruções de comando que conduzem às acções correspondentes.

No caso dos reconhecedores automáticos que executam a transcrição ortográfica do sinal de fala[2], fazem-no avaliando uma relação funcional, previamente definida,

[1] Por exemplo, os sistemas de reconhecimento e de verificação do falante baseiam-se no reconhecimento de determinadas características da voz.

[2] Como é o caso dos sistemas desenvolvidos neste trabalho.

entre um espaço de representação dos padrões acústicos, assim como das dependências entre eles, e um espaço onde estão definidas todas as sucessões de palavras que são possíveis.

1.2.2 Dificuldades

O simples facto da fala ser tão natural ao homem, e além disso se estabelecer através de um sinal acústico unidimensional, poderia sugerir a conclusão apressada que o reconhecimento automático de fala constituiria um problema cuja solução plena seria fácil, ou pelo menos acessível. Pelo contrário, desde há algumas décadas, logo após as desilusões nos primeiros passos da tecnologia da fala, se verificou que é um problema extremamente complexo e difícil.

Curiosamente, essa dificuldade está relacionada com a facilidade que caracteriza a comunicação oral entre os homens. Na sua qualidade de ser social, o homem tem ao longo dos milénios sujeitado a fala e a língua, entendendo-a como o código [10] que rege a fala, a um processo continuado de evolução, adaptando-a continuamente às suas necessidades e disponibilidades. Além disso, ao longo de toda a sua vida aprende a utilizar e exercita naturalmente as suas faculdades de falar e ouvir. Estas são, portanto, actividades muito bem adaptadas, com todas as implicações que daí resultam, às capacidades cognitivas do homem. Logo, não será de estranhar que exista uma enorme dificuldade em emular essa actividade utilizando máquinas que apresentam, no actual estádio tecnológico, capacidades cognitivas muito rudimentares.

Esta acessibilidade tem também contribuído para que a fala seja um instrumento de comunicação extremamente rico, permitindo a transmissão de "mensagens" que estão para além do que se pode considerar como sendo informação útil à descodificação do conteúdo linguístico. Por exemplo, um segmento de sinal de fala com a duração de um segundo poderá conter, em média, cerca de três palavras, cada qual ortografada com cerca de cinco caracteres, portanto a transcrição ortográfica do conteúdo linguístico da fala poderia ser codificada a cerca de 75 *bps*[1]. Ora, frequentemente o sinal de fala é amostrado a cerca de 8 kHz e quantificado com 8 *bits* por amostra, pois caso esses parâmetros sejam significativamente mais baixos verifica-se a degradação do resultado do reconhecimento. Assim, a taxa de informação que poderá ser representada é algumas centenas de vezes superior à informação contida na representação ortográfica. Este breve exercício aritmético [13] ilustra quão redundante, na perspectiva do reconhecimento do conteúdo linguístico, é a informação contida no sinal da fala. A dificuldade reside precisamente em conseguir extrair a informação que se encontra diluída no sinal da fala e é essencial à descodificação linguística.

[1] 75 *bits por segundo*, considerando um alfabeto com o número de caracteres não superior a 32.

Isto contribui também para que o sinal de fala seja ambíguo e confuso. Além disso, um discurso natural está recheado de hesitações e utilização de palavras que não constam do vocabulário definido para a tarefa. Pode até ocorrer que o sinal da fala seja incompleto, isto é, parte da informação essencial associada à mensagem pode ser transmitida de outras formas ou pode ser implícita e comum ao falante e ao ouvinte.

Diversas outras características da fala têm sido também apontadas como importantes causas, sobretudo por se manifestarem em simultâneo, da dificuldade da tarefa do reconhecimento.

Sem dúvida que a sua grande variabilidade, que se verifica ao nível da realização acústica de cada unidade simbólica elementar mas também segundo a dimensão temporal associada à sucessão desses símbolos, é uma dessas características. Com efeito, o sinal da fala é muito dependente do falante e, embora geralmente em menor escala, varia também com um único falante em função da velocidade do discurso, do seu estado físico ou psíquico ou de outras condicionantes mais ou menos fortuitas.

Acresce que o sinal da fala é em geral contínuo, não estando portanto claramente definidas as fronteiras entre as unidades simbólicas. Além disso, fluindo as palavras continuamente, são acentuados os efeitos de co-articulação, verificando-se a assimilação e até a eliminação de alguns sons.

O ambiente acústico adverso em que os reconhecedores automáticos da fala operam, como sucede frequentemente, constitui uma enorme dificuldade. Até mesmo os melhores sistemas apresentam uma degradação muito significativa do seu desempenho quando operam em ambientes apenas ligeiramente, do ponto de vista humano, mais hostis. Os diferentes tipos de ruído ambiente e as limitações e alterações diversas no canal de transmissão são um grande desafio ao reconhecimento automático da fala.

Uma outra contrariedade, de natureza diferente mas também importante na tecnologia do reconhecimento da fala, reside na grande dificuldade em compreender e abordar globalmente um problema que apresenta múltiplas vertentes. O problema do reconhecimento da fala é multidisciplinar, envolvendo áreas do conhecimento tão diversas como a psicologia, a acústica, o processamento do sinal, a informática ou a linguística.

1.2.3 Aplicações

As aplicações da tecnologia do reconhecimento da fala são bastante diversificadas [23]. Na área das telecomunicações, são de destacar os sistemas de informação, de acesso a dados ou serviços por via telefónica. Existem também múltiplas aplicações em escritórios, desde *sistemas de ditado* para edição de documentos até controlo de sistemas informáticos através de comandos por voz. Em

ambientes fabris existem também inúmeras aplicações importantes para a tecnologia do reconhecimento da fala. Em algumas actividades profissionais específicas, como por exemplo em medicina, são de grande interesse os sistemas de reconhecimento automático da fala na produção de diferentes tipos de relatórios que utilizam vocabulários e gramáticas relativamente restritas. Muitos utensílios e aparelhos de utilização doméstica são também um alvo para aplicações comerciais da tecnologia da fala. Para além destas e de muitas outras aplicações, que são interessantes por constituírem factores económicos ou de comodidade e lazer mais ou menos relevantes, refiram-se também os sistemas desenvolvidos para o auxílio de indivíduos com deficiências auditivas.

O facto de geralmente as aplicações exigirem que os reconhecedores automáticos apresentem taxas de reconhecimento que se aproximem das alcançadas pelo homem e, por outro lado, dadas as limitações que subsistem na tecnologia da fala, torna-se necessário introduzir determinadas restrições na utilização dos sistemas.

É assim compreensível que os sistemas iniciais tivessem sido desenvolvidos para aplicações de fala isolada com um vocabulário extremamente limitado, permitindo apenas um falante e operando em ambiente pouco ruidoso.

Esse problema específico foi em termos práticos plenamente resolvido mas novos desafios foram sendo lançados, como por exemplo um sistema com as mesmas características mas capaz de operar com vocabulários maiores ou com uma baixa relação sinal ruído. Embora existam sistemas mais robustos do que outros, todos eles se mostram muito sensíveis à qualidade do ambiente acústico, sobretudo se comparados com o homem, pelo que esta continua a ser uma restrição significativa.

Apesar de continuarem a existir aplicações de grande interesse para os sistemas de reconhecimento de palavras isoladas, como são muitos sistemas de comando por voz, a falta de naturalidade que lhe está associada tornou imperioso eliminar essa restrição. Numa perspectiva histórica, os primeiros sistemas de reconhecimento de fala contínua foram desenvolvidos em larga medida tentando adaptar técnicas de reconhecimento de fala isolada[1].

Houve necessidade de manter fortes restrições quanto à variabilidade do sinal e à perplexidade da tarefa, ou seja, no início os sistemas eram dependentes do falante e continuavam a ser apenas capazes de operar com vocabulários muito pequenos. Além disso, um vocabulário relativamente grande exigiria a utilização de técnicas mais agressivas de redução do volume de cálculos[2] para poder obter um funcionamento em tempo real, aproximações que obviamente se poderiam reflectir negativamente no desempenho dos reconhecedores.

Gradualmente e em grande medida devido também ao acentuado progresso da tecnologia dos computadores, permitindo a execução em tempo real de algoritmos cada

[1] Solução parcialmente adoptada nos chamados sistemas de fala ligada.
[2] Como as técnicas de *prunning*.

vez mais poderosos, têm vindo a ser atenuadas as limitações impostas às tarefas de reconhecimento de fala contínua. Os melhores sistemas independentes do falante são capazes de utilizar vocabulários da ordem das dezenas de milhar de palavras e ainda assim apresentar taxas de erro aceitáveis, desde que a relação sinal ruído seja relativamente elevada e não ocorram outras perturbações significativas do sinal da fala durante a sua transmissão. No entanto, estes resultados apenas têm vindo a ser alcançados usando limitações linguísticas ao nível sintáctico, associadas geralmente a gramáticas bastante restritivas, que reduzem muito significativamente a perplexidade[1] da tarefa do reconhecimento. Porém, essas gramáticas são demasiado artificiais na maior parte das aplicações e sobretudo quando a fala é espontânea.

A tabela 1.1, apresentando resultados do desempenho de sistemas de reconhecimento automático da fala para determinadas aplicações em ambiente acústico favorável, ilustra alguns dos aspectos acabados de referir.

Fala	Vocabulário	Gramática (perplexidade)	Dependência do falante	Taxa de erro (%)
isolada	10 dígitos	não (10)	sim	0.0
			não	0.1
	39 α-dígitos	não (39)	sim	4.5
			não	7.0
contínua	991 palavras	sim (60)	não	3.0
	1.800 palavras	sim (25)		3.0
	20.000 palavras	sim (145)		12.0

Tab. 1.1 - Taxas de palavras erradas para diversos sistemas de reconhecimento.

Por outro lado, muitas tarefas em reconhecimento da fala apresentam elas próprias restrições ao nível semântico e, eventualmente, pragmático que deveriam também ser utilizadas pelo reconhecedor automático. Porém, apesar do esforço de investigação que tem sido desenvolvido nesta área, continua a ser muito difícil dotar os reconhecedores com mecanismos eficientes para representar e tratar essa informação de nível mais elevado.

[1] Estimada como o número médio de ramificações em cada ponto de decisão, na perspectiva do processo de descodificação ser a pesquisa sobre um grafo de frases possíveis.

1.3 Sistemas de reconhecimento automático da fala

1.3.1 Função

Dado um sinal de fala definido num intervalo de tempo, a função de um reconhecedor automático da fala é produzir a respectiva sucessão[1] correcta de símbolos ou instruções de comando para executar o conjunto de acções correspondentes. A informação relevante para esta transformação encontra-se codificada no sinal através de uma sucessão de padrões acústicos, portanto, um reconhecedor automático de fala pode ser visto essencialmente como um reconhecedor de padrões[2] que apresentam também uma dimensão temporal.

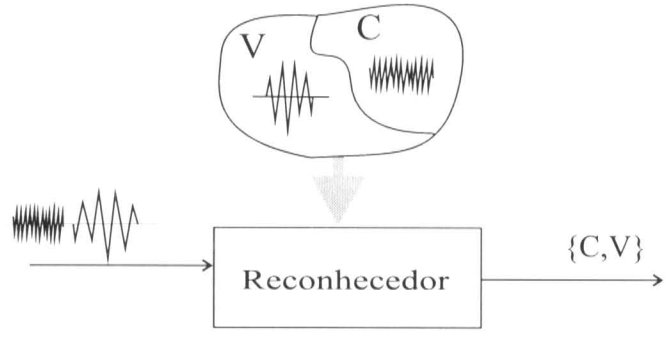

Fig. 1.1 - Reconhecedor automático de sucessões com consoantes e vogais.

Embora este problema seja significativamente mais complexo do que uma operação de classificação de padrões estáticos, esta perspectiva é considerada na figura 1.1 pois permite abordar de maneira simples alguns conceitos gerais. Assim, assumindo que cada padrão acústico pertence a uma classe única[3], o reconhecedor faz a partição do espaço abstracto de representação do sinal em regiões disjuntas, cada qual com uma única identidade de classe.

1.3.2 Estrutura base

Um reconhecedor de padrões, e em particular um reconhecedor automático de fala, é constituído por duas partes fundamentais: um módulo de análise do sinal e um módulo de classificação.

---

[1] Nos reconhecedores de palavras isoladas essa sucessão consiste num único termo.

[2] Em sentido lato, abrangendo as dependências entre os padrões que compõem a sequência se estes forem entendidos como entidades isoladas.

[3] Nos chamados sistemas *fuzzy* não é necessariamente assim.

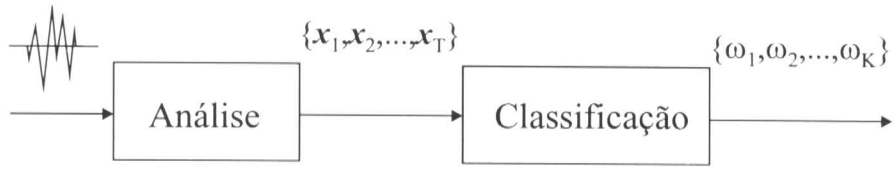

Fig. 1.2 - Estrutura básica de um reconhecedor automático de fala.

O módulo de análise converte o sinal de entrada numa representação mais adequada ao processo de classificação.

Seja $s(t)$ um sinal de fala definido num determinado intervalo de tempo. O módulo de análise transforma esse sinal numa sucessão de T vectores de dimensão D:

$$\boldsymbol{X}_1^T = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_T\}, \quad \boldsymbol{x}_i \in \Re^D \tag{1.1}$$

Este módulo deve ser desenhado com o objectivo de extrair, do sinal original, apenas a informação que é relevante para o processo de classificação. Assim, as componentes do vector de características:

$$\boldsymbol{x}_i = \left(x_i^1, x_i^2, \ldots, x_i^D\right) \tag{1.2}$$

devem favorecer a separação entre as classes, isto é, a representação dos vectores no espaço de decisão deve apresentar uma variação pequena no interior de cada classe e grande entre as diferentes classes.

Para além de facilitar a discriminação entre as classes, o módulo de análise determina geralmente uma representação mais compacta do sinal inicial permitindo acelerar o processo de classificação e, eventualmente, a utilização de algoritmos mais poderosos. Por outro lado, uma escolha criteriosa ou baseada na evidência experimental do intervalo de tempo que decorre regularmente entre os sucessivos vectores de características evita a eventual perda de informação, relevante para o processo de classificação, devido ao aumento da granulosidade temporal. Baseando-se tipicamente os vectores de características na análise espectral do sinal, é fundamental que durante esses intervalos de tempo se verifique a sua quase-estacionaridade.

O módulo de classificação transforma a sucessão de vectores de características numa sucessão de símbolos[1] pertencentes a um vocabulário Γ e relacionados com as

[1] Em grande parte dos sistemas esses símbolos são as representações ortográficas das palavras pertencentes ao vocabulário de uma determinada língua.

classes padrão:

$$W_1^K = \{\omega_1, \omega_2, ..., \omega_K\}, \quad \omega_i \in \Gamma_\omega \tag{1.3}$$

O cerne do classificador é o modelo, constituído por um conjunto de funções discriminantes, que define as fronteiras entre as diferentes classes. Essas funções estabelecem portanto as regras numéricas, sintácticas ou híbridas que determinam o resultado da classificação.

É importante ter presente que, contrariamente ao que sucede nos classificadores estáticos, a aplicação das funções discriminantes é feita com base numa estrutura capaz de suportar a dimensão temporal associada a esta operação de classificação. Relacionado com isto, um aspecto essencial reside no facto de o classificador dever alinhar no tempo as duas sucessões, a dos vectores de características e a dos símbolos à saída, que são geralmente assíncronas.

Apesar da distinção feita entre a análise do sinal e a sua classificação, é muitas vezes conveniente que o desenvolvimento dos respectivos módulos seja integrado para poder obter um sistema que apresente um bom desempenho. Para além da questão óbvia de compatibilizar, aos diferentes níveis, os procedimentos que geram a sucessão dos vectores de características e os que a classificam, é muito importante desenhar cada módulo tendo também como objectivo potenciar as qualidades e reduzir os efeitos das deficiências do outro. Por exemplo, um sistema baseado num classificador que apresente incapacidade para modelar a correlação existente entre os sucessivos vectores de características deve ser compensado introduzindo termos no vector de características que representem em alguma medida essa correlação.

1.3.3 Conhecimento apriorístico e treino

Embora seja uma característica comum à generalidade dos sistemas de engenharia, interessa realçar que para desenvolver sistemas automáticos de reconhecimento de fala eficazes é fundamental dispor de conhecimento apriorístico, neste caso particular proveniente de áreas tão diversas como a neurologia ou a informática.

Um exemplo concreto da utilização desse conhecimento consiste na utilização da escala Mel na análise do sinal acústico, em resultado de estudos realizados na área da psico-acústica sobre o processo de percepção auditiva humana. Outro exemplo é a escolha da topologia e complexidade das estruturas utilizadas nos modelos acústicos[1] com base no conhecimento existente sobre a estrutura acústica do sinal da fala.

[1] Por exemplo, o número de estados e a topologia dum modelo escondido de Markov, ou o número de nós na camada escondida duma *rede neuronal* do tipo perceptrão.

Mas é também claro que não é possível desenvolver satisfatoriamente todas as componentes do sistema de reconhecimento automático da fala usando apenas conhecimento apriorístico. Mesmo que existisse uma teoria do reconhecimento automático da fala capaz de formular os princípios fundamentais a partir dos quais poderiam ser pesquisadas soluções estruturais óptimas para os problemas, certamente isso seria insuficiente pois no sistema continuariam a existir alguns parâmetros por definir. É geralmente seguido um procedimento, chamado treino do sistema, que utiliza a informação contida num conjunto de exemplos do sinal da fala para determinar esses parâmetros livres. É assim necessário que o sistema consiga apreender a informação real contida no conjunto de dados de treino e é importante que o faça de maneira automática. No caso do treino ser supervisionado, cada exemplo encontra-se anotado de acordo com o que se supõe ser a sua classe correcta. Em alguns sistemas o treino pode ser parcialmente efectuado segundo uma técnica não-supervisionada, sendo automaticamente destrinçadas algumas características evidenciadas pelo conjunto de treino.

Após o treino, o desempenho do sistema pode ser avaliado testando-o com exemplos de sinal da fala que não foram utilizados durante o processo de treino. Um conjunto desses exemplos é submetido ao sistema, comparando-se então os resultados do reconhecimento com as respectivas identidades de classe, que devem ser conhecidas de antemão. O desempenho assim medido é um indicador directo da capacidade de generalização do sistema.

Assim, o desenvolvimento dos sistemas automáticos de reconhecimento da fala pode ser entendido, numa perspectiva geral e simplificada, como um processo apresentando duas etapas. Primeiro, o conhecimento apriorístico permite seleccionar as técnicas a utilizar e determinar as restrições estruturais do sistema. Então, através de um procedimento de treino, as fronteiras de decisão no espaço de representação do sinal são ajustadas, segundo um determinado critério, à informação contida num conjunto de exemplos conhecidos do sinal da fala. O objectivo final é, obviamente, obter um módulo de análise do sinal e um conjunto de funções discriminantes que permitam cometer menos erros de classificação em dados não utilizados no treino, isto é, o sistema deve apresentar uma boa capacidade de generalização.

1.3.4 Níveis de conhecimento

Para que a generalidade dos sistemas de reconhecimento automático da fala e em especial os utilizados em aplicações mais exigentes possam alcançar um desempenho aceitável é necessário que sejam capazes de processar eficientemente a informação que aos diferentes níveis se encontra associada ao sinal da fala.

Nestes sistemas são geralmente identificados dois níveis de processamento, um deles acústico e o outro linguístico.

Ao nível acústico o reconhecedor tenta relacionar directamente segmentos[1] do sinal da fala com modelos acústicos que representam internamente esses segmentos. Essas correspondências são determinadas apenas com base em regras, de natureza numérica, de similitude entre as características físicas do sinal da fala e os modelos acústicos, pelo que este nível é geralmente considerado básico.

Ao nível linguístico, diferentes patamares de abstracção do sinal da fala podem ser considerados: lexical, sintáctico, semântico e pragmático. Supondo, como sucede frequentemente, que as unidades básicas são definidas ao nível do fonema, é necessário utilizar informação lexical para poder construir estruturas mais complexas como são as palavras. O conhecimento sintáctico define as regras que regem a construção de sucessões de palavras, ou frases. O léxico e a sintaxe são dependentes da língua. Ao nível semântico, tal como o nome indica, é analisada a significação das palavras e a evolução do seu sentido no contexto da mensagem. Finalmente, o nível pragmático aborda a significação da mensagem no contexto em que ela é produzida.

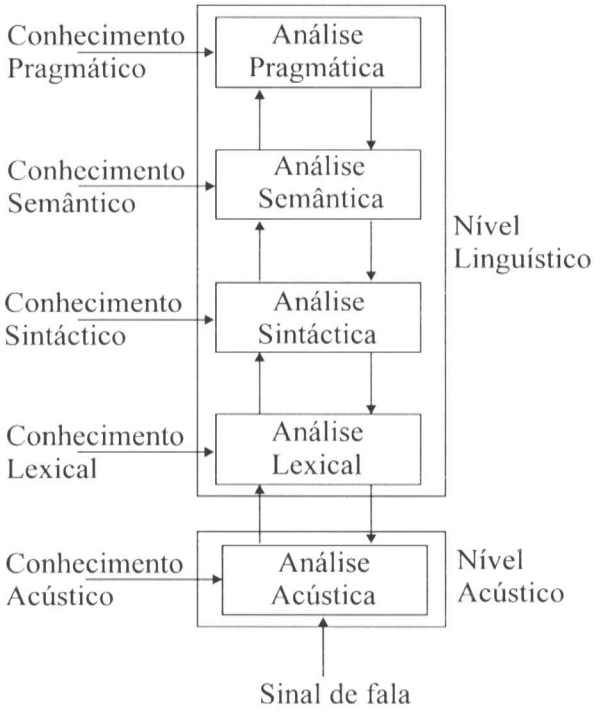

Fig. 1.3 - Modelo Peircian de um reconhecedor automático da fala.

Segundo o modelo Peircian, válido em grande parte dos reconhecedores automáticos da fala, estes diferentes níveis de processamento estão hierarquicamente estruturados.

[1] Nos sistemas de reconhecimento de fala contínua, frequentemente esses segmentos são as realizações acústicas de unidades elementares da língua como são os fonemas.

Ainda de acordo com este modelo, a informação pode fluir nessa estrutura em sentido ascendente ou descendente[1]. Assim, por exemplo, uma hipótese de reconhecimento estabelecida ao nível lexical pode ser submetida ao nível imediatamente superior, para verificação da sua correcção sintáctica, ou ao nível inferior, para avaliar a sua consistência face ao sinal observado.

Diversas outras estruturas, não hierárquicas ou híbridas, têm sido investigadas com o objectivo de criar mecanismos eficientes de interacção das múltiplas fontes do conhecimento. Fazendo uma brevíssima referência histórica, alguns dos sistemas que prestaram grande atenção a este problema conduziram a estruturas extremamente complexas mas que se revelaram ineficientes. Em parte, esse fracasso terá resultado do facto dessa complexidade acarretar enormes dificuldades no processamento da informação. Além disso, posteriormente, concluiu-se que em alguns casos era dada uma exagerada preponderância aos níveis de conhecimento mais elevados em detrimento do nível acústico[2].

No actual estádio tecnológico, os melhores resultados em sistemas de reconhecimento de fala contínua têm sido obtidos integrando de maneira equilibrada, numa única estrutura de informação, o conhecimento acústico, lexical e sintáctico. Nesses sistemas, geralmente é dada grande atenção ao processamento acústico de maneira a conseguir extrair do sinal a informação relevante para o processo de classificação. Em termos práticos, o léxico estabelece restrições na modelação da pronunciação das palavras e a sintaxe reduz a perplexidade da tarefa.

Sobretudo em aplicações mais exigentes, como é o caso dos sistemas de reconhecimento de fala espontânea, é plenamente reconhecida a necessidade de incorporar e interagir conhecimento aos níveis semântico e pragmático. Uma dificuldade daí decorrente é a articulação consistente do conhecimento definido em níveis tão diversos, especialmente quando a natureza da representação em cada um deles é também diferente.

1.3.5 Modelos estatísticos e regras simbólicas

A representação do conhecimento nos reconhecedores automáticos de fala suscita, naturalmente, questões que são comuns à generalidade dos problemas de reconhecimento de padrões [11]. Em alguns desses problemas os padrões evidenciam uma natureza estatística essencial permitindo, portanto, o processamento da informação, facilmente quantificável, com base em estruturas estatísticas. Pelo contrário, existem outros problemas em que a informação fundamental para o processo de reconhecimento é veiculada através de relações estruturais, dificilmente

[1] Respectivamente, nos chamados reconhecedores *bottom-up* e *top-down*.

[2] Situação frequente durante a década de 70, quando se registou um grande interesse em aplicar técnicas de inteligência artificial ao problema do reconhecimento automático da fala.

quantificáveis mas passíveis de descrição através de estruturas de informação adequadas, entre unidades elementares de informação a que se associam símbolos. Em inúmeros problemas, a fronteira entre estas duas realidades pode ser bastante difusa e ambas as vertentes se podem manifestar, eventualmente em níveis diferentes de abstracção da informação.

É o que sucede no reconhecimento automático da fala contínua onde, por exemplo, podem coexistir modelos estatísticos que representam o conhecimento acústico directamente associado aos vectores de características do sinal, de facto apresentando uma distribuição estatística que pode ser captada, e modelos baseados em regras estruturais, embutidos no sistema para representar o conhecimento lexical[1].

Assim, estas duas abordagens primárias têm vindo a ser consideradas no reconhecimento automático da fala. Relativamente à abordagem simbólica, refiram-se apenas alguns dos sub-sistemas mais relevantes: ao nível mais baixo de informação, as regras acústico-fonéticas; alguns léxicos; os interpretadores semânticos; e os analisadores pragmáticos, no topo da hierarquia da informação. Em relação à outra abordagem, são de referir, pela sua enorme relevância nos actuais sistemas: os modelos acústicos de natureza estatística; e as gramáticas também estatísticas. Os sistemas híbridos tentam combinar estas duas abordagens aproveitando as melhores propriedades de cada uma delas.

Embora a abordagem simbólica seja em geral considerada mais satisfatória do ponto de vista intelectual[2], o certo é que o desempenho dos sistemas que favorecem esta via tem sido, de maneira consistente, pior que o dos sistemas baseados em técnicas estatísticas. Um factor que tem sido apresentado como um dos causadores deste relativo fracasso consiste, simplesmente, na dificuldade que as abordagens simbólicas têm em suportar informação incerta, situação trivial no reconhecimento da fala. Além disso, as técnicas automáticas existentes para definir as regras simbólicas são por norma muito deficientes, uma das razões pelo que se recorre geralmente a peritos humanos para as determinar, situação que por sua vez acarreta outras dificuldades importantes. Este factor é particularmente favorável aos sistemas estatísticos, pois eles dispõem de poderosos algoritmos de aprendizagem ou treino.

No entanto, apesar do relativo fracasso dos sistemas baseados essencialmente em regras simbólicas, é reconhecido o interesse em integrar essas técnicas nos reconhecedores. Por exemplo, aos níveis semântico e pragmático continua a ser dada predominância às técnicas baseadas na lógica simbólica.

---

[1] No entanto, até os léxicos mais simples que apenas definem um conjunto de regras determinísticas são geralmente construídos com base também em considerações de natureza estatística.

[2] Pois existindo de algum modo a noção de que os reconhecedores automáticos emulam essa actividade perceptiva humana, será eventualmente mais reconfortante admitir processos baseados em relações simbólicas do que "apenas" em calculadores numéricos.

1.3.6 Sistema estatístico integrado para reconhecer fala contínua

1.3.6.1 Perspectiva histórica

O reconhecimento de fala contínua é actualmente uma área muito importante da ciência e da tecnologia do reconhecimento automático da fala. No entanto, como foi já referido, os sistemas iniciais estavam limitados a aplicações muito simples baseadas no reconhecimento de palavras pronunciadas isoladamente e pertencentes a um pequeno léxico, constituído no máximo por algumas poucas dezenas de palavras.

Esses primeiros sistemas assentavam na utilização de uma base de representações parametrizadas das palavras do vocabulário e num mecanismo de avaliação e comparação de distâncias nesse espaço de representação[1]. Durante o treino dos sistemas, era seleccionado pelo menos um exemplar para representar cada palavra do vocabulário. Durante o reconhecimento, a cada palavra era atribuída a identidade do exemplar menos distante[2]. A distância entre uma palavra a reconhecer e um qualquer exemplar era calculada somando as distâncias parcelares, medidas segundo determinada métrica definida no espaço de representação acústica, entre os respectivos vectores de características cujas sucessões se encontravam já linearmente alinhadas.

Um progresso muito importante relativamente aos primeiros sistemas consistiu na introdução de uma técnica dinâmica de deformação temporal, conhecida por algoritmo DTW[3], capaz de absorver as distorções temporais típicas do sinal da fala. Assim, permitindo a expansão ou compressão de partes do sinal de modo a conseguir a menor distância entre as sucessões de vectores de características das palavras a reconhecer e dos padrões de referência, foi possível obter um sucesso assinalável[4].

Foram sendo desenvolvidos sistemas mais complexos na tentativa de resolver os problemas colocados por aplicações também cada vez mais exigentes. Embora de forma algo redutora, poder-se-iam definir dois vectores de dificuldade que se manifestaram com agudeza e que exigiram, e continuam ainda a exigir, grande esforço de investigação mas cujos frutos se revelaram determinantes na evolução da tecnologia do reconhecimento automático da fala.

A variabilidade característica do sinal da fala foi uma dessas dificuldades. Tornou-se evidente ao tentar usar os chamados reconhecedores DTW, existentes para um único falante, em aplicações independentes do falante. Verificou-se então uma degradação muito acentuada das taxas de reconhecimento. Inicialmente, foram experimentadas soluções baseadas na utilização de um grande número de exemplares para cada palavra do vocabulário na tentativa de cobrir melhor o espaço acústico de cada palavra. Em princípio, se o número desses exemplares fosse tão grande que permitisse cobrir praticamente todas a realizações possíveis de cada palavra, então os

---

[1] Conhecidos, em Inglês, por *template matching systems*.
[2] Segundo o critério, em Inglês, *nearest neighbour* (1-NN)
[3] Do Inglês, *dynamic time warping*.
[4] Ainda actualmente esta abordagem é utilizada em algumas aplicações simples.

reconhecedores deste tipo poderiam apresentar taxas de reconhecimento elevadas. Porém, o número de exemplares possíveis é necessariamente bastante reduzido devido aos custos computacionais. Por essa razão e também pelo facto de os reconhecedores DTW apresentarem geralmente uma fraca resposta às realizações acústicas que não são observadas durante a fase de treino, isto é, terem uma fraca capacidade de generalização, esta abordagem foi sendo abandonada.

Então, surgiu um processo bastante mais eficiente para absorver as variações acústicas do sinal da fala. Poder-se-ia afirmar, de maneira simplista, que consistia na integração de métodos estatísticos nos reconhecedores DTW existentes. Para cada palavra, o conjunto de exemplares era então substituído por uma distribuição estatística capaz de descrever um sinal que evolui no tempo. No reconhecimento, o penoso processo de comparação do sinal de entrada com inúmeros exemplares era substituído por um processo de avaliação das diferentes distribuições estatísticas, sendo a palavra reconhecida como aquela cujo modelo estatístico obtivesse maior verosimilhança. Assim começavam a ser utilizados os modelos escondidos de Markov na tecnologia do reconhecimento automático da fala. Esses modelos, assim como as inúmeras variantes e generalizações que se seguiram, constituem ainda hoje a tecnologia basilar na maior parte dos sistemas de maior complexidade e que exigem um elevado desempenho. Refira-se que as redes neuronais artificiais, outro paradigma capaz de representar informação acústica presente no sinal da fala, tem vindo também a ser usado com particular sucesso em determinadas aplicações do reconhecimento da fala.

A outra dificuldade estava relacionada com a necessidade de suportar vocabulários cada vez maiores e, além disso, reconhecer sucessões de palavras pronunciadas de forma contínua. Relativamente ao aumento da dimensão dos vocabulários, houve necessidade de recorrer a unidades acústicas elementares ao nível do fonema limitando assim o número de modelos. Foram portanto introduzidos léxicos para definir as pronunciações possíveis para as palavras. Para resolver o problema do reconhecimento de fala contínua houve necessidade de introduzir restrições ao nível sintáctico, limitando a escolha das palavras possíveis em cada ponto da frase. Para o efeito foram introduzidas gramáticas muito restritivas. Apesar das restrições impostas, as taxas de reconhecimento alcançadas mantinham-se baixas, o que em grande parte se devia à modelação bastante deficiente dos padrões acústico-fonéticos, agora apresentando ainda uma maior variabilidade[1].

Estes fracos resultados constituíram um estímulo para que parte significativa do esforço de investigação fosse orientado para a procura de soluções alicerçadas em técnicas de inteligência artificial, por essa altura, uma área em franca expansão. No entanto, as abordagens baseadas em regras simbólicas, apesar de geralmente apresentarem estruturas muito complexas para processar a informação aos diferentes níveis, também não alcançaram resultados interessantes. Foram sendo preteridas e no

[1] Só bastante mais tarde seriam introduzidas as técnicas de modelação acústica com contexto.

essencial apenas se mantiveram activas no processamento de informação aos níveis semântico e pragmático[1].

Foram então surgindo alguns sistemas[2] que apresentavam uma abordagem nova [9] ao problema e que alcançavam níveis de desempenho muito interessantes. No essencial, estes reconhecedores podem ser caracterizados por integrarem numa única estrutura o conhecimento acústico, lexical e sintáctico. Um das vantagens principais é permitir, na fase do reconhecimento, a utilização de algoritmos eficazes, baseados num princípio da pesquisa de grafos, como o algoritmo Viterbi.

Combinando as técnicas de modelação estatística, baseadas nos modelos escondidos de Makov, com técnicas de representação dos conhecimentos acústico, lexical e sintáctico numa mesma estrutura tornou-se então possível enfrentar com relativo sucesso o problema do reconhecimento automático da fala contínua. Apesar da natural evolução desses sistemas, que nesta perspectiva poderão ser designados por sistemas estatísticos integrados, no essencial os seus fundamentos funcionais e estruturais têm-se mantido.

1.3.6.2 Reconhecimento integrado de padrões

Embora a generalidade dos reconhecedores de fala contínua que utilizam uma abordagem baseada em modelos estatísticos e estruturas do conhecimento integradas apresentem características essenciais que são comuns, por outro lado é claro que existem aspectos dependentes das particularidades de cada aplicação[3]. Assim, antes de prosseguir interessa referir que os reconhecedores de fala contínua estudados e desenvolvidos ao longo deste trabalho têm como função efectuar a transcrição ortográfica do sinal da fala, isto é, produzir de maneira automática a sucessão de palavras, pertencentes a um vocabulário definido numa determinada língua, representadas no sinal acústico. Interessa também acrescentar que esses reconhecedores não utilizam conhecimento semântico ou pragmático e portanto não fazem qualquer processamento de informação nesses níveis.

Um sistema integrado de reconhecimento automático de fala contínua executa o reconhecimento de padrões integrando num único processo as acções de classificação que ocorrem a diferentes níveis. No caso dos sistemas baseados em conhecimento acústico, lexical e sintáctico, todas as realizações acústicas das unidades fonéticas, palavras e frases são consideradas como padrões constituídos por sucessões de vectores de características e o reconhecedor executa a classificação desses padrões,

---

[1] Áreas de investigação de que muito se espera mas, por enquanto, não têm conseguido satisfazer essas expectativas.

[2] O sistema HARPY (Lowerre, 1976) é, nesse sentido, frequentemente apontado como uma marco histórico.

[3] Por exemplo, os reconhecedores fonéticos de fala contínua dispensam, naturalmente, informação lexical.

definidos nos diferentes níveis, de maneira integrada. A Figura 1.4 ilustra como nesses sistemas o conhecimento sintáctico, lexical e acústico se pode encontrar representado numa única estrutura de informação.

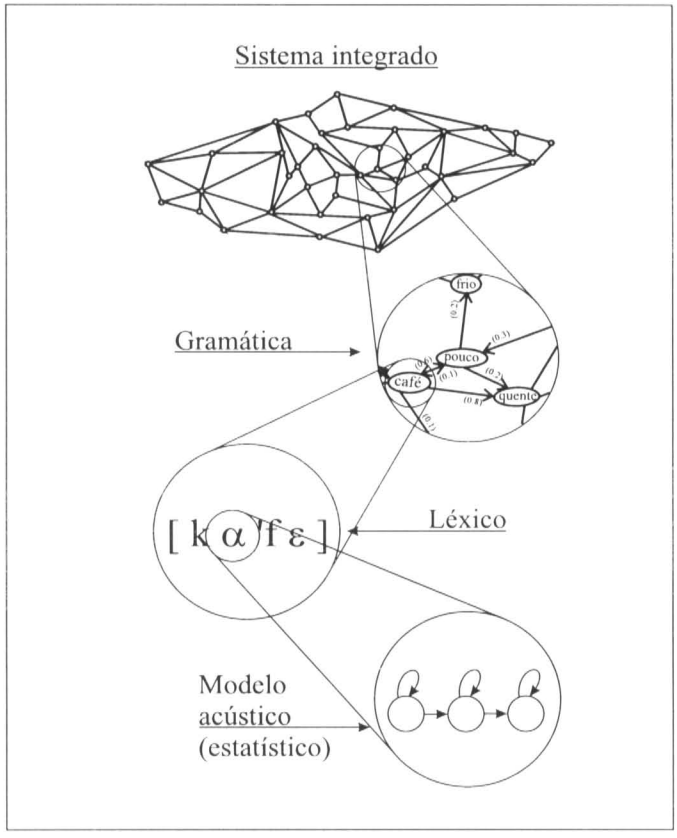

Fig. 1.4 - Estrutura do conhecimento no reconhecedor estatístico integrado

1.3.6.3 Reconhecimento estatístico de padrões

Num sistema de reconhecimento estatístico a natureza dos classificadores operando em diferentes níveis é no essencial estatística. Para além de algumas propriedades já referidas e que se reflectem na supremacia actual destes sistemas, uma das vantagens dos classificadores estatísticos reside no facto de serem suportados por uma base matemática sólida. Isso é patente no conceito de classificação óptima que num enquadramento estatístico pode ser definido com rigor.

Assim, seja $\boldsymbol{X}$ uma sucessão de vectores de características, $\boldsymbol{X}_1^T$ definida em (1.1), extraída do sinal da fala e seja W_c a frase, constituída pela sucessão de palavras W_1^K definida em (1.3), resultante do respectivo processo de classificação. Essa

classificação é determinada pelas funções discriminantes que definem as fronteiras, num espaço abstracto que integra as representações do sinal da fala aos diferentes níveis, entre as classes possíveis.

A questão da classificação óptima pode então ser formulada nos seguintes termos: conhecida uma sucessão de vectores, $\boldsymbol{X}$, pertencente a uma das classes definidas, W_c, $c = 1,2,\ldots,C$, qual é a regra que conduz à probabilidade mínima de errar a classificação?

A resposta a esta questão é: a sucessão de vectores $\boldsymbol{X}$ deve ser atribuída à classe W_c que apresenta, de entre todas, a maior probabilidade *a posteriori*, isto é:

$$P(W_c|\boldsymbol{X}) > P(W_j|\boldsymbol{X}), \quad \forall j=1,\ldots,C, \; j \neq c \qquad \Rightarrow \qquad \boldsymbol{X} \rightarrow W_c \tag{1.4}$$

Um classificador que estabeleça esta regra é conhecido como classificador Bayesiano. Qualquer outro classificador apresenta, em média, maior erro de classificação.

Em geral não é possível modelar directamente as probabilidades *a posteriori* para as diferentes classes. Apenas podem ser estimadas, a partir dos dados e de algum conhecimento apriorístico. Portanto, a correcção dessas estimativas é, em última análise, o factor determinante no desempenho do classificador.

A regra de Bayes permite escrever:

$$P(W_j|\boldsymbol{X}) = \frac{P(\boldsymbol{X}|W_j)P(W_j)}{P(\boldsymbol{X})}, \quad \forall j = 1,\ldots,C \tag{1.5}$$

Assim, sendo $P(\boldsymbol{X})$ constante durante o reconhecimento, a resposta à questão da classificação óptima, dispensando o cálculo directo das probabilidades *a posteriori*, é a seguinte:

$$P(W_c)P(\boldsymbol{X}|W_c) > P(W_j)P(\boldsymbol{X}|W_j), \quad \forall j=1,\ldots,C, \; j \neq c \qquad \Rightarrow \qquad \boldsymbol{X} \rightarrow W_c \tag{1.6}$$

Para cada classe, a função discriminante do classificador passa a ser definida pelo produto de dois modelos estatísticos:

$P(W_c)$, $c = 1,2,\ldots,C$ determina a probabilidade *a priori* de observar a classe, isto é, a probabilidade da frase ocorrer não sendo conhecido o sinal da fala;

$p(\boldsymbol{X}|W_c)$, $c = 1,2,\ldots,C$ determina a probabilidade[1] de observar a sucessão de vectores de características condicionada ao conhecimento da classe, isto é, a verosimilhança do sinal da fala sendo conhecida a frase.

Para além de contornar a impossibilidade prática de construir um modelo que calcule directamente as probabilidades *a posteriori* para cada classe, esta factorização torna possível decompor o problema da modelação estatística em dois sub-problemas:

- a modelação linguística, associada a $P(W)$;
- a modelação acústica, associada a $p(\boldsymbol{X}|W)$.

No entanto, qualquer destes sub-problemas apresenta muitas dificuldades que estão ainda em grande parte por resolver[1].

1.3.6.4 Gramática

A experiência tem mostrado claramente que a introdução de restrições linguísticas no processo de reconhecimento, reduzindo desse modo a perplexidade sintáctica associada à tarefa, é muito importante para alcançar bons resultados.

Nestes sistemas são geralmente usadas gramáticas estocásticas que estabelecem a probabilidade de ocorrência das frases não sendo conhecido o sinal acústico. As chamadas gramáticas regulares são usadas em algumas aplicações, apesar de eventualmente serem demasiado artificiais, por permitirem a fácil criação de um modelo conjunto acústico-linguístico baseado nos modelos escondidos de Markov. Outras gramáticas a destacar, pelo seu reconhecido mérito, são as conhecidas por N-Gram que definem a probabilidade de ocorrência de cada palavra em função das N-1 palavras imediatamente anteriores. Diversos outros modelos linguísticos têm sido propostos, alguns deles capazes de modelar dependências bastante complexas entre as palavras.

Os modelos linguísticos apresentam em geral um elevado número de parâmetros livres que exigem enormes bases de texto para o seu treino, efectuado também por esta razão independentemente do treino dos modelos acústicos.

---

[1] Se $\boldsymbol{X}$ é uma variável aleatória contínua, a verosimilhança é efectivamente uma função densidade de probabilidade, $p(\boldsymbol{X}|W_c)$.

1.3.6.5 Unidades acústicas

A dimensão do vocabulário utilizado numa qualquer tarefa de reconhecimento de fala é um factor de dificuldade muito importante. Aumentando a dimensão do léxico torna-se maior a confusão entre as palavras. Além disso, em sistemas de grande vocabulário[2] não é praticável modelar cada palavra individualmente devido às limitações das bases de dados existentes para treinar os respectivos modelos acústicos e também às restrições do espaço de memória para armazenar esses mesmos modelos.

Os problemas decorrentes dessas limitações são geralmente atenuados aplicando o princípio da partilha dos parâmetros, isto é, o sistema é definido de maneira a que alguns parâmetros possam ser utilizados em diferentes partes. A solução adoptada em grande parte dos sistemas de reconhecimento de fala contínua consiste em utilizar modelos acústicos definidos ao nível fonético, uma escolha natural dado o conhecimento apriorístico relativamente vasto que existe sobre os fonemas. É claro que o aumento dos dados disponíveis para treinar cada modelo tem, como contrapartida, a perda de nitidez desses mesmos modelos uma vez que são treinados com um grande número de exemplos ocorrendo em contextos eventualmente muito diferentes e, portanto, apresentando realizações acústicas significativamente distintas. Assim, é necessário ponderar a escolha das unidades elementares de fala de maneira a que elas possam ser treinadas com robustez mas, por outro lado, sejam sensíveis aos efeitos de co-articulação que se verificam no sinal da fala e que são determinantes para o desempenho dos sistemas.

Nos sistemas mais simples a escolha recai frequentemente nos fonemas modelados sem contexto, permitindo uma grande partilha de parâmetros dado que a generalidade das línguas naturais têm um reduzido número de fonemas, mas por outro lado conduzindo a modelos relativamente difusos.

Sistemas mais evoluídos apresentam habitualmente unidades elementares também ao nível do fonema mas modeladas com contexto, isto é, para cada unidade são treinados diversos modelos que cobrem as variações mais importantes na sua realização acústica devidas ao contexto em que se encontram. Esta variação está em grande parte dependente dos sons que se encontram imediatamente à esquerda e à direita, pelo que é frequente utilizar unidades acústicas designadas tri-fonemas que consideram este contexto. Outras técnicas utilizam modelos de fonemas dependentes da palavra, face à evidência experimental de que a realização acústica de cada unidade está também normalmente dependente da identidade da palavra a que o unidade pertence. Apenas devem ser considerados isoladamente os contextos que são de facto relevantes pois caso contrário o número de unidades acústicas elementares pode tornar-se incomportável.

---

[1] Além disso, tal como será analisado posteriormente, alguns desses problemas estão estreitamente relacionados com esta factorização, como é a deficiente capacidade de discriminação dos modelos treinados com base no critério da maximização das verosimilhanças.

[2] Actualmente, um sistema de reconhecimento automático de fala é considerado de grande vocabulário quando é capaz de suportar pelo menos alguns milhares de palavras.

Tem-se conseguido controlar, com resultados assinaláveis, o número de modelos acústicos utilizando também conhecimento apriorístico. Geralmente utilizam-se regras acústico-fonéticas de substituição, supressão e inserção de fonemas baseadas no contexto. Com o mesmo objectivo, têm sido também utilizadas com sucesso técnicas estatísticas que no essencial fundem segundo um determinado critério de proximidade algumas variantes dos modelos acústicos eventualmente geradas de maneira automática.

1.3.6.6 Léxico

Independentemente dos modelos acústicos elementares que são utilizados, é necessário definir as palavras do vocabulário com base nessas unidades. É essa a função do léxico.

Em muitos sistemas o léxico apresenta apenas uma transcrição fonética[1] por palavra. No entanto, a fala contínua não adere geralmente às formas canónicas de pronunciação. É caracterizada por uma acentuada co-articulação entre as palavras, sobretudo na fala espontânea, para além de existirem outros factores de variabilidade na pronunciação como, por exemplo, os diferentes dialectos. Se essas diferenças não forem consideradas no treino, assim como no reconhecimento, as correspondências entre os modelos acústicos e as observações reais serão deficientes e portanto o desempenho de todo o sistema será negativamente afectado.

Assim, os reconhecedores de fala contínua mais elaborados têm vindo a utilizar algumas técnicas desenvolvidas para acomodar a variabilidade na pronunciação das palavras. Em alguns desses sistemas as diferentes pronunciações são suportadas por modelos estocásticos, treinados de maneira automática, ou através de um conjunto de regras fonológicas extraídas do conhecimento apriorístico.

Tem-se verificado, no entanto, que acrescentar apenas pronunciações alternativas aos léxicos não conduz a uma boa solução, pois a confusão acústica do léxico aumenta e assim os eventuais ganhos podem ser suplantados pelas perdas. Este é mais um dos problemas que está ainda claramente em aberto, não existindo uma solução plenamente satisfatória.

1.3.6.7 Modelos acústicos

Para obter uma boa aproximação ao classificador Bayesiano é fundamental que o modelo acústico global seja capaz de determinar com suficiente correcção a probabilidade de observar uma qualquer sucessão de vectores de características

[1] Admitindo a situação frequente em que as unidades elementares estão definidas ao nível do fonema.

condicionada ao conhecimento da frase, isto é, a verosimilhança global. Esta necessidade de modelar o melhor possível as dependências estatísticas que, ao nível acústico, se verificam no sinal da fala e que são essenciais no processo de reconhecimento é sobretudo evidente quando o modelo não é treinado discriminativamente, ou seja, quando aprende apenas a representar bem as realizações da sua classe mas não aprende a discriminar contra as realizações das outras classes.

O modelo acústico global é criado concatenando, de acordo com a gramática e o léxico do sistema, os modelos acústicos correspondentes às unidades elementares definidas geralmente ao nível fonético. Assim se pode compreender como as restrições impostas ao nível linguístico são também importantes na qualidade da modelação acústica. Apesar dos modelos linguístico e acústico serem por norma treinados isoladamente, devem articular-se de maneira a permitir a correspondência eficaz entre os modelos acústicos e as observações reais, factor essencial no desempenho do sistema. É portanto fundamental que a gramática e o léxico sejam capazes de representar as dependências sintácticas e lexicais que se verificam na realidade.

Por outro lado, é claro que a qualidade dos modelos acústicos, considerados isoladamente, é fundamental para o desempenho do reconhecedor. Actualmente, os modelos escondidos de Markov constituem a tecnologia dominante. Embora de forma algo simplista, pode afirmar-se que isso se deve à sua razoável capacidade em representar distribuições estatísticas que evoluem no tempo, também ao facto de permitirem algoritmos com estrutura bastante simples e obviamente porque em geral obtêm resultados ao melhor nível.

Nos modelos escondidos de Markov a evolução temporal é modelada por uma estrutura, conhecida por cadeia, composta por estados que se encontram ligados, segundo uma determinada topologia, através de transições às quais estão associadas probabilidades. Em cada estado é definida uma função[1], conhecida por densidade de observação, que modela a distribuição estatística de um segmento quase estacionário de sinal da fala. Na maior parte dos sistemas, essa função é definida parametricamente através da combinação linear de funções de Gauss. Podem também ser utilizados modelos não paramétricos, como por exemplo os histogramas. Em conclusão, o processo estocástico associado aos modelos escondidos de Markov pode ser interpretado como uma distribuição estatística variável ao deslocar-se de um estado para outro com determinada probabilidade de transição.

As fraquezas mais significativas que os modelos escondidos de Markov têm revelado na tecnologia do reconhecimento automático da fala estão identificadas e algumas soluções têm sido propostas.

Fazendo por agora apenas uma brevíssima referência a duas dessas deficiências, podem ser destacadas a dificuldade em modelarem correctamente a duração dos segmentos da fala e a correlação existente entre os sucessivos vectores de

[1] É assim nos modelos escondidos de Markov do tipo Moore, nos modelos do tipo Mealy os segmentos do sinal são "observados" nas transições entre estados.

características. Grande esforço tem sido também realizado no sentido de melhorar a resolução dos modelos sem deteriorar a sua robustez, introduzindo topologias e funções densidade de observação mais complexas mas tendo presente o princípio da partilha de parâmetros.

Em resultado desse esforço de investigação têm surgido inúmeras variantes dos modelos escondidos de Markov *standard*. Outros formalismos estatísticos têm também sido desenvolvidos, sendo de realçar os chamados modelos de segmentos[1] que constituem uma poderosa generalização dos modelos escondidos de Markov. São também de destacar soluções híbridas que tentam combinar a habilidade dos modelos escondidos de Markov na modelação de sinais que evoluem no tempo com a capacidade de classificação característica das redes neuronais artificiais.

1.3.6.8 Treino

O treino do sistema consiste no processo de determinação do valor dos parâmetros livres, cujo conjunto é aqui designado θ, com base na informação contida num conjunto de exemplos, T. Cada exemplo é constituído por uma sucessão de vectores de características, $\boldsymbol{X}$, e a correspondente sucessão de palavras, W_c, ou seja:

$$\mathsf{T} = \left\{ \left(\boldsymbol{X}^n, W_c^n \right), n = 1,.., N_{\mathsf{T}} \right\} \tag{1.7}$$

De acordo com a regra de classificação óptima definida pelo classificador Bayesiano, os parâmetros devem ser determinados de maneira a maximizar a probabilidade *a posteriori* estendida ao conjunto de treino:

$$\theta = \arg\max_{\theta} \prod_{n=1}^{N_{\mathsf{T}}} P_{\theta}\left(W_c^n | \boldsymbol{X}^n \right) \tag{1.8}$$

Frequentemente aceita-se uma solução aproximada[2] que permite reduzir muito a complexidade do procedimento de treino. Assim, embora seja reconhecida a consequente perda de capacidade discriminativa do sistema, os parâmetros passam a ser determinados de acordo com o critério de maximização da probabilidade conjunta de $\boldsymbol{X}$ e de W_c estendida ao conjunto de treino:

$$\theta = \arg\max_{\theta} \prod_{n=1}^{N_{\mathsf{T}}} P_{\theta}\left(\boldsymbol{X}^n | W_c^n \right) P_{\theta}\left(W_c^n \right) \tag{1.9}$$

[1] Ou, em Inglês, *segmental models*.
[2] Posteriormente, esta questão será analisada mais detalhadamente.

Além desta aproximação, o modelo linguístico é em geral treinado independentemente do modelo acústico, utilizando para o efeito enormes bases de dados de texto.

A utilização dos modelos escondidos de Markov na modelação acústica apresenta a grande vantagem do treino dispensar a tarefa árdua de segmentar e anotar a base de dados acústicos completa, sendo suficiente iniciar isoladamente os modelos acústicos com alguns exemplos de treino. Então, um procedimento iterativo baseado no algoritmo Baum-Welch ou no algoritmo Viterbi utiliza a base de dados completa para treinar embebidamente os modelos.

Em alguns sistemas, posteriormente a este procedimento de treino e por vezes simultaneamente à execução das respectivas tarefas de reconhecimento, os parâmetros são novamente adaptados[1] utilizando um conjunto de dados reduzido. O objectivo é essencialmente a redução das diferenças existentes entre as características acústicas que efectivamente se verificam durante a operação do reconhecedor e aquelas que o conjunto utilizado no treino do sistema apresenta.

1.3.6.9 Reconhecimento

O reconhecimento, ou classificação, de uma sucessão de vectores de características, $\boldsymbol{X}$, consiste no processo de determinação da sucessão de palavras, W_c, correspondente. De acordo com a regra de classificação definida em (1.6), relativa ao classificador Bayesiano, $\boldsymbol{X}$ é atribuído à classe W_c que, de entre todas, apresenta o valor mais elevado de $P(\boldsymbol{X}|W_c)P(W_c)$.

O problema consiste, então, em pesquisar o espaço das frases possíveis de maneira a encontrar aquela que permite a maior probabilidade conjunta $P(\boldsymbol{X},W_c)$. Uma busca exaustiva é apenas praticável em problemas de pequeno vocabulário. Por outro lado, a estrutura integrada do sistema permite a utilização de algoritmos bastante eficientes de pesquisa, como o algoritmo Viterbi, baseados em técnicas de programação dinâmica. Uma técnica usual por ser extremamente simples e ao mesmo tempo ser capaz de reduzir muito o espaço de busca sem deteriorar significativamente as taxas de reconhecimento é a chamada pesquisa em feixe[2]. Inúmeras outras técnicas mais elaboradas têm sido apresentadas com o objectivo de acelerar o processo de reconhecimento. Por exemplo, alguns sistemas utilizam um processo de reconhecimento em dois, ou mais, passos de maneira que é feita uma descodificação inicial relativamente rápida num espaço de pesquisa mais limitado. A lista de hipóteses resultante é então, no segundo passo, pesquisada utilizando a representação acústica completa com o objectivo de seleccionar a melhor frase.

---

[1] Este procedimento é conhecido, em Inglês, por *(on-line) adaptive learning*.

[2] Do Inglês, *beam-search*.

1.3.6.10 O *estado da arte*

Os actuais reconhecedores de fala contínua que apresentam melhores taxas de reconhecimento, em aplicações com léxicos até cerca de 60.000 palavras e são independentes do falante, utilizam arquitecturas integradas de conhecimento e baseiam a decisão em formalismos estatísticos.

A representação do sinal acústico é geralmente obtida através de um procedimento baseado na análise espectral do sinal, sendo associadas determinadas características dinâmicas.

O modelo acústico é baseado na tecnologia dos modelos escondidos de Markov. No entanto, alguns sistemas utilizam uma abordagem híbrida baseada nos modelos escondidos de Markov e em redes neuronais artificiais na modelação das densidades de observação. Os modelos acústicos elementares são geralmente definidos ao nível fonético e são modelados atendendo ao contexto determinado pelos sons vizinhos[1]. Curiosamente, nestes sistemas de grande complexidade tem-se verificado que as topologias mais simples dos modelos escondidos de Markov, frequentemente do tipo Bakis e com apenas três estados, são as que conduzem aos melhores resultados.

No modelo linguístico são utilizadas frequentemente gramáticas do tipo 3-Gram, ou até mesmo gramáticas capazes de modelar dependências entre palavras mais distantes. O léxico pode ser composto por modelos estocásticos e utilizar também regras fonológicas no sentido de modelar a variabilidade na pronunciação das palavras.

É importante referir que enormes bases de dados, acústicos e de texto, são utilizados para treinar estes sistemas.

[1] Em Inglês, *tri-phones* (ou até mesmo *quadri-phones*, em alguns sistemas).

Capítulo 2

Sistema básico de reconhecimento de fala contínua

Neste capítulo é apresentado o sistema de reconhecimento automático da fala que foi desenvolvido para servir de base a este trabalho. Inicialmente é feita uma breve caracterização do sistema. Em seguida são descritas as opções tomadas, os procedimentos utilizados no seu desenvolvimento e os fundamentos considerados essenciais, em particular a teoria dos modelos escondidos de Markov. Finalmente, são apresentados resultados obtidos com o treino e teste em bases de dados *standard*.

2.1 Introdução

Com a construção deste sistema base pretendeu-se obter um reconhecedor automático de fala contínua, independente do falante e com um vocabulário de média dimensão, que estabelecesse uma referência a utilizar ao longo do trabalho.

Um outro objectivo importante consistiu, naturalmente, na aquisição de conhecimentos sobre a utilização de técnicas correntes em sistemas *standard* para aplicações do mesmo tipo. Ao longo do desenvolvimento do sistema foram efectuadas muitas experiências confrontando técnicas alternativas ou ajustando parâmetros que permitissem melhorar o seu desempenho relativamente às taxas de reconhecimento, à velocidade de processamento e ao espaço de memória.

O sistema é baseado na tecnologia dos modelos escondidos de Markov e utiliza unidades acústicas elementares, modeladas sem contexto, ao nível do fonema.

As opções gerais tomadas relativamente a este sistema inicial foram ponderadas tendo presente os objectivos essenciais já referidos, em particular a definição de um ponto de partida interessante para o resto do trabalho, e também procurando um compromisso razoável entre o seu nível de complexidade e o seu desempenho.

Um dos exemplos mais significativos de restrição da complexidade neste sistema é a não utilização de modelação acústica com contexto. Apesar de ser uma técnica cujo mérito é reconhecido, foi dispensada por duas razões principais. Em primeiro lugar porque, pretendendo-se essencialmente com este trabalho fazer uma avaliação

comparativa dos sistemas que usam múltiplas características do sinal de entrada relativamente aos sistemas baseados apenas numa análise *standard* do sinal acústico, constitui um problema por si só muito interessante fazer esse estudo para sistemas com modelação acústica sem contexto. Além disso, existe a forte convicção de que grande parte das conclusões e conhecimento em geral extraídas destes sistemas usando modelos acústicos sem contexto são extensíveis aos que utilizam explicitamente essa informação, embora algumas outras possam não se verificar pelo menos relativamente a algumas situações em particular. Em segundo lugar, a redução criteriosa da complexidade dos sistemas foi decisiva na rentabilização do esforço de investigação, o que obviamente abrange os recursos informáticos disponíveis. Esse aspecto é claramente sentido quando, por exemplo, é necessário esperar centenas de horas para treinar um sistema, mas é-o ainda mais quando é necessário escrever os programas para o desenvolver. É importante salientar, a este respeito, que os programas que suportam este sistema, assim como as inúmeras experiências realizadas, foram integralmente desenvolvidos "de raiz" sem recorrer a ferramentas de investigação específicas para reconhecimento da fala[1]. Foi apenas utilizado um compilador *standard* de ANSI C e o pacote de *software* Matlab.

Apesar dos aspectos referidos, e sendo certo que é muito mais fácil melhorar um sistema relativamente deficiente, pretendeu-se obter um reconhecedor capaz de um desempenho aceitável numa tarefa *standard*. Esse objectivo foi claramente alcançado, o sistema foi treinado com a base de dados *Resource Managment* e o teste conhecido por *February'89* realizado com uma simples gramática *word-pair* atingiu uma taxa de erro de palavra de 6,48%.

2.2 Tarefa de reconhecimento

Este sistema foi desenvolvido para o reconhecimento de fala contínua numa aplicação com um vocabulário de média dimensão e capaz de operar com múltiplos falantes. A tarefa que se apresenta pode ser formulada nos termos que se seguem.

Seja o sinal de fala que se pretende reconhecer, já amostrado e quantificado, definido portanto num domínio discreto e apresentando também valores discretos, representado por:

$$s(k), \qquad k=1, 2, \ldots \tag{2.1}$$

e correspondente à realização acústica de uma sucessão de palavras:

[1] Por exemplo, o *HTK Toolkit*

$$W_1^K = \{\omega_1, \omega_2, \ldots, \omega_K\} \tag{2.2}$$

válida nesta tarefa.

Seja L o modelo linguístico definido para esta tarefa. Ao nível sintáctico é utilizada uma gramática *word-pair*, estabelecendo as palavras que podem suceder a uma qualquer pertencente também ao vocabulário, Γ_ω, com 991 palavras.

Seja M o modelo associado à realização acústica das frases nesta tarefa de reconhecimento. Este modelo é previamente apreendido pelo sistema através de um procedimento de treino utilizando uma base de dados relativamente à qual se deve salientar: 1) as frases resultam da leitura da respectiva representação ortográfica, efectuada de maneira "normal", nomeadamente, sem pausas forçadas entre as palavras; 2) o sinal da fala é transmitido num ambiente acusticamente favorável, com ausência de ruído ou distorção significativos, o mesmo se verificando em relação ao sistema de captação do sinal.

Seja, ainda, verificada a seguinte condição durante o reconhecimento: a frase é pronunciada por um falante não necessariamente conhecido do sistema, ou seja, possivelmente não interveniente no treino dos parâmetros livres.

Então, o objectivo definido para esta tarefa consiste em determinar a sucessão de palavras correcta, isto é, aquela que corresponde ao sinal acústico de entrada:

$$s(k) \xrightarrow[L,\,M]{} W_1^K \tag{2.3}$$

2.3 Bases de dados

O desenvolvimento do sistema foi feito utilizando a base de dados *Resource Managment* [58], a qual foi desenhada para uma tarefa consistindo no reconhecimento de instruções de comando e pesquisa de informação numa base de dados naval. A língua utilizada nesta base de dados é o Inglês, tal como é falado em diferentes regiões dos EUA. Dois exemplos típicos de frases são: *DRAW TRACKS FOR ALL USN CRUISERS*; *WHAT IS THE DATE AND HOUR OF ARRIVAL IN PORT FOR ARKANSAS*.

Uma vantagem imediata associada à utilização desta base de dados resulta do facto dela constituir uma referência bem conhecida, tendo servido de suporte ao desenvolvimento de muitos sistemas ao longo de pouco mais de uma década. Embora

essa longevidade acarrete algumas limitações que podem ser importantes no desenvolvimento de alguns sistemas, por exemplo no que se refere à relativa exiguidade do vocabulário, a sua utilização neste trabalho não apresenta dificuldades significativas.

O vocabulário desta base de dados é constituído por 991 palavras cujo léxico é bastante confuso devido à existência de muitas palavras com realização acústica semelhante. Exemplos dessa dificuldade [59] verificam-se entre as palavras *WHAT*, *WHAT'S* e *WAS* ou *THE* e *A* ou ainda *FOUR* e *FOURTH*, entre muitas outras. Grande parte dos substantivos existem no singular, no plural e também na forma possessiva, como por exemplo *SUBMARINE*, *SUBMARINES* e *SUBMARINE'S*.

Outra dificuldade, particularmente importante, reside na ocorrência frequente de palavras tais como *A*, *AND*, *OF*, *THE*, *IN* e *TO*. Geralmente estas palavras são articuladas de maneira bastante deficiente, muito dependente do contexto[1] e a sua pronunciação é débil, sendo até com alguma frequência eliminadas [60]. Além disso, durante o processo de reconhecimento, sucede com alguma frequência (§2.7.2) não ser possível, através da gramática, eliminar a ambiguidade que lhes está associada.

Ao nível gramatical, para modelar um total de 900 frases diferentes existentes nesta tarefa estão também associadas à base de dados diversas gramáticas *standard*. A gramática utilizada no desenvolvimento deste sistema consiste apenas num conjunto de regras que estabelecem as palavras que podem suceder a qualquer palavra do vocabulário e impõe um grau de perplexidade (§2.7.2) estimado em aproximadamente 60 palavras por ramificação. Apesar deste valor definir uma tarefa que ao nível gramatical não se pode considerar particularmente difícil, refira-se que muitos sistemas utilizando esta, e outras, base de dados são desenvolvidos com gramáticas bastante mais restritivas.

As frases que constituem esta base de dados são lidas pelo falante e não resultam de um processo espontâneo de geração de fala. Refira-se, no entanto, que no âmbito deste trabalho esta característica não constitui uma eventual limitação a considerar.

Um aspecto que interessa também referir é a elevada relação sinal-ruído e a ausência de distorção significativa do sinal, em resultado da qualidade acústica ambiente e dos sistemas de captação e registo.

O sinal está amostrado à frequência 16 KHz e encontra-se quantificado com 16 bits por amostra.

No desenvolvimento do sistema foi utilizada parte da *Resource Managment* dedicada a sistemas independentes do falante, consistindo em aproximadamente 2½ horas de sinal de fala contendo 2880 frases de treino, correspondentes a 72 falantes cada qual pronunciando 40 frases.

---

[1] Assim, em alguns sistemas [59] estas palavras são modeladas com particular atenção.

Na avaliação do sistema foi utilizado um conjunto de teste, conhecido por *February'89*, considerado particularmente difícil e consistindo em 300 frases pronunciadas por 10 falantes, nenhum deles pertencente ao conjunto de treino.

2.4 Estrutura do reconhecedor

Este reconhecedor apresenta uma estrutura global similar à de grande parte dos sistemas estatísticos integrados utilizados em aplicações de fala contínua e vocabulário com média ou grande dimensão.

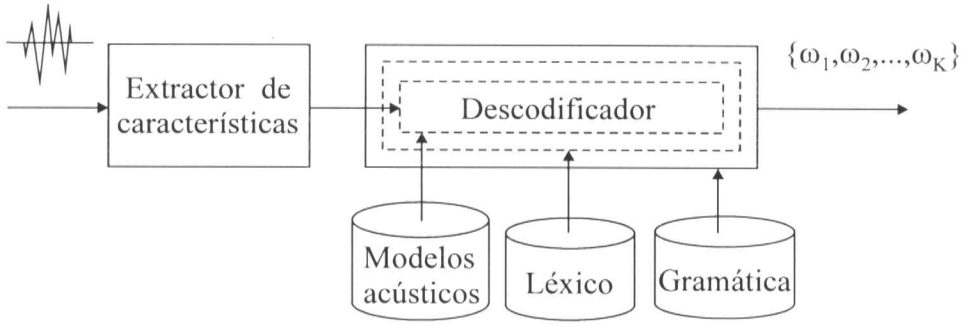

Fig. 2.1 - Diagrama de blocos do reconhecedor

O módulo extractor de características converte o sinal acústico numa sucessão de vectores de características obtidos com base na análise cepstral do sinal.

O módulo descodificador é baseado na pesquisa de uma estrutura integrada de informação, operação realizada pelo algoritmo Viterbi, de que resulta a sucessão de palavras que compõem a frase reconhecida. Essa estrutura é definida por uma gramática do tipo *word-pair*, um léxico de pronunciação única e, ao nível acústico, por modelos escondidos de Markov do tipo semi-contínuo que modelam unidades elementares do tipo fonético.

2.5 Análise do sinal

2.5.1 Introdução

Sendo o processo de decisão que ocorre num reconhecedor automático da fala efectuado com base nos dados fornecidos pelo extractor de características, é óbvia a importância deste módulo.

Por princípio, pretende-se que a análise do sinal acústico conduza a uma representação que preserve toda a informação necessária para o processo de classificação mas seja insensível a informação não relevante. Esse objectivo é também importante ao favorecer uma acentuada redução da dimensão do espaço de representação acústica.

Têm sido propostas muitas técnicas de análise que abrem então espaço a inúmeras variantes cujos méritos dependem, naturalmente, da natureza da aplicação e também das características dos módulos subsequentes do reconhecedor.

Fazendo uma rápida consulta por algumas publicações sobre este assunto verifica-se que a larga maioria das técnicas de análise assenta na extracção de determinadas características a partir de sectores do sinal com pequena duração e onde se regista a sua quase estacionaridade. No entanto, alguns sistemas [42] utilizam uma abordagem que aponta numa direcção radicalmente diferente, isto é, o sinal da fala é modelado como uma sucessão de transições, designados "eventos", separados por períodos de relativa estacionaridade. Isso decorre do conhecimento apriorístico segundo o qual as partes mais ricas do sinal da fala e também as mais robustas em condições adversas, do ponto de vista da percepção humana, correspondem às transições entre os períodos de quase estacionaridade [43].

Nos sistemas baseados na modelação de características quase estacionárias, em segmentos do sinal, têm vindo a ser investigadas técnicas que utilizam janelas de análise com duração muito longa, da ordem das décimas de segundo. De acordo com experiências realizadas na área da psico-acústica, o sistema auditivo humano recorre também à análise integrada do sinal sobre períodos da ordem dos 200 ms [44]. Esta questão[1] é particularmente interessante na perspectiva deste trabalho, na medida em que características capazes de representar dependências estatísticas de longo alcance podem ser simultaneamente utilizadas com as representações *standard* do sinal acústico [45].

Na classe dos métodos de análise ditos "tradicionais" podem ser destacados dois grupos [46]: os baseados no espectro de predição linear associado ao modelo de produção acústica da fala, por exemplo os coeficientes de predição linear (LPC); e os baseados no espectro de Fourier do sinal da fala, dos quais interessa realçar os coeficientes Mel-cepstrais (MFCC) (§2.5.2.4). É curioso verificar que em bastantes sistemas os MFCC são calculados a partir dos LPC.

Neste sistema são utilizados os coeficientes Mel-cepstrais, calculados a partir da representação espectral de Fourier do sinal à saída de um banco de filtros. Esta opção foi tomada em grande medida devido aos resultados existentes que atestam a consistência e desempenho assinaláveis destas características em sistemas e aplicações similares a esta. Além disso, não exigem uma elevada complexidade computacional e permitem uma representação bastante compacta.

---

[1] Tal como a questão relacionada com a modelação das transições (referida no parágrafo anterior).

Para além da escala Mel, em alguns sistemas, designados genericamente por modelos de auditório, são utilizadas técnicas que emulam outras propriedades conhecidas da percepção auditiva humana [47,48] e que se revelam úteis[1] nos reconhecedores automáticos.

Na generalidade dos sistemas construídos para operar em condições não-laboratoriais a robustez das características é uma questão de enorme importância [51]. Nesse sentido, têm vindo a ser desenvolvidas características mais eficazes no processamento de sinal da fala contaminado com diferentes tipos de ruído ou distorção associadas ao canal de transmissão [50,52]. Em alguns casos as representações existentes são transformadas, como sucede com uma técnica extremamente simples e bastante eficaz na normalização do canal, aplicada aos coeficientes Mel-cepstrais e consistindo apenas na subtracção, a cada coeficiente, do respectivo valor médio [54]. Em alguns sistemas são utilizadas soluções que abrangem simultaneamente os modelos acústicos, como por exemplo a normalização do canal acabada de referir integrada numa técnica conhecida por *parallel model combination* [53]. No âmbito deste trabalho a questão da robustez das características não foi praticamente considerada[2], tal como sucede em muitos sistemas desenvolvidos em ambiente laboratorial.

Frequentemente, na parte final do módulo de análise as características existentes são transformadas de maneira a que a informação linguística que lhes está associada apresente uma maior capacidade de discriminação no módulo de descodificação [37,57]. Por exemplo, os modelos escondidos de Markov utilizados em alguns sistemas apresentam graves limitações na modelação da correlação existente entre os sucessivos vectores de características, pelo que nesta fase podem ser usadas técnicas, algumas delas relativamente simples, para atenuar este problema [41]. Outros exemplos de transformações frequentemente utilizadas são a redução de dimensão da representação [40,56] e a descorrelação [36] entre os elementos do vector de características. A redução da dimensão do vector de características pode ser uma operação essencial, sobretudo em aplicações em tempo real, ao tornar menos complexa e portanto mais expedita a tarefa do reconhecimento, ou então, por exemplo, permitindo a utilização de técnicas mais sofisticadas nos módulos subsequentes.

Antes de terminar acrescente-se apenas que a avaliação da representação do sinal da fala é geralmente feita integrando o respectivo módulo de extracção de características num sistema de reconhecimento e comparando os resultados, para uma determinada tarefa, com os obtidos por outros módulos de análise utilizados como referência. Um procedimento alternativo [55] consiste em utilizar métodos que medem, segundo um determinado critério, a separação entre as classes no espaço de características, fornecendo assim um indicador da capacidade discriminativa dessa representação.

---

[1] Como é previsível, a existência de uma operação realizada pelo sistema auditorial humano não é condição suficiente para a sua "utilidade" num reconhecedor automático.

[2] A base de dados utilizada no desenvolvimento do sistema apresenta uma elevada relação sinal-ruído e não se verificam assinaláveis "diferenças de canal" entre os conjuntos de treino e de teste.

2.5.2 O módulo extractor de características

2.5.2.1 Estrutura e resultado

Este módulo de análise, cujo diagrama de blocos é apresentado na Figura 2.2, considera que o sinal da fala já se encontra amostrado e quantificado, tal como está disponível na base de dados *Resource Managment*.

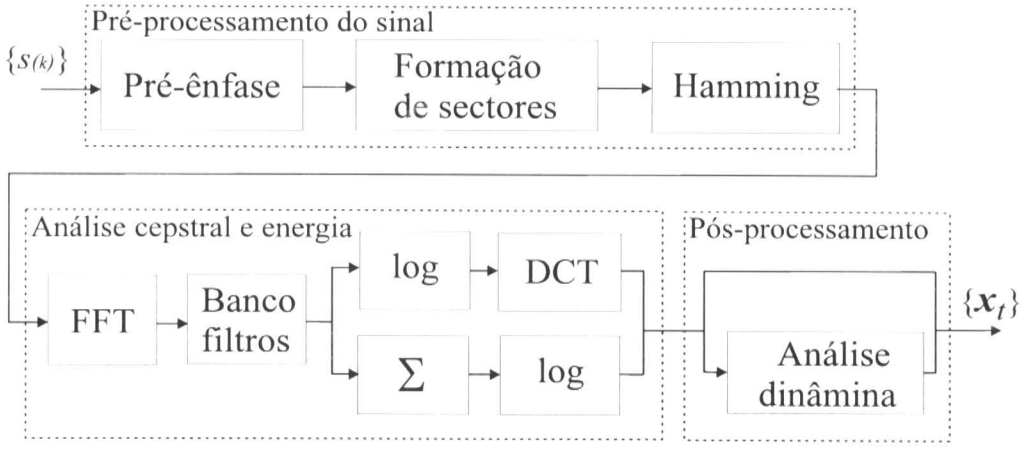

Fig. 2.2 - Diagrama de blocos do extractor de características

Na parte inicial do módulo de análise, aqui referida por pré-processamento, o sinal é preparado para a extracção dos coeficientes Mel-cepstrais e termo de energia: é realizada a operação de pré-ênfase e são formados os sectores a que vão corresponder os vectores de características.

Em seguida, para cada sector do sinal são extraídos os respectivos coeficientes Mel-cepstrais utilizando um método baseado num banco de filtros. A energia correspondente a cada sector é também obtida através da análise espectral.

Na parte final são simplesmente acrescentados ao vector de características alguns termos que explicitam informação local sobre a respectiva trajectória neste espaço de representação acústica.

À saída deste módulo está então disponível uma representação do sinal acústico que se apresenta à entrada, consistindo numa sucessão de vectores de características definidos num espaço 39-dimensional e regularmente intervalados de 10 ms:

$$s(k) \rightarrow \boldsymbol{X}_1^T \tag{2.4}$$
$$\boldsymbol{X}_1^T = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_T\}$$
$$\boldsymbol{x}_i \in \Re^{39}$$
$$t_{real}\Big|_{\boldsymbol{x}_{i+1}} - t_{real}\Big|_{\boldsymbol{x}_i} = 10ms, \quad \forall_i$$

2.5.2.2 Pré-ênfase do sinal

Logo após a amostragem e quantificação, o sinal da fala é submetido a um filtro linear de primeira ordem com o objectivo de nivelar a representação espectral do sinal[1]. A função de transferência desse filtro é:

$$H(z) = 1 - \alpha z^{-1}, \qquad \alpha = 0{,}95 \tag{2.5}$$

O valor do parâmetro α encontra-se tipicamente no intervalo aproximado [0,91 ; 0,99], verificando-se experimentalmente que neste intervalo o seu ajuste preciso não é crítico pois não se altera significativamente o desempenho do sistema.

A evidência experimental mostra também que a utilização deste filtro melhora de maneira consistente o desempenho dos reconhecedores. Com este sistema foi realizada uma pequena experiência para comparar a taxa do erro de palavra (*WER*) aplicando o filtro de pre-ênfase num caso e dispensando-o no outro. Utilizando os modelos obtidos após uma iteração de treino, sem pré-ênfase o *WER* alcançado foi igual a 21,24% e com α igual a 0,95 o *WER* foi 17,95%. Tendo em ambos os casos os modelos sido iniciados e treinados por um procedimento idêntico, é significativa esta redução do erro em aproximadamente 15 $\%_{rel}$[2].

Como evidencia a figura 2.3, mostrando a energia média em cada filtro passa-banda pertencente ao banco de filtros usado no módulo de análise, definido de acordo com a escala Mel, o filtro de pré-ênfase aumenta relativamente a energia nas bandas espectrais de maior frequência. A energia média[3] foi estimada com base em aproximadamente 25000 vectores de características correspondentes a uma amostra aleatória de 100 frases pertencentes ao conjunto de treino.

---

[1] Em alguns sistemas que utilizam os LPC, o filtro de pré-ênfase tem também como função atenuar eventuais problemas de instabilidade numérica.

[2] A redução percentual relativa é calculada da seguinte maneira: $\%_{rel} = 100\left(1 - \frac{\%_{final}}{\%_{inicial}}\right)$

[3] A estimativa da log-energia média, correspondente à média geométrica da energia, conduziria a conclusões semelhantes.

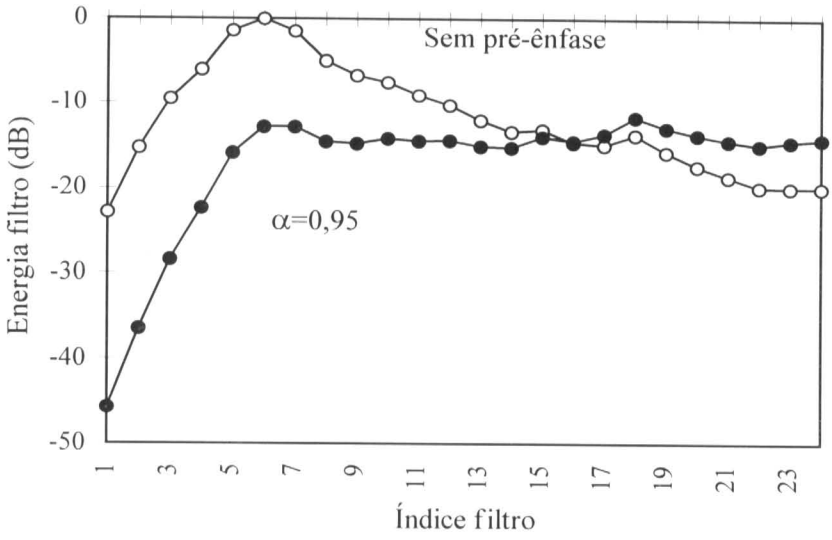

Fig. 2.3 - Estimativa da energia média no banco de filtros.

É interessante fazer aqui uma brevíssima referência ao modelo de produção acústica do sinal da fala [6] cujo diagrama de blocos é apresentado na figura 2.4. Neste modelo básico são distinguidas duas excitações de natureza diferente. Uma das funções de excitação, relativa à produção de sons vocálicos, é periódica e consiste num trem de impulsos que são então moldados por um filtro que modela a glote. A outra função de excitação, correspondente aos sons não vocálicos, consiste num gerador de ruído cuja distribuição espectral é constante. Em ambos os casos, o sinal da fala resulta da convolução do sinal obtido pela fonte de excitação com a resposta impulsional do filtro correspondente ao tracto vocal e radiação acústica nos lábios.

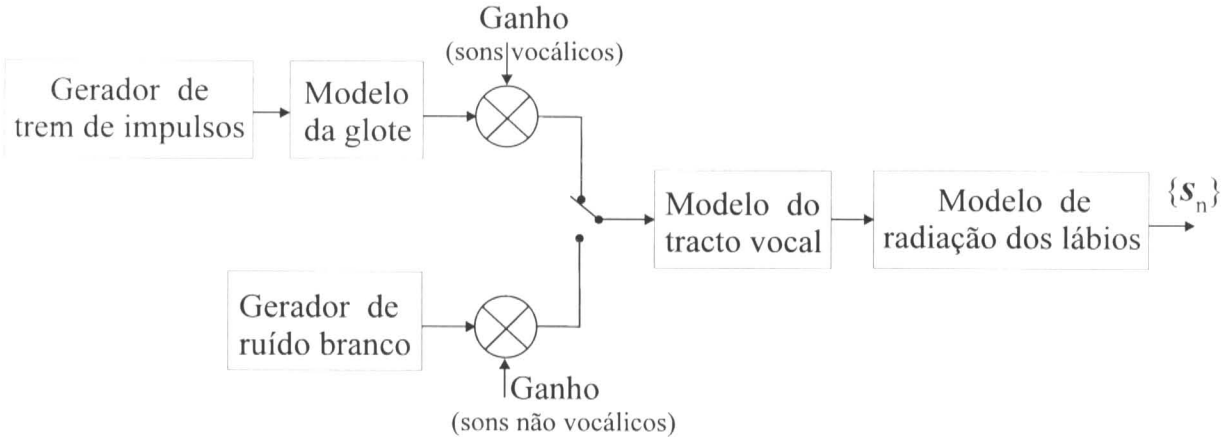

Fig. 2.4 - Modelo de produção acústica da fala.

A produção dos sons vocalizados à saída do tracto vocal apresenta uma resposta em frequência que regista uma queda de 12dB/oitava. Essa resposta é parcialmente compensada pela impedância de radiação dos lábios que apresenta um efeito de filtragem passa-alto estimado em aproximadamente 6dB/oitava. Assim, a resposta em frequência resultante para o modelo correspondente aos sons vocalizados é de aproximadamente -6dB/oitava. A figura 2.5 mostra como essa resposta é compensada pelo filtro de pré-ênfase. A energia média entre os filtros passa-banda 5 e 18, correspondente às vogais e calculada a partir duma segmentação automática em classes "grossas"[1] realizada sobre os cerca de 25000 vectores de características já referidos, é relativamente independente da frequência central do filtro.

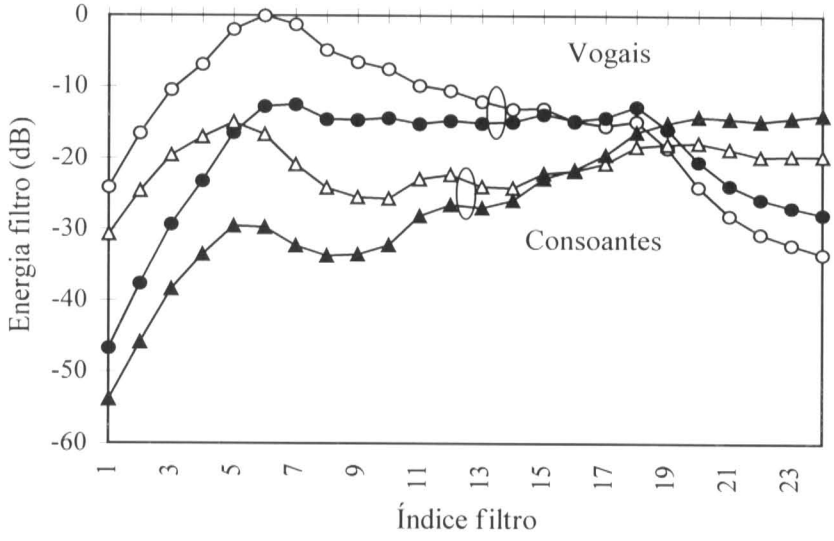

Fig. 2.5 - Estimativa da energia média no banco de filtros considerando independentemente os sons vogais e consoantes (os símbolos a cheio correspondem à aplicação de pré-ênfase).

Por outro lado, relativamente aos sons não vocalizados, como são as consoantes, esta compensação acentua a já natural preponderância das componentes espectrais de mais elevada frequência, resultado que é também claro na figura 2.5. Portanto, o filtro de pré-ênfase não deveria ser aplicado aos sons não vocálicos [6]. No entanto, isso obrigaria a efectuar primeiro uma classificação do sinal em sons vocálicos e sons não vocálicos, razão pela qual neste sistema, e tal como sucede em geral, o filtro de pré-ênfase é permanentemente utilizado.

[1] Foram consideradas três classes de sons: as vogais, as consoantes e outros sons.

2.5.2.3 Formação de sectores e aplicação da janela Hamming

A sucessão de vectores de características é determinada analisando uma sucessão respectiva de sectores do sinal, a qual é portanto necessário definir.

Seja $s(k)$ o sinal à saída do filtro de pré-ênfase e seja $h(k)$ uma função também de variável discreta, conhecida por janela Hamming:

$$h(k) = 0{,}54 - 0{,}46\cos\left(2\pi\frac{k-1}{L-1}\right), \quad 1 \le k \le L \tag{2.6}$$

Então, a sucessão de sectores $\{s_n(k), \quad n = 1,2,\ldots, \quad k = 1,2,\ldots,L\}$ é definida por:

$$\{s_n(k) = s(k + nD + \Delta)\, h(k), \quad n = 1,2,\ldots, \quad k = 1,2,\ldots,L\} \tag{2.7}$$

Neste sistema, o valor dos parâmetros L e D é, respectivamente, 400 e 160. Assim, a sucessão de sectores pode ser entendida como o resultado de observações do sinal em intervalos regulares de 10 ms[1], sendo cada observação realizada através de uma janela Hamming com a duração de 25 ms. Cada sector encontra-se, portanto, sobreposto durante 15 ms ao sector imediatamente anterior.

A duração de cada sector tem um valor que é típico em sistemas baseados na extracção de características através da análise espectral do sinal, que se supõe quase estacionário por partes, isto é, quase estacionário durante determinados períodos.

Relativamente a esses períodos de quase estacionaridade, que desde logo impõem restrições à duração de cada sector, de acordo com alguns resultados [29, 30] para obter um sinal de fala inteligível não é necessário modelar o movimento dos articuladores do tracto vocal humano com taxas superiores a 16 Hz.

A duração de cada sector é definida de modo a obter um compromisso razoável entre a resolução nas representações temporal e espectral do sinal. Refira-se, a este propósito, que uma característica conhecida do sistema auditivo humano, e aproveitada em alguns sistemas automáticos de reconhecimento de fala [28], consiste na dependência da sua resolução temporal relativamente à banda de frequência do sinal acústico.

De acordo com uma propriedade da transformada de Fourier, a transformada do produto de $s_n^*(k) = s(k + nD + \Delta)$ por uma janela $h(k)$:

[1] Relembre-se que o valor da frequência de amostragem do sinal é 16 KHz.

$$S_n(e^{j\omega}) = \sum_{m=-\infty}^{+\infty} s_n^*(m)h(m)e^{-j\omega m} \tag{2.8}$$

é igual ao resultado da convolução das transformadas de cada uma dessas sucessões.

$$S_n(e^{j\omega}) = S_n^*(e^{j\omega}) \otimes H(e^{j\omega}) \tag{2.9}$$

Logo, o espectro $S_n(e^{j\omega})$ extraído de um sector do sinal aproxima-se mais da efectiva representação espectral do sinal nesse sector, $S_n^*(e^{j\omega})$, uma vez mais supostamente quase-estacionário, se a transformada de Fourier da janela, $H(e^{j\omega})$, se aproximar de um impulso, isto é, se a janela tiver uma maior duração.

Ainda relativamente ao estabelecimento de um minorante para a duração dos sectores, refira-se que se pretende posteriormente[1] integrar neste módulo de análise um estimador do período do harmónico fundamental da excitação associada aos sons vocálicos. Para extrair essa informação através da análise espectral, esse sistema necessita que a janela tenha uma duração maior do que o período dessa excitação, limite que em geral é condicionado pelas vozes masculinas que exigem um valor para a duração da janela de pelo menos 10 a 20 ms.

Por outro lado, é óbvio que só uma janela relativamente breve permite uma resolução temporal aceitável. O facto de apenas durante segmentos relativamente breves do sinal da fala se manterem quase inalteradas as suas características espectrais, estabelece à partida um limite superior para a duração da janela.

A comparação de espectrogramas do sinal acústico obtidos através de janelas com diferente duração [27] evidencia alguns dos aspectos acabados de referir. Usando janelas com a duração relativamente curta, perde-se a representação dos harmónicos de excitação nos sons vocálicos e regista-se também o alisamento do espectro correspondente aos outros sons. Em contrapartida, a resolução temporal das *formantes* é notável. Usando uma janela com duração significativamente maior registam-se claramente os harmónicos da excitação e aumenta a definição espectral do ruído, mas é perdida a resolução temporal.

Um outro aspecto a considerar está relacionado com a forma da janela aplicada a cada sector. Neste sistema é utilizada a janela Hamming, definida na equação 2.6, ao invés de simplesmente usar a janela rectangular, com o objectivo de reduzir o fenómeno de Gibbs [26], segundo o qual o truncamento abrupto do sinal num dos domínios de representação origina oscilações no outro domínio. Comparando a largura de banda das representações espectrais das janelas rectangular e Hamming [26]

[1] No sistema reconhecedor *multi-stream*.

verifica-se que ela é aproximadamente metade no caso da janela rectangular. Assim, esta janela permite uma maior resolução do sinal no domínio das frequências. No entanto, isto é geralmente insuficiente para compensar a desvantagem associada à bastante maior irregularidade da curva do espectro no caso de usar a janela rectangular, em resultado também da maior amplitude da transformada de Fourier desta janela fora da banda passante. Este efeito reflecte-se na suavidade das trajectórias do sinal no espaço de representação acústica. A maior suavidade da curva correspondente à janela Hamming é evidente na figura 2.6 que mostra a evolução da energia, para uma única frase do conjunto de treino, no filtro passa-banda triangular cuja frequência central é igual a 156 Hz e pertencente ao banco de filtros usado posteriormente na análise do sinal.

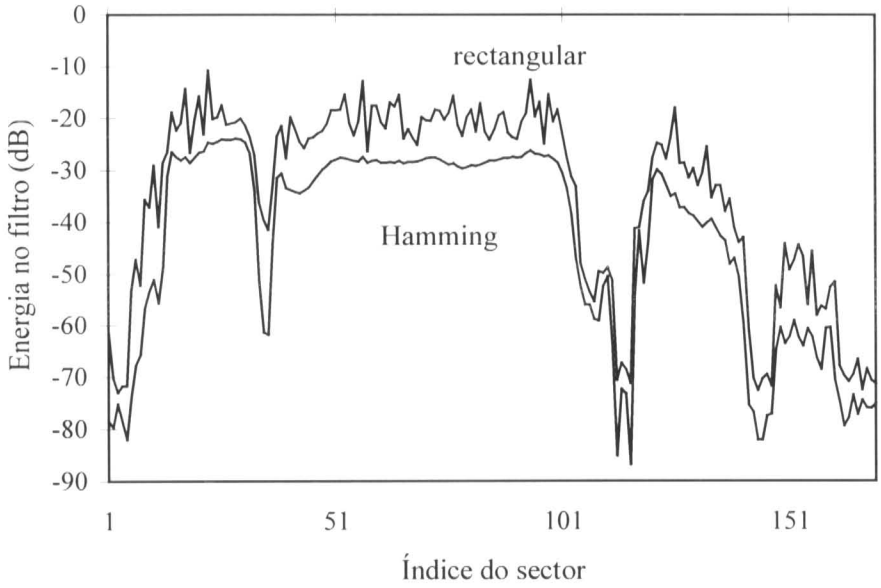

Fig. 2.6 - Evolução da energia num filtro passa-banda usando dois tipos de janela (correspondente à frase "change area alert pac").

O facto de neste sistema os filtros passa-banda terem uma largura de banda que é progressivamente maior à medida que a respectiva frequência central aumenta, contribui para uma menor diferença de suavidade, entre as curvas de energia correspondentes às janelas rectangulares e Hamming, nos filtros colocados em frequências mais elevadas. Ainda assim, a trajectória do vector de características final[1] no espaço acústico apresenta um comportamento mais suave quando se utiliza a janela Hamming, sendo esta característica decisiva na escolha deste tipo de janela.

[1] Os MFCC são obtidos descorrelacionando as energias à saída dos filtros passa-banda.

Devido às amostras do sinal serem muito atenuadas nas extremidades da janela Hamming é necessário sobrepor parcialmente os sectores. Neste sistema é utilizado o valor típico de 15 ms para a sobreposição entre sectores consecutivos.

2.5.2.4 Cálculo dos coeficientes Mel-cepstrais

Para cada sector do sinal são calculados os respectivos coeficientes cepstrais utilizando um procedimento heurístico baseado na energia em bandas espectrais espaçadas de acordo com a escala Mel.

Transformada de Fourier do sinal

Inicialmente, a cada sector do sinal são apensos os zeros[1] necessários para obter uma sucessão com 512 termos e assim poder calcular a transformada discreta de Fourier através de um eficiente algoritmo do tipo transformada rápida de Fourier (FFT) [26].

Obtém-se assim o espectro de potência do sinal:

$$\Gamma_n(k) = \left| \sum_{l=1}^{L} s_n(l) e^{-j\frac{2\pi}{512}k(l-1)} \right|^2 \quad , \quad k = 1,2,\ldots,512 \tag{2.10}$$

Banco de filtros

O espectro de potência é aplicado a um banco de filtros com forma triangular como é mostrada na figura 2.7.

A posição dos 24 filtros passa-banda encontra-se definida de acordo com a escala Mel. Essa escala resulta de estudos na área da psico-acústica [31] que indicam que a percepção humana da frequência associada aos sons não segue uma escala linear para as frequências. De entre as diversas formas funcionais [32] que têm sido utilizadas para modelar os dados obtidos experimentalmente relacionando a frequência real e a perceptível, f_{mel}, uma função frequentemente utilizada é:

[1] Um procedimento designado, em Inglês, por *zero padding*.

$$f_{mel} = 2595 \log_{10}\left(1 + \frac{f_{real}}{700}\right) \tag{2.11}$$

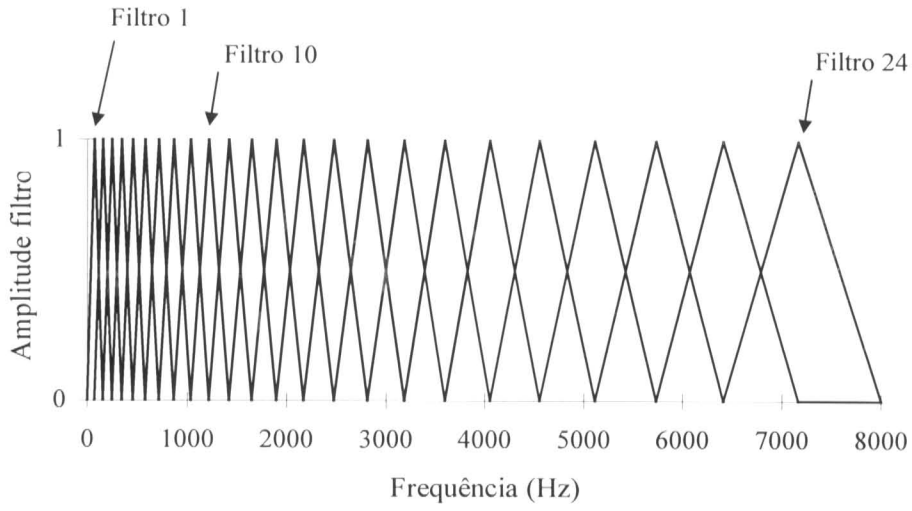

Fig. 2.7 - Banco de filtros.

Assim, na escala Mel aproximada por esta função, os filtros deste sistema encontram-se linearmente espaçados. Na escala real de frequências, à semelhança do que sucede com o ouvido humano, este sistema é dotado de uma resolução espectral cuja dependência da frequência é não linear. Essa resolução é relativamente maior para as frequências mais baixas, o que permite uma melhor representação do conteúdo relevante do espectro para a tarefa do reconhecimento [25].

À saída de cada filtro está então disponível uma estimativa da energia do sinal nessa banda, obtida a partir da soma pesada, através dos respectivos coeficientes $H_i(k)$, das componentes do espectro:

$$y_n(i) = \sum_k \Gamma_n(k) H_i(k) \quad , \quad i = 1,\ldots,24 \tag{2.12}$$

Um número significativo de sistemas utiliza um banco de filtros definido de acordo com uma outra escala, designada escala Bark [34], também com raízes no sistema auditivo mas baseada num critério diferente para determinação da posição e largura dos filtros passa-banda. O conceito subjacente a essa escala consiste na determinação de larguras de banda críticas para diferentes frequências de maneira que em cada uma dessas bandas não é possível distinguir as componentes espectrais de sons complexos. Serve também esta referência para salientar que apesar das diversas técnicas inspiradas no conhecimento existente sobre o sistema auditivo humano, elas

não constituem necessariamente soluções finais ou óptimas para alcançar o objectivo que é minimizar as taxas de erro dos reconhecedores automáticos da fala. Assim, com base nesse critério objectivo, para este problema em particular continuam a ser investigadas técnicas de optimização do desenho do banco de filtros [35].

Em todo o caso, em resultado da utilização de qualquer destes bancos de filtros é conseguida uma grande redução da dimensão da representação espectral do sinal.

Compressão logarítmica

Outro resultado fundamental do modelo de auditório consiste na compressão da energia do sinal. Esse procedimento é simulado neste sistema submetendo a energia associada a cada filtro a uma transformação logarítmica, resultando assim o vector Ψ_n cujas componentes são:

$$\psi_n(i) = \log(y_n(i)), \quad i = 1,\ldots,24 \tag{2.13}$$

Também têm sido utilizadas outras formas funcionais na compressão da energia, como é o caso da função sigmóide [33] que permite reduzir significativamente uma dificuldade típica da compressão logarítmica que consiste no ênfase atribuído às partes do sinal com muito baixo nível de energia e relativamente ruidosos.

Neste sistema é adicionado um valor constante à energia em cada filtro, antes do cálculo do logaritmo, com o objectivo de evitar problemas numéricos em segmentos do sinal onde se verifica o quase-silêncio. Durante o desenvolvimento do sistema foi realizada uma breve experiência com o objectivo de ajustar o valor dessa constante, δ, reduzindo a irregularidade das curvas da energia durante os períodos de quase silêncio. Tal como esperado, um valor grande de δ suavizava a energia nos períodos de silêncio mas aumentava significativamente a taxa do erro. Para valores inferiores a essa referência o desempenho do sistema mostrou-se pouco sensível à variação de δ. Experimentou-se também ajustar δ dependendo do filtro passa-banda, sendo para cada filtro calculado o histograma do respectivo valor da log-energia, como o exemplo típico mostrado na figura 2.8 calculado com todos os sectores do conjunto de treino.

Por inspecção do histograma relativo a cada filtro obteve-se o intervalo aproximado da energia nos sectores correspondentes aos períodos de quase-silêncio. No filtro correspondente à figura 2.8 esse intervalo está definido numa vizinhança do valor, aproximadamente, -60 dB, a abcissa do pico mais à esquerda no histograma. Com base nessa informação foram então testados valores δ dependentes do filtro passa-banda, tendo numa grande parte desses testes resultado a redução das taxas do erro. No entanto, em nenhum dos casos essa redução foi estatisticamente significativa.

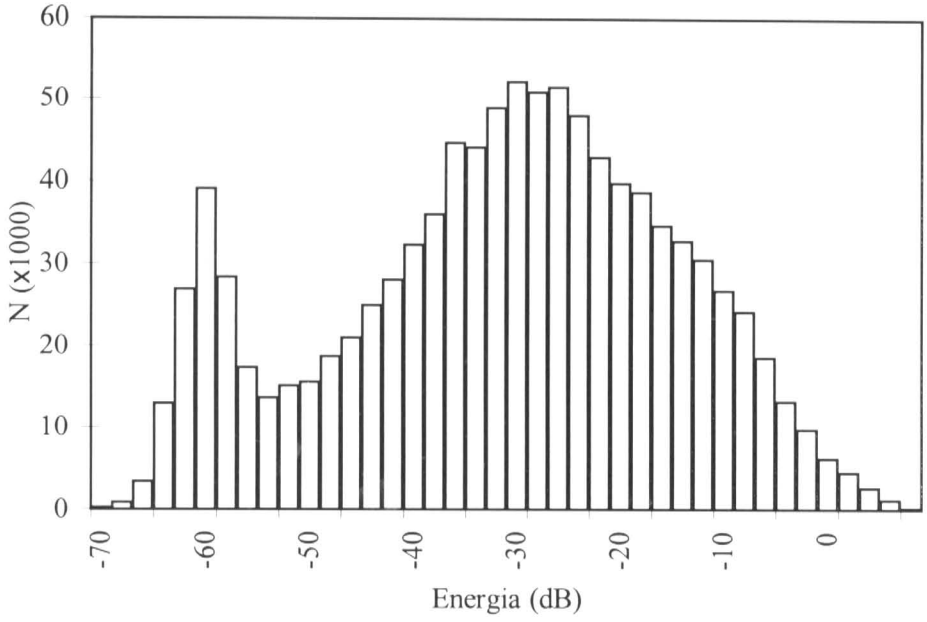

Fig. 2.8 - Histograma da energia no filtro com frequência central 1646 Hz.

É oportuno fazer aqui uma breve referência a uma ocorrência acidental, devida a um lapso ao escrever o código, verificada durante a fase inicial do desenvolvimento deste sistema: era calculado o logaritmo do espectro de potência e só depois esse resultado era submetido ao banco de filtros; no entanto, e enquanto esse erro não foi detectado, foram obtidos alguns resultados para as taxas de reconhecimento em alguns testes preliminares; verificou-se posteriormente que esses resultados eram próximos dos obtidos após a "correcção" dos programas. Esses resultados são aparentemente surpreendentes uma vez que antes da correcção era comprimido o produto dos termos, isto é, calculava-se o $\log_{10}\left(\prod_k \Gamma_n(k) H_i(k)\right)$. No entanto, em alguns sistemas de análise de características, e de acordo com resultados publicados [33], é conveniente antecipar a transformação logarítmica, aplicando-a aos termos do somatório apresentado na equação 2.12.

Descorrelação

Uma característica desta representação espectral do sinal da fala, desde logo previsível dada a configuração do banco de filtros, é a grande correlação que se verifica entre as componentes de Ψ. Esse facto pode ser confirmado através da figura 2.9 que mostra a matriz correlação do vector Ψ, calculada a partir de aproximadamente 25000 vectores de características extraídos aleatoriamente do conjunto de treino.

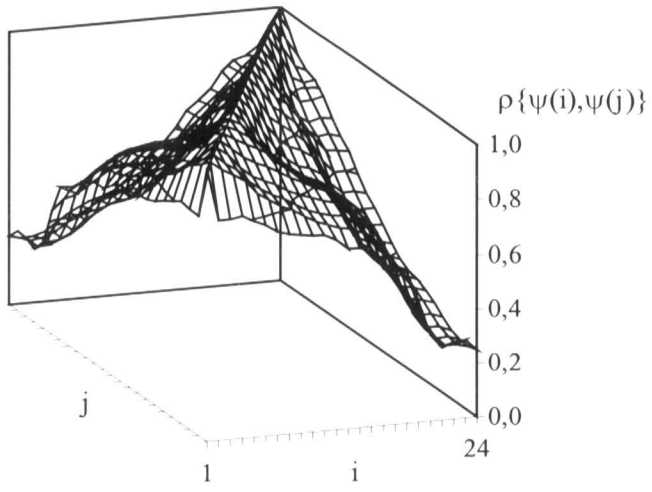

Fig. 2.9 - Correlação das log-energias nos filtros passa-banda.

Nos sistemas baseados em modelos acústicos que não modelam essa correlação, como sucede em alguma medida neste caso, torna-se necessário descorrelacionar as componentes do vector Ψ.

Para o efeito, neste sistema é utilizada uma transformação linear conhecida por transformada discreta, ou inversa, do coseno e definida com base numa família de funções discretas ortogonais entre si. O vector Ψ é assim transformado num outro vector, **C**, cujas componentes são calculadas da seguinte maneira:

$$c_n(i) = \sum_{k=1}^{24} \psi_n(k) \cos\left(i\left(k - \frac{1}{2} \right) \frac{\pi}{24} \right), \quad i = 1,\ldots,12 \tag{2.14}$$

A figura 2.10 mostra a matriz correlação de **C**, o vector dos coeficientes Mel-cepstrais, calculada a partir dos mesmos vectores de características utilizados no cálculo da matriz correlação do vector Ψ e apresentada na figura 2.9. Comparando as figuras 2.9 e 2.10 é clara a redução da magnitude média da correlação entre os termos nesta nova representação acústica.

O nível de correlação que subsiste é aceitável para este sistema, apesar das funções de Gauss multivariável utilizadas para cobrir o espaço acústico não modelarem explicitamente a correlação existente entre as componentes do vector de características, uma vez que têm matriz das covariâncias diagonal. A utilização de modelos escondidos de Markov do tipo semi-contínuo permite que o número de

funções de Gauss correspondentes a cada estado seja relativamente grande[1], sendo assim satisfatoriamente modelada a distribuição real dos vectores de características associados a cada estado sem utilizar funções de Gauss com matriz das covariâncias completa.

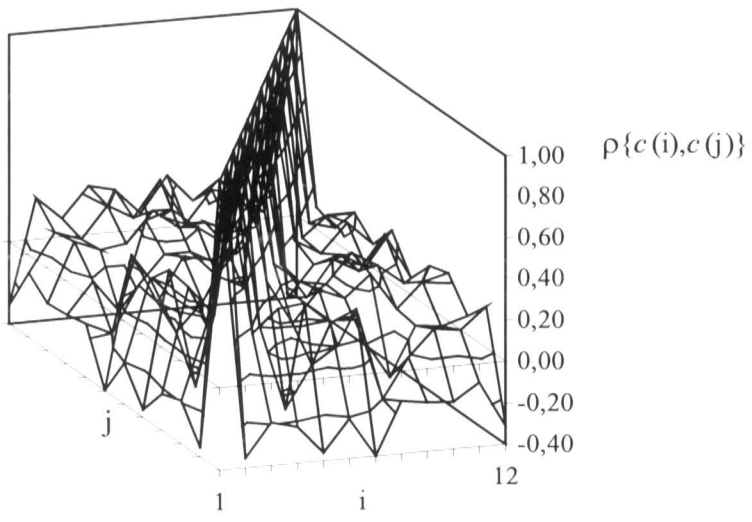

Fig. 2.10 - Correlação dos coeficientes Mel-cepstrais.

Em alguns sistemas são utilizadas outras técnicas para descorrelacionar as componentes dos vectores de características, como é o exemplo bem conhecido da transformação Karhunen-Loève. A aplicação dessa transformação, também linear, ao vector com as log-energias no banco de filtros resultaria numa representação globalmente descorrelacionada, isto é, a matriz correlação calculada com base no conjunto (de treino) completo seria a matriz identidade[2]. Por outro lado, a aplicação dessa transformação apenas com base no cálculo da matriz das covariâncias global não assegura a descorrelação local do vector de características, objectivo que de facto seria muito importante alcançar e relativamente ao qual têm sido investigadas diversas técnicas [36]. Refira-se que a transformada discreta do coseno também não resolve directamente este problema da descorrelação local das componentes do vector de características.

Apesar da existência de resultados assinalando alguma superioridade de outros métodos de descorrelação relativamente à transformada discreta do coseno [37], as razões essenciais que motivaram a escolha desta solução foram a consistência e

[1] Da ordem das centenas, número bastante maior ao que geralmente se verifica nos sistemas baseados nos HMM contínuos onde, também por esta razão, são frequentemente utilizadas funções de Gauss com matriz das covariâncias completa.

[2] Também relativamente às covariâncias resultaria a matriz identidade, uma vez que esta transformação normaliza as componentes do vector de características pelos respectivos desvios padrão.

qualidade dos resultados que permite alcançar também em sistemas com características similares às deste e o facto de ser uma técnica cuja execução é relativamente simples.

Uma característica também interessante da transformada discreta do coseno, embora não seja um exclusivo seu, consiste no facto de permitir reduzir substancialmente a dimensão da representação acústica do sinal sem que ocorra a perda significativa de informação essencial ao processo de classificação.

Conceitos sobre a análise cepstral

Antes de terminar este capítulo é interessante acrescentar uma breve referência a alguns conceitos que contribuem para fundamentar a utilização dos coeficientes Mel-cepstrais. O processo de cálculo dos MFCC acabado de apresentar, frequente em reconhecedores automáticos da fala, tem as suas raízes no conhecimento apriorístico sobre o sistema auditivo humano e também, como o próprio nome sugere, numa técnica relevante de processamento do sinal, designada análise cepstral[1] [27], cuja utilização no reconhecimento automático da fala a seguir se resume.

De acordo com o modelo de produção acústica referido anteriormente (§2.5.2.2), o sinal da fala resulta da convolução do sinal gerado pela fonte de excitação com a resposta impulsional do filtro correspondente ao tracto vocal. No reconhecimento automático da fala é importante desconvolucionar estas duas componentes do sinal, para poder assim isolar a informação que está associada à configuração do tracto vocal, aquela que interessa extrair por ser a que veicula o conteúdo linguístico do sinal. A análise cepstral é uma técnica utilizada com esse objectivo.

Sejam $E\left(e^{j\omega}\right)$ e $H\left(e^{j\omega}\right)$ as transformadas de Fourier, respectivamente, da excitação e do filtro. Então, efectuando a transformada de Fourier do sinal de fala obtém-se:

$$S\left(e^{j\omega}\right) = E\left(e^{j\omega}\right)H\left(e^{j\omega}\right) \tag{2.15}$$

Aplicando a transformação logarítmica a este resultado, e considerando a representação de fase nula do sinal[2] [6], resulta:

[1] Que por sua vez faz parte de uma classe de métodos frequentemente designados desconvolução homomórfica.

[2] No caso do chamado cepstrum complexo utilizar-se-ia o logaritmo complexo, $\log|z| + i\arg(z)$.

$$\log\left|S\left(e^{j\omega}\right)\right| = \log\left|E\left(e^{j\omega}\right)\right| + \log\left|H\left(e^{j\omega}\right)\right| \tag{2.16}$$

Verifica-se, assim, que neste novo domínio de representação estas duas componentes do sinal da fala se encontram linearmente combinadas. Além disso, no segundo membro da equação 2.16 o termo correspondente ao modelo do tracto vocal apresenta uma variação com a frequência relativamente lenta, ao contrário do que sucede com o outro termo. Logo, aplicando a transformada (inversa) de Fourier a $\log\left|S\left(e^{j\omega}\right)\right|$ resulta uma nova representação onde as duas componentes do sinal acústico estão definidas em regiões diferentes do domínio e são portanto facilmente separáveis através dum processo muito simples de filtragem[1].

2.5.2.5 Cálculo da energia

Um dos elementos do vector de características é o logaritmo da energia do sinal em cada segmento. Este termo é calculado a partir da energia à saída do banco de filtros:

$$e_n = \log_{10}\left(\sum_{k=1}^{24} y_n(k)\right) \tag{2.17}$$

Durante o desenvolvimento deste sistema tentou-se averiguar o desempenho relativo de soluções alternativas para a representação da energia. Para tornar as experiências mais expeditas, os modelos acústicos foram treinados com uma única iteração após a sua iniciação e não foi utilizada a base de dados completa[2]. No entanto, o volume de dados foi considerado suficiente para permitir optar por outras representações na eventualidade de se verificarem diferenças significativas e positivas nos resultados.

Numa dessas experiências, tal como sucede em alguns reconhecedores foi utilizado o coeficiente Mel-cepstral de ordem zero, ou seja, $c_n(0) = \sum_{k=1}^{24} \log_{10}\left(y_n(k)\right)$. Apesar da assinalável correlação existente entre $c_n(0)$ e e_n, verificou-se uma degradação significativa da taxa de palavras erradas em consequência desta substituição.

Foram também efectuadas experiências com a normalização da log-energia, utilizando alguns critérios definidos a partir de diferentes medidas de energia máxima

[1] Ou, em Inglês, *"liftering"*.
[2] Por selecção aleatória, 720 das 2840 frases do conjunto de treino e 100 das 300 frases de teste.

ou mínima ao nível da frase. No entanto, nos testes realizados o melhor resultado verificou-se sem aplicar qualquer técnica de normalização.

2.5.2.6 Análise dinâmica

À entrada do módulo de análise dinâmica o vector de características é composto por 12 coeficientes Mel-cepstrais e por um termo representando a energia. A sucessão de eventos acústicos presentes no sinal da fala é assim representada por uma trajectória definida no espaço $\Re^{13}$.

Os modelos acústicos do sistema deveriam modelar essa trajectória e não apenas as posições instantâneas no espaço de representação acústica definidas pelo vector de características, pois como é evidente um qualquer ponto pode pertencer a diferentes trajectórias.

Mas os modelos escondidos de Markov apresentam grandes limitações a esse nível, sendo portanto necessário introduzir em cada vector de características alguma informação relativa à forma da trajectória nas proximidades do respectivo ponto. De facto, a topologia dos HMM define uma estrutura temporal capaz de modelar a evolução do vector de características de um estado para o seguinte. No entanto, em cada estado esses vectores são consideradas estatisticamente independentes, portanto também descorrelacionados, e identicamente distribuídos, partilhando a mesma função de distribuição probabilística. Consequentemente, apenas é possível modelar, em alguma medida, a trajectória do sinal na região acústica abrangida por cada estado se essa informação for explicitamente incorporada no vector de características.

Este problema da utilização de informação de contexto, ao nível do sector, é de grande importância. Assim, têm sido apontados diversos métodos que conduzem a uma clara melhoria do desempenho dos reconhecedores [38, 41], desde soluções com base no processamento dos vectores de características, como por exemplo agrupando-os em janelas de contexto e posteriormente submetendo-os a uma forte redução da dimensão, até soluções ao nível do modelo acústico, como é o exemplo de alguns modelos híbridos que utilizam redes neuronais artificiais. É oportuno fazer aqui uma breve referência a uma abordagem claramente orientada para solucionar este problema da modelação da trajectória. São os chamados modelos de segmentos[1], um formalismo estatístico que constitui uma poderosa generalização dos HMM [49] e que tem alcançado resultados globalmente muito promissores[2].

Neste sistema é utilizado um método muito frequente devido à sua consistência nos resultados e também à sua grande simplicidade. Baseia-se no cálculo de uma

[1] Em Inglês, *(stochastic) segmental models*.

[2] Embora apresentem também desvantagens importantes, por exemplo, uma enorme exigência de dados acústicos para treino dos parâmetros, em número muito elevado, e também uma grande complexidade computacional.

regressão linear efectuada sobre as características nos sectores vizinhos [39]. No entanto esta técnica não está isenta de defeitos, refira-se por exemplo o problema potencial, notado por diversos investigadores [19], que consiste na maior sensibilidade do reconhecedor relativamente à celeridade do discurso da fala.

De acordo com este método, para cada um dos parâmetros, x_n^i, $i = 1,2,\ldots,13$, do vector de características existente é inicialmente calculado um novo termo, apropriadamente designado "primeira derivada pesada", utilizando a fórmula:

$$\Delta x_n^i = \frac{\sum_{k=-2}^{2} k\, x_{n+k}^i}{\sum_{k=-2}^{2} k^2}, \quad i = 1,2,\ldots,13 \tag{2.18}$$

Com base nos parâmetros, Δx_n^i, $i = 1,2,\ldots,13$ são em seguida calculadas as "segundas derivadas" dos parâmetros originais através da fórmula:

$$\Delta^2 x_n^i = \frac{\sum_{k=-1}^{1} k\, \Delta x_{n+k}^i}{2}, \quad i = 1,2,\ldots,13 \tag{2.19}$$

É assim composto o vector de características utilizado neste sistema, o qual se encontra definido num espaço $\Re^{39}$:

$$\boldsymbol{x}_n = \left(x_n^1,\ldots,x_n^{13},\Delta x_n^1,\ldots,\Delta x_n^{13},\Delta^2 x_n^1,\ldots,\Delta^2 x_n^{13}\right) \tag{2.20}$$

A figura 2.11 mostra uma representação gráfica da matriz correlação deste vector, tornando claro que se verifica uma correlação significativa entre, por exemplo, os termos estáticos e as respectivas segundas derivadas. Tal como foi referido anteriormente, poder-se-ia utilizar uma transformação para tentar descorrelacionar as componentes deste vector, operação que no entanto não é realizada neste sistema.

Em alguns sistemas é feita a projecção do vector de características num espaço de menor dimensão utilizando geralmente uma transformação linear [40]. Neste sistema isso não foi feito, apesar da dimensão do vector de características ser já assinalável, porque se considerou essa como uma questão lateral face aos objectivos do trabalho. No entanto, em diversas ocasiões durante o desenvolvimento do sistema ficou patente a redundância e falta de capacidade discriminativa de alguns dos elementos do vector de características. Referindo apenas um exemplo, durante as sucessivas iterações de treino do sistema pôde verificar-se que particularmente as variâncias, utilizadas na

parametrização das funções de Gauss[1], correspondentes às segundas derivadas dos MFCC de ordem mais elevada não eram, em termos médios, adaptadas de acordo com a tendência geral que consistia na redução sensível do seu valor. Isto é, com o decorrer do processo de treino, a generalidade das componentes do vector de características aumentava o seu peso no argumento das exponenciais relativamente às componentes do vector correspondentes às segundas derivadas dos MFCC de ordem mais elevada, os quais apresentavam efectivamente uma distribuição globalmente mais difusa.

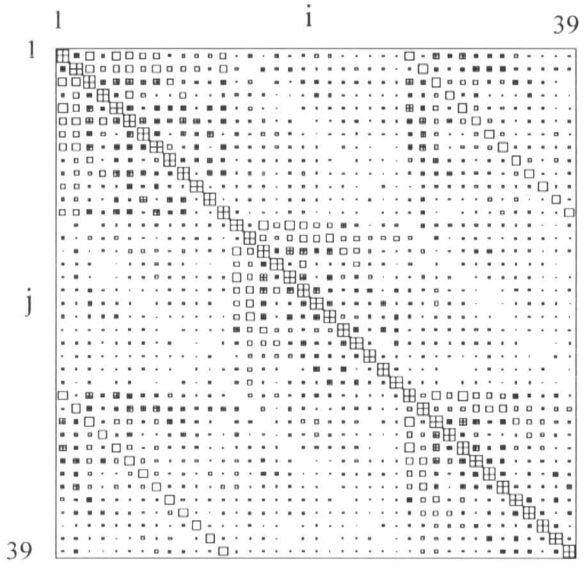

Fig. 2.11 - Correlação do vector de características, $\rho\left(x_n^i, x_n^j\right)$.

2.6 Modelação acústica

2.6.1 Introdução

O problema que se coloca à saída do módulo de análise do sinal é a classificação correcta da sucessão de vectores de características. De acordo com a formulação apresentada (§2.2), esta transformação é feita com base nos modelos linguístico, L, e acústico, M, associados a esta tarefa e integrados no reconhecedor:

$$\boldsymbol{X}_1^T \xrightarrow[L,\,M]{} W_1^K \tag{2.21}$$

Neste sistema o sinal da fala é considerado essencialmente como um padrão estocástico, sendo portanto utilizada uma abordagem baseada no reconhecimento estatístico de padrões, aquela que por norma apresenta melhores resultados (§1.3.5).

[1] As matrizes das covariâncias são diagonais.

De acordo com o critério de classificação Bayesiano (§1.3.6.3), uma sucessão genérica de vectores de características $\boldsymbol{X}$ é reconhecida como a realização acústica da sucessão de palavras W que maximiza, no universo de frases possíveis, a probabilidade conjunta, $P(\boldsymbol{X},W)$. Factorizando esta probabilidade pode obter-se o resultado equivalente $P(\boldsymbol{X}|W)P(W)$, tornando possível decompor a função discriminante do classificador no produto de dois modelos estatísticos (§1.3.6.3). O modelo linguístico, L, determina a probabilidade *a priori* de ocorrência da frase, independentemente do sinal acústico, isto é, $P(W)$. O modelo acústico, M, determina a probabilidade de observar a sucessão de vectores de características condicionada ao conhecimento da frase, ou seja, $P(\boldsymbol{X}|W)$.

A modelação acústica, o problema que se encontra agora em discussão, é assim decisiva no desempenho do classificador, pois uma boa aproximação ao classificador Bayesiano apenas é possível se M, o modelo acústico global, determinar com suficiente correcção a verosimilhança global (§1.3.6.7).

A abordagem deste problema durante a fase de desenvolvimento do sistema apresenta normalmente duas etapas que antecedem o treino do sistema. Em primeiro lugar, é necessário escolher o conjunto, Γ_u, de unidades acústicas elementares necessárias para compor qualquer sucessão de palavras possível na tarefa, isto é, $W=\{u_1,u_2,\ldots,u_M\}$, $\forall u \in \Gamma_u$. Em seguida, é necessário definir o formalismo estatístico para modelar cada uma dessas unidades, ou seja, a estrutura matemática que permite calcular $P(\boldsymbol{X}|u)$, $\forall u \in \Gamma_u$.

A exposição que se segue trata estes problemas no caso particular deste reconhecedor.

2.6.2 Unidades fonéticas

2.6.2.1 Fundamentos

Se a base de dados para treino, assim como o espaço de memória e velocidade de processamento do sistema, fossem praticamente ilimitados, poder-se-ia definir, treinar e utilizar um modelo estatístico para cada uma das frases permitidas pelo modelo linguístico. Essa disponibilidade "ilimitada" de recursos permitiria modelar com rigor as diversas dependências estatísticas que se verificam no sinal da fala, por exemplo a co-articulação entre palavras[1], o que certamente resultaria em taxas de reconhecimento muito interessantes. No entanto, por razões óbvias, essa não é uma solução.

A solução é definir um conjunto de unidades acústicas elementares:

[1] Refira-se que a co-articulação entre palavras é também parcialmente modelada em alguns sistemas baseados em modelos acústicos elementares ao nível fonético.

$$\Gamma_u = \left\{u^1, u^2, ..., u^L\right\} \tag{2.22}$$

e criar um modelo acústico para cada uma dessas unidades:

$$u^i \leftrightarrow M^i, \quad i=1,...,L \tag{2.23}$$

de maneira a que o modelo acústico correspondente a uma qualquer frase possa ser construído concatenando, segundo determinadas regras, esses modelos elementares:

$$M_W = M_1 \oplus M_2 \oplus ... \oplus M_M \tag{2.24}$$

Assim, o modelo acústico global, M, é formado com base num conjunto de modelos acústicos correspondentes a unidades elementares de fala (§1.3.6.7) definidas a um nível que, por enquanto, se poderá designar "sub-frase".

Consistindo esta tarefa no reconhecimento da sucessão de palavras associada ao sinal da fala, a escolha imediata para essas unidades acústicas recairia provavelmente sobre as próprias palavras. Mais importante que este aspecto seria a natural robustez que esses modelos acústicos exibiriam ao problema importante da co-articulação no interior de cada palavra.

No entanto, a base de dados *Resource Managment* é manifestamente exígua para treinar com eficácia um modelo acústico dedicado a cada uma das 991 palavras do vocabulário. Contudo, deve reconhecer-se que possivelmente o volume de dados seria suficiente para treinar as palavras mais frequentes, entre as quais se encontrariam algumas das responsáveis por um número significativo de erros (§2.3) e que portanto exigiriam melhor modelação. No entanto, optou-se por utilizar um único tipo de unidades acústicas em todo o sistema pelo que esta eventual solução não foi seguida.

A solução aqui utilizada para obter modelos acústicos robustos está assente no princípio da partilha dos parâmetros (§1.3.6.5) aplicado ao nível sub-palavra.

Em grande medida devido ao vasto conhecimento existente sobre o funcionamento fonético da língua, os fonemas são geralmente escolhidos como unidades elementares da fala. Uma vantagem importante que decorre directamente dessa escolha é a facilidade na construção do léxico. Apesar disso, em alguns sistemas as unidades elementares são definidas sem utilizar esse conhecimento linguístico, recorrendo antes a algum mecanismo automático de agrupamento[1] de sons baseado num critério objectivo definido a partir das distâncias acústicas [61]. O conjunto das unidades utilizadas neste sistema baseia-se nos fonemas *standard* definidos em Inglês,

[1] Em Inglês, *clustering*.

tendo sido posteriormente submetido a algumas alterações durante o desenvolvimento de um sistema[1] de reconhecimento para a mesma aplicação.

Os fonemas podem ser modelados utilizando informação de contexto, isto é, considerando os efeitos dos sons anterior e posterior, a identidade da palavra em que ocorrem e até informação prosódica. Muitos sistemas utilizam com grande sucesso tri-fonemas [59], ou seja, fonemas modelados considerando a identidade dos fonemas anterior e posterior. Frequentemente esses modelos são robustecidos [69,70,77] combinando-os com os respectivos modelos obtidos com contexto apenas em um dos extremos, os bi-fonemas, e também com os modelos sem contexto, como sucede com a técnica conhecida por *deleted interpolation*. Essa e outras técnicas [56] tornam-se indispensáveis quando os recursos de dados acústicos de treino são relativamente escassos ou os contextos são mais abrangentes, como sucede em alguns sistemas de muito grande vocabulário que utilizam penta-fonemas [68]. Em alguns sistemas os modelos fonéticos modelam também a co-articulação entre palavras contíguas [68], o que aumenta muito a sua complexidade computacional. Existem diversas técnicas [78,79,80,81] que permitem considerar apenas os contextos que são de facto relevantes no desempenho do sistema.

Na construção deste sistema optou-se pela modelação sem contexto, restringindo assim significativamente a sua complexidade. Por outro lado, tal como seria previsível esta acentuada partilha de parâmetros acarreta uma relativa falta de nitidez dos modelos.

A figura 2.12 apresenta um exemplo, baseado no modelo acústico da consoante africada /jh/, que ilustra este fenómeno. Refira-se, muito resumidamente, que esta classe de sons é representada por um modelo escondido de Markov do tipo semi-contínuo com três estados, dois deles dedicados aos extremos do segmento de sinal e o outro à parte central. As três curvas do gráfico representam os coeficientes, ordenados segundo valores decrescentes, das misturas gausseanas nesses estados. A curva com o traço mais grosso corresponde ao estado que modela a parte central do /jh/. Por simples inspecção do gráfico verifica-se que no estado central a mistura é dominada por um número relativamente reduzido de funções de Gauss, sendo o modelo substancialmente mais difuso nos extremos. Em certa medida, o facto do estado central focar uma região mais restrita do espaço acústico pode dever-se também a aí o fonema se encontrar na sua fase mais estacionária. Assim, foi realizada uma pequena experiência para comparar os coeficientes da mistura gausseana no primeiro estado semeando o modelo de duas maneiras diferentes: sem qualquer contexto; utilizando um contexto à esquerda único. Confirmou-se que no primeiro caso o estado inicial era significativamente mais desfocado, apresentando grande insensibilidade aos efeitos de co-articulação.

[1] Desenvolvido no laboratório ELIS, Universidade de Ghent (onde este sistema foi, parcialmente, também desenvolvido).

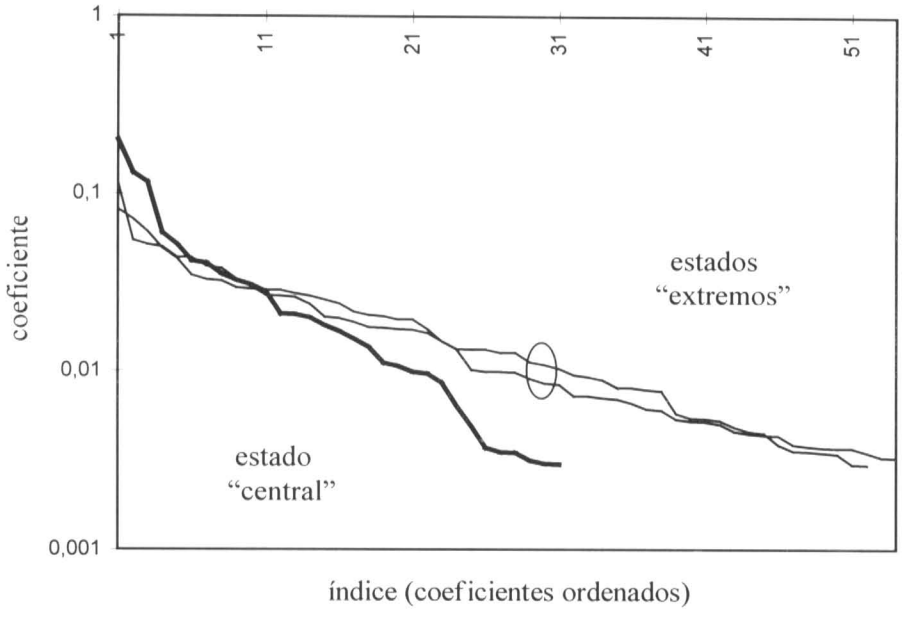

Fig. 2.12 - Curvas dos coeficientes (ordenados decrescentemente) das misturas gausseanas nos três estados do modelo acústico do fonema /jh/.

Embora se verifique alguma degradação do desempenho do sistema (§1.3.6.5) devido à não utilização de modelos com contexto, este mantém-se a um nível perfeitamente aceitável. Além disso, a modelação com contexto é dispensável no âmbito dos objectivos delineados para este trabalho (§2.1).

Apesar do que acaba de ser exposto, deve ser salientado que o conjunto de unidades acústicas elementares utilizadas neste sistema permite considerar, também através do léxico, o problema do contexto em algumas situações particularmente críticas nesta tarefa. Por exemplo, para modelar o som correspondente à consoante oclusiva associada à letra "*B*" estão definidas dois modelos acústicos, notados pelos símbolos /bcl/ e /b/. De maneira simplificada, pode dizer-se que /bcl/ corresponde à parte inicial do som, durante o período de oclusão, e /b/ modela a parte restante. Assim, quando a parte inicial do som é fortemente afectada por uma consoante nasal imediatamente anterior, como sucede em "*NUMBER*", o "*B*" é modelado apenas por /b/. Pelo contrário, quando a realização deste som é nítida e completa, como por exemplo na palavra "*BAY*", o "*B*" é modelado por /bcl+b/. Ou ainda, por exemplo, quando a pronunciação é débil e incompleta, apresentando apenas a parte inicial do som, como sucede em "*SUBMARINE*", o "*B*" é modelado apenas por /bcl/.

2.6.2.2 Descrição

Neste sistema são utilizadas 52 unidades acústicas elementares que constituem uma extensão do vocabulário fonético geralmente considerado para a língua inglesa[1], razão pela qual serão designadas unidades fonéticas ou, abreviadamente, fonemas. Na generalidade das línguas[2] estão definidos diversos critérios para classificar os fonemas [10], ou outras unidades da fala. A taxinomia que aqui se considera reflecte, essencialmente, o tipo de excitação e a maneira de articulação.

Vogais

As vogais resultam da repercussão da onda acústica, associada ao fluxo de ar que à saída da glote apresenta uma pressão periódica, nas partes superiores do canal de expiração. Essencialmente são o volume e a forma da cavidade bucal, determinados pela posição da língua e dos lábios e do grau de abertura da boca, que estabelecem o timbre característico de cada vogal. É possível distinguir as vogais entre si através da localização das três primeiras frequências *formantes* [6]. Normalmente as vogais apresentam uma amplitude acústica relativamente elevada e a sua duração é bastante variável, tipicamente pertencente ao intervalo [40 ms, 400 ms].

Neste sistema são consideradas catorze unidades acústicas correspondentes a vogais em língua inglesa[3].

Símbolo	Exemplo
/iy/	*AREA*
/ih/	*IT*
/ey/	*EIGHT*
/eh/	*EDIT*
/ae/	*AND*
/er/	*FIRST*
/ax/	*AS*

Símbolo	Exemplo
/axr/	*NEVER*
/ah/	*MUCH*
/uw/	*DO*
/uh/	*COOK*
/ow/	*GO*
/ao/	*ALL*
/aa/	*ARE*

Tab. 2.1 - Unidades acústicas correspondentes a sons vogais.

[1] Frequentemente são considerados 39 fonemas essenciais no Inglês mas, por exemplo, alguns reconhecedores para determinadas tarefas em Inglês utilizam léxicos com 47 símbolos de natureza fonética.

[2] Não sendo o Português e o Inglês excepções.

[3] A notação simbólica aqui seguida é baseada na notação *standard* utilizada na base de dados TIMIT.

É de salientar que para além das treze vogais geralmente consideradas no Inglês, é utilizada a unidade notada pelo símbolo /axr/ e correspondente a um som vocálico cuja modelação independente se revelou positiva em alguns sistemas utilizando esta base de dados e idêntica tarefa.

Ditongos e semivogais

Este grupo de sons de natureza vocálica caracteriza-se pela sua não estacionaridade, em resultado da alteração do tracto vocal durante a sua emissão. Para este grupo de sons estão neste sistema definidas oito unidades acústicas elementares.

Símbolo	Exemplo
/ay/	*BY*
/aw/	*COUNT*
/oy/	*EMPLOYED*
/r/	*CORAL*

Símbolo	Exemplo
/l/	*CLEAR*
/el/	*VESSEL*
/w/	*WEST*
/y/	*YORKTOWN*

Tab. 2.2 - Unidades acústicas correspondentes a ditongos e semivogais.

Nos ditongos, que neste sistema são /ay/, /aw/ e /oy/, verifica-se uma transição claramente perceptível entre duas vogais. No caso das semivogais, a realização acústica apresenta um determinado período de estacionaridade, numa região vocálica, mas verifica-se também uma fase transitória.

Consoantes nasais

As consoantes nasais são sons cuja excitação é vocálica, encontrando-se a cavidade bucal exteriormente fechada e as fossas nasais incorporadas no tracto vocal.

Símbolo	Exemplo
/m/	*HIM*
/n/	*IN*
/en/	*REASONS*
/ng/	*LONG*

Tab. 2.3 - Unidades acústicas correspondentes a consoantes nasais.

Alguns fonemas não nasais encontram-se nasalizados devido ao efeito de co-articulação. Isso sucede especialmente com vogais que precedem um som nasal, verificando-se a descida do véu palatino ainda durante a produção da vogal. Neste sistema é utilizada a unidade acústica /en/ para modelar uma dessas situações que é particularmente sensível nesta tarefa de reconhecimento.

Consoantes fricativas

Estes sons resultam da turbulência que se verifica na passagem de um fluxo contínuo de ar por uma região de estrangulamento do tracto vocal. A localização desse estrangulamento determina os diferentes sons fricativos, que no caso deste sistema são nove.

Símbolo	Exemplo
/f/	*FAR*
/th/	*MONTH*
/s/	*ITS*
/sh/	*SHE*
/hh/	*HAS, WHO*

Símbolo	Exemplo
/v/	*NEVER*
/dh/	*THE*
/z/	*WAS*
/zh/	*VISUAL*

Tab. 2.4 - Unidades acústicas correspondentes a consoantes fricativas.

As consoantes fricativas correspondentes aos símbolos /f/, /th/, /s/ e /sh/ são designadas não-vocálicas porque resultam da excitação associada apenas à turbulência no local de constrição. Na parte da direita da tabela 2.4 são apresentadas as respectivas consoantes fricativas vocálicas. Estas resultam da excitação provocada pela turbulência do ar no local de constrição e também pelo fluxo periódico de ar devido à vibração da glote.

Consoantes oclusivas

As consoantes oclusivas são sons não contínuos que resultam da libertação brusca de energia acumulada pela pressão que o ar exerce num estrangulamento total localizado no tracto vocal. Podem ser identificadas três fases na produção destes sons: 1) aumento da pressão do ar no local de estrangulamento; 2) breve período fricativo devido à libertação brusca do ar; 3) breve período durante o qual ocorre um fluxo contínuo de ar no tracto vocal antecipando a produção da vogal subsequente.

Símbolo	Exemplo
/p/	*OPEN*
/t/	*STEAM*
/k/	*TOKYO*

Símbolo	Exemplo
/b/	*ABOVE*
/d/	*DALE*
/g/	*FIGURE*

Tab. 2.5 - Unidades acústicas correspondentes a consoantes oclusivas.

Na parte esquerda da tabela 2.5 são apresentadas as consoantes oclusivas não vocálicas e na parte direita as vocálicas. A diferença mais significativa entre estes dois tipos de sons consiste no nível de vibração das paredes do tracto vocal[1] que ocorre nas consoantes oclusivas vocálicas durante a fase inicial ou imediatamente após a libertação do ar.

Símbolo	Exemplo
/pcl/	*OPEN* /ow/ /pcl/ /p/ /ax/ /n/ *DEPTHS* /dcl/ /d/ /eh/ /pcl/ /s/ *LAMPS* /l/ /ae/ /m/ /p/ /s/
/tcl/	*TWO* /tcl/ /t/ /uw/ *WHICH* /w/ /ih/ /tcl/ /ch/ *HECTOR* /hh/ /eh/ /kcl/ /t/ /axr/
/kcl/	*ECHO* /q/ /eh/ /kcl/ /k/ /ow/ *HECTOR* /hh/ /eh/ /kcl/ /t/ /axr/ *BANGKOK* /bcl/ /b/ /ae/ /ng/ /k/ /aa/ /kcl/ /k/
/bcl/	*BAY* /bcl/ /b/ /ey/ *SUBMARINE* /s/ /ah/ /bcl/ /m/ /axr/ /iy/ /n/ *NUMBER* /n/ /ah/ /m/ /b/ /axr/
/dcl/	*DAY* /dcl/ /d/ /ey/ *WOULD* /w/ /uh/ /dcl/ *ENDING* /q/ /eh/ /n/ /d/ /ih/ /ng/
/gcl/	*GAS* /gcl/ /g/ /ae/ /s/ *CODAG* /kcl/ /k/ /ow/ /dcl/ /d/ /ae/ /gcl/ *LONGER* /l/ /ao/ /ng/ /g/ /axr/

Tab. 2.6 - Extensão do conjunto das unidades acústicas correspondentes a consoantes oclusivas.

Neste sistema são utilizadas outras unidades acústicas, para além das mencionadas na tabela 2.5, destinadas também à modelação de consoantes oclusivas. São em número de seis e foram introduzidas na consequência de resultados[2] atestando

[1] Isso é facilmente perceptível, por exemplo, colocando uma mão em contacto com a garganta.
[2] Resultados obtidos pelo grupo de trabalho do laboratório de processamento da fala ELIS.

a vantagem que apresentam nesta tarefa de reconhecimento ao permitirem modelar melhor as consoantes oclusivas. Essas unidades, apresentadas na tabela 2.6, geralmente emparelham com as anteriores (tabela 2.5), tal como reflecte a notação simbólica utilizada[1]. No entanto, como evidenciam os exemplos apresentados, o seu conteúdo acústico é significativamente diferente. Um traço comum dessa diferença reside no facto de estas últimas corresponderem geralmente à fase inicial das consoantes oclusivas.

Para cada unidade são apresentados três exemplos. O primeiro corresponde à realização clara e completa da consoante oclusiva, sendo então utilizado o par /x-closure/, /x/ na transcrição fonética do som. O segundo exemplo corresponde geralmente a uma situação em que a pronunciação da consoante é débil e incompleta, apresentando apenas a parte inicial do som. O terceiro exemplo ilustra ocorrências em que a parte inicial do som, durante o período de oclusão, é fortemente afectada pelo fonema imediatamente anterior, como sucede, com relativa frequência, no caso das consoantes nasais.

Quando as consoantes oclusivas ocorrem entre vogais, a sua realização acústica pode alterar-se de maneira razoavelmente independente do tipo das vogais vizinhas. Neste sistema é utilizada a unidade acústica com o símbolo /dx/, relativamente à qual se apresentam alguns exemplos na tabela 2.7, para modelar este fenómeno significativo na base de dados *Resource Managment*.

Símbolo	Exemplos
/dx/	*ADDED* /q/ /ae/ /dx/ /ax/ /dcl/ *BIDDLE* /bcl/ /b/ /ih/ /dx/ /el/ *CITY* /s/ /ih/ /dx/ /iy/ *EIGHTY* /q/ /ey/ /dx/ /iy/ *FOOTER* /f/ /uh/ /dx/ /axr/ *LETTER* /l/ /eh/ /dx/ /axr/ *RADAR* /r/ /ey/ /dx/ /aa/ /r/

Tab. 2.7 - Extensão do conjunto das unidades acústicas correspondentes a consoantes oclusivas.

Consoantes africadas

Durante a emissão destes sons ocorre uma transição entre duas consoantes, com origem numa do tipo oclusiva e com destino numa fricativa.

[1] Por exemplo, ao símbolo /b/ corresponde o símbolo /bcl/ (ou /b-closure/).

Símbolo	Exemplo
/ch/	*CHINA*
/jh/	*JUNE*

Tab. 2.8 - Unidades acústicas correspondentes a consoantes africadas.

Outras unidades acústicas elementares

Refira-se que neste sistema são utilizadas ainda mais duas classes acústicas para além das já referidas. Uma delas consiste na oclusão da glote, relativamente à qual são na tabela 2.9 apresentados alguns exemplos.

Símbolo	Exemplos
/q/	*ADD* /q/ /ae/ /dcl/ *EIGHTH* /q/ /ey/ /tcl/ /th/ *OFF* /q/ /ao/ /f/

Tab. 2.9 - Unidadade acústica correspondente à oclusão da glote.

A "52ª" unidade acústica, notado pelo símbolo /-/, é definida para modelar os períodos de silêncio, ou, mais exactamente, de quase-silêncio.

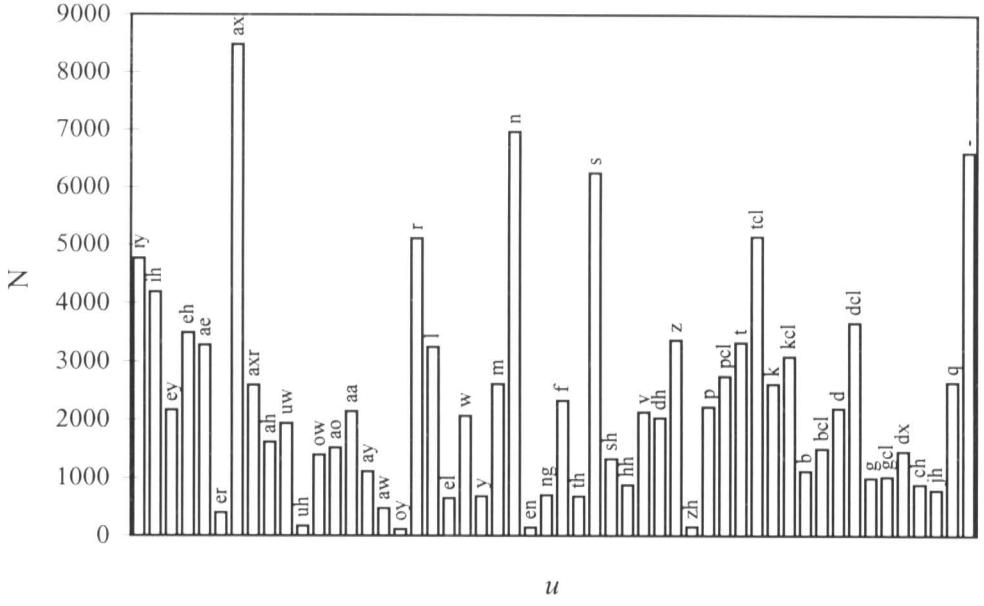

Fig. 2.13 - Frequência das unidades acústicas na base de dados de treino.

A figura 2.13 mostra o número de ocorrências de cada unidade acústica, estimada por um procedimento de segmentação automática, na base de dados utilizada no treino do sistema.

2.6.3 Modelos escondidos de Markov

2.6.3.1 Introdução

A questão que se coloca pode ser apresentada da seguinte maneira: como construir um modelo estatístico M, associado a uma unidade elementar de fala definida ao nível fonético, referenciada pelo símbolo u, que determine a probabilidade de uma sucessão de vectores de características, $\boldsymbol{X}_1^T$, corresponder à realização acústica de u, isto é, calcule correctamente $P\left(\boldsymbol{X}_1^T | u\right)$?

Por "correctamente" entende-se, no essencial, que o modelo deve apresentar valores de probabilidade elevados sempre que observa segmentos do sinal correspondentes à sua classe e, além disso, valores de probabilidade reduzidos para realizações das restantes classes.

Nesta secção irá ser tratada a primeira parte da resposta à questão apresentada, isto é, a selecção de um formalismo estatístico para modelar u, apenas numa secção posterior (§2.9) será discutido o processo de treino dos parâmetros definidos. No entanto, é essencial referir aqui que neste sistema os modelos não são treinados com base num critério discriminativo, pelo que se acentua a exigência quanto à sua precisão. Significa isto que não sendo os modelos directamente treinados para "observar mal" as realizações de outras classes, então devem ser capazes de modelar com acrescida precisão as ocorrências da sua própria classe, de maneira a ser legítimo esperar o aumento da distância entre o modelo correcto e os seus rivais e portanto permitir que o modelo acústico, no seu todo, mantenha uma capacidade discriminativa aceitável.

Admitindo que se opta pelos modelos escondidos de Markov (HMM), por exemplo se se pretender modelar um fonema cuja realização acústica apresenta muitas alterações ao longo do tempo, então, o número de estados do respectivo modelo deve ser grande. Ou se o sistema é utilizado para falantes com características acústicas muito diversas então o espaço acústico deve ser coberto por um grande número de funções de Gauss.

Claro que o aumento da precisão dos modelos implica geralmente uma maior complexidade e um sistema com mais parâmetros livres, portanto exige maior potência de cálculo e também acarreta a falta de robustez dos modelos se a base de dados não acompanhar esse aumento.

É geralmente aceite que os HMM constituem actualmente a base tecnológica dominante no reconhecimento de fala contínua [19,118], estabelecendo um razoável compromisso entre a precisão da modelação e a complexidade do sistema. Esta conclusão é sobretudo sustentada pelos resultados práticos obtidos em inúmeros reconhecedores baseados nos HMM.

Este foi, em termos gerais, o argumento mais relevante que desde o início pesou na escolha dos HMM para a modelação acústica neste sistema. Deve no entanto referir-se que antes de iniciar o desenvolvimento deste sistema foi feito algum trabalho, essencialmente ao nível da formação, no sentido de adquirir sensibilidade na utilização de outras tecnologias com aplicação directa na modelação acústica.

Uma das abordagens experimentadas baseia-se num paradigma das redes neuronais artificiais (ANN), que embora não sejam geralmente consideradas formalismos estatísticos em sentido mais estrito, são efectivamente capazes de captar e representar dependências estatísticas que se verificam no sinal da fala [104].

Foram desenvolvidos alguns protótipos, utilizando a tecnologia conhecida por *perceptrão* de múltiplas camadas (MLP), para uma aplicação simples consistindo no reconhecimento de dígitos isolados. Em algumas versões foram utilizados mapas de Kohonen integrados em módulos de pré-classificação e redução da dimensão do espaço acústico inicial.

No decurso desse trabalho resultou também a apresentação de uma generalização do algoritmo *backpropagation* de treino dos MLP [91]. Essa generalização baseia-se no algoritmo-ε [197], que por sua vez é também uma extensão do procedimento Δ^2d'Aitken.

Esta breve incursão pelos MLP foi naturalmente impulsionada pelas vantagens, relativamente aos HMM, que têm sido apontadas [7,96,105,106,107] e entre as quais se destacam: o critério de treino *standard* dos MLP é discriminativo, existindo diversas técnicas [100,101] capazes de acelerar significativamente esse processo tornando-o suportável mesmo em sistemas de grande dimensão[1]; em princípio, os MLP podem estabelecer qualquer transformação não linear da entrada, o que é muito importante nos reconhecedores da fala, cujas superfícies de decisão apresentam uma forma muito complexa; os MLP, e os ANN em geral, dispensam grande parte das assunções sobre diferentes componentes do sistema que em outras abordagens tecnológicas é necessário formular previamente; os MLP podem acomodar muito facilmente entradas contextuais e de realimentação.

[1] A estrutura fortemente paralela e regular dos ANN favorece também arquitecturas dedicadas que podem ser com relativa facilidade realizadas em *hardware*.

A principal desvantagem dos ANN no reconhecimento automático da fala é a inaptidão em suportar adequadamente a dimensão temporal do sinal da fala, onde é necessário classificar padrões de comprimento variável e apresentando distorções no alinhamento temporal. Têm sido apresentadas algumas soluções para esse problema [108,109], como é o exemplo dos TDNN[1], mas apesar do seu comprovado potencial em muitas aplicações [95,102], o certo é que não se têm revelado uma solução inteiramente consistente em sistemas de maior dimensão [7].

Um especial destaque deve ser dado a uma abordagem híbrida que integra a tecnologia dos HMM, na modelação temporal do sinal, com a dos ANN, na modelação local do espaço acústico, em substituição das tradicionais misturas gausseanas [103] e permitindo uma utilização mais eficaz da informação de contexto. Esses sistemas têm obtido resultados assinaláveis, capazes de superar os sistemas baseados apenas nos HMM [7].

Uma outra abordagem que mereceu atenção na fase inicial deste trabalho, tendo então sido ponderada a hipótese de a considerar na modelação acústica deste sistema, são os chamados modelos estocásticos de segmentos (SSM) [97].

Estes modelos são uma generalização dos HMM e constituem um formalismo capaz de modelar de forma explícita segmentos, ou sequências de vectores de características, do sinal da fala. Esta aptidão é de grande relevância, pois é reconhecida a incapacidade dos HMM *standard* modelarem convenientemente as dependências estatísticas entre os sucessivos vectores de características, o que é inerente à assunção básica de independência condicional das observações, dada a sequência de estados (§2.6.3.2). Apesar do grande número de extensões dos HMM que têm sido propostas para solucionar esse problema [22,38,76,77,126,127,128,129,131,132], o certo é que este continua ainda a constituir uma das suas principais limitações.

A maior desvantagem dos SSM é o grande número de parâmetros livres do sistema e a sua grande complexidade computacional. Apesar da utilização de técnicas sofisticadas de partilha de parâmetros e também de aceleração dos algoritmos de pesquisa e reconhecimento [98], o treino dos modelos e a tarefa do reconhecimento apresentam dificuldades muito maiores comparando com os HMM.

Face ao que se acaba de expor e atendendo também aos objectivos principais deste sistema, anteriormente referidos (§2.1), os HMM foram seleccionados para a modelação acústica. Outro factor que também influenciou esta escolha foi o facto de na altura em que se iniciou o desenvolvimento deste sistema já existir um maior conhecimento teórico e também prático relativamente a esta tecnologia. No essencial essa experiência havia sido adquirida através do desenvolvimento de alguns protótipos,

[1] Do Inglês, *time delay neural network*.

para o reconhecimento de dígitos isolados e também dígitos ligados, baseados nos HMM do tipo discreto, semi-contínuo e contínuo.

2.6.3.2 Formalização dos modelos

Um modelo escondido de Markov pode ser entendido como uma máquina de estados finita à qual estão associados dois processos estocásticos concorrentes [2].

Cadeia de Markov

Um dos processos estocásticos, frequentemente designado cadeia de Markov, é aqui representado pela variável aleatória q_t. Esta variável está definida para valores de t discretos e o seu valor identifica, através de um número natural, um dos S estados do modelo:

$$q_t \in \Gamma_q, \qquad t = 1,2,\ldots,T \qquad (2.25)$$
$$\Gamma_q = \{1,2,\ldots,S\}$$

Pode então afirmar-se que no instante t a cadeia de Markov se encontra no estado q_t.

A probabilidade de uma cadeia de Markov de primeira ordem se encontrar, num qualquer instante $t = 1,2,\ldots,T$ no estado genérico j, depende apenas do estado em que se encontra no instante imediatamente anterior. Ficam assim definidas as probabilidades de transição entre estados:

$$a_{i,j} = P(q_t = j | q_{t-1} = i) \qquad (2.26)$$

É conveniente arbitrar mais dois estados que servem apenas como um artifício para definir facilmente o início e o fim do processo. Considera-se que imediatamente antes de iniciar o processo da cadeia de Markov esta encontra-se no estado "0", $q_0 = 0$, logo, da equação (2.26) resulta que a probabilidade do modelo ocupar o estado k no instante inicial[1] é $a_{0,k} = P(q_1 = k)$. Imediatamente depois de terminar o processo da cadeia de Markov esta encontra-se no estado $q_{T+1} = S + 1$, pelo que apenas os

[1] É importante notar que por enquanto apenas se está a considerar a cadeia de Markov, independentemente da "observação" dos vectores de características.

estados com probabilidade de transição $a_{k,S+1}$ não nula podem ocupar o estado k no instante final.

As probabilidades definidas na equação (2.26) obedecem à seguinte condição[1]:

$$\sum_{j=1}^{S+1} a_{i,j} = 1, \quad i = 0,1,\ldots,S \tag{2.27}$$

Uma vez que a cadeia de Markov é invariante no tempo, estas probabilidades de transição entre os estados são estacionárias.

As probabilidades de transição são geralmente organizadas na forma matricial, dando então origem à matriz (S+2)x(S+2) das probabilidades de transição, $A = \left[a_{i,j}\right]$. Esta matriz estabelece a topologia da cadeia de Markov, por exemplo se $a_{I,J} = 0$ significa que a cadeia não apresenta uma transição directa com origem no estado I e destino no estado J.

Associado à cadeia de Markov surge o conceito de caminho, ou sucessão de estados:

$$Q^g = \left\{q_1^g, q_2^g, \ldots, q_T^g\right\}, \quad q_t^g \in \Gamma_q \tag{2.28}$$

A probabilidade de um caminho particular, Q^g, pode ser assim calculada:

$$P\left(Q^g\right) = \prod_{t=1}^{T+1} a_{q_{t-1}^g, q_t^g} \tag{2.29}$$

A equação (2.29) evidencia uma característica da cadeia de Markov de 1ª ordem, com implicações importantes na modelação da estrutura temporal do sinal da fala, que consiste no facto da forma funcional do modelo de permanência num estado ser exponencial:

$$P\left(q_t = q_{t+1} = \ldots = q_{t+\tau} = i\right) = a_{i,i}^{\tau} \tag{2.30}$$

[1] Além de, naturalmente, terem um valor positivo.

Densidades de probabilidade de observação

O outro processo estocástico modela a verosimilhança associada às "observações" locais que o modelo escondido[1] de Markov realiza sobre o sinal da fala, cuja sucessão de vectores de características, $\boldsymbol{X}_1^T$, adere à cadeia de Markov. Assim, para cada estado da cadeia de Markov está definida uma função densidade de probabilidade, ou verosimilhança, de um vector de características ser observado por esse estado[2]:

$$b_j(\boldsymbol{x}_t) = p(\boldsymbol{x}_t | q_t = j) \tag{2.31}$$

É importante realçar que nos modelos escondidos de Markov *standard* as densidades de probabilidade de observação são condicionalmente independentes das observações anteriores, isto é:

$$p(\boldsymbol{x}_t | q_t = j, X_1^{t-1}) = p(\boldsymbol{x}_t | q_t = j) \tag{2.32}$$

Esta assunção de independência acarreta uma limitação acentuada da memória dos modelos e portanto uma deficiente modelação da correlação existente entre os sucessivos vectores de características.

Em contrapartida, reduz muito significativamente a complexidade dos modelos. A verosimilhança da sucessão de observações associada a um caminho conhecido, Q^g, é dada por:

$$p(\boldsymbol{X}_1^T | Q^g) = \prod_{t=1}^{T} b_{q_t^g}(\boldsymbol{x}_t) \tag{2.33}$$

Em qualquer estado, a função densidade de probabilidade é frequentemente definida combinando linearmente funções de Gauss[3]:

$$b_j(\boldsymbol{x}_t) = \sum_m c_{m,j} N(\boldsymbol{x}_t; \mu_{m,j}, \Sigma_{m,j}) \tag{2.34}$$

[1] O termo "escondido", em HMM, reflecte o facto do modelo observar a sucessão de vectores de características, emitindo a respectiva verosimilhança, mas sem explicitar a cadeia de Markov.

[2] É assim nos modelos escondidos de Markov do tipo Moore, os quais apresentam alguma economia na parametrização relativamente aos modelos do tipo Mealy, onde as observações são realizadas na transição entre os estados.

[3] Não é assim, por exemplo, nos HMM do tipo discreto.

Verosimilhança total

A densidade de probabilidade conjunta do modelo de Markov observar uma sucessão $\boldsymbol{X}_1^T$ e utilizar um caminho particular Q^g com comprimento T nessa observação pode ser obtida a partir dos resultados correspondentes às equações (2.29) e (2.33):

$$p\left(\boldsymbol{X}_1^T, Q^g\right) = \prod_{t=1}^{T}\left[a_{q_{t-1}^g, q_t^g} b_{q_t^g}\left(\boldsymbol{x}_t\right)\right] a_{q_T^g, S+1} \tag{2.35}$$

A densidade probabilística do modelo de Markov observar a sucessão $\boldsymbol{X}_1^T$ é então obtida somando o resultado da equação (2.35) para todos os caminhos possíveis, $Q^g \in Q^*$:

$$p\left(\boldsymbol{X}_1^T\right) = \sum_{Q^g \in Q^*} p\left(\boldsymbol{X}_1^T, Q^g\right) \tag{2.36}$$

A utilização directa da equação (2.36) para calcular $p\left(\boldsymbol{X}_1^T\right)$ não é de modo algum praticável porque o número de caminhos possíveis é muito grande, apresentando uma dependência exponencial relativamente a T.

Algoritmo *"forward"*

O algoritmo *"forward"* permite calcular eficientemente $p\left(\boldsymbol{X}_1^T\right)$. No essencial consiste num procedimento baseado no cálculo recursivo da probabilidade conjunta do modelo observar a sucessão parcial dos vectores de características $\boldsymbol{X}_1^t = \left\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_t\right\}$ e de se encontrar no estado j no instante t, ou seja, a probabilidade *"forward"* é assim definida:

$$\alpha_{j,t} = p\left(\boldsymbol{X}_1^t, q_t = j\right) \tag{2.37}$$

A iniciação do algoritmo *"forward"* é executada pela instrução:

$$\alpha_{j,1} = b_j\left(\boldsymbol{x}_t\right) a_{0,j} \quad \text{para} \quad j=1,\ldots,S \tag{2.38}$$

Então, é recursivamente executada a instrução:

$$\alpha_{j,t} = b_j(\boldsymbol{x}_t)\sum_{i=1}^{S}\alpha_{i,t-1}a_{i,j} \quad \text{para} \quad t=2,...,T \quad \text{e} \quad j=1,...,S \tag{2.39}$$

Terminada a sucessão de vectores de característica, basta extrair a probabilidade *"forward"* no estado final da cadeia de Markov:

$$p\left(\boldsymbol{X}_1^T\right) = \alpha_{S+1,T+1} = \sum_{i=1}^{S}\alpha_{i,T}a_{i,S+1} \tag{2.40}$$

Apesar da simplicidade deste procedimento, a sua realização acarreta algumas dificuldades numéricas[1].

2.6.3.3 Aplicação dos modelos

Cada uma das 52 unidades elementares da fala em que se baseia este sistema (§2.6.2.2) é representada pelo respectivo modelo escondido de Markov. Esses modelos acústicos elementares, M^i, calculam o resultado definido em (2.36), isto é, as respectivas verosimilhanças na observação de $\boldsymbol{X}_1^T$:

$$\mathsf{M}^i: \quad \boldsymbol{X}_1^T \rightarrow p\left(\boldsymbol{X}_1^T \middle| u^i\right), \qquad \forall u^i \in \Gamma_u \tag{2.41}$$

Os HMM dispensam a caracterização prévia e completa da estrutura acústica do sinal que modelam. Significa isto que é possível em alguma medida modelar essa informação, relativamente a uma base de dados acústicos, utilizando procedimentos automáticos de treino.

Contudo, como facilmente se depreende a partir da formalização destes modelos, é vantajoso utilizar algum conhecimento apriorístico sobre o sinal a modelar. Basta ter presente que o processo estocástico associado a um HMM pode ser interpretado como uma distribuição estatística estacionária em cada estado mas variável ao deslocar-se de um estado para outro com determinada probabilidade de transição. Por outro lado, relativamente ao sinal da fala, em particular às unidades fonéticas elementares aqui definidas, ele pode considerar-se quase-estacionário por partes.

Assim, duas questões fundamentais no desenvolvimento dos HMM são: 1) como definir a topologia da cadeia de Markov para conseguir modelar da melhor maneira a variabilidade e a estrutura temporal do sinal associado aos modelos acústicos? 2) em

[1] Esses problemas serão discutidos aquando da apresentação dos procedimentos de descodificação e treino destes modelos.

cada estado, como modelar a distribuição estatística correspondente às diferentes realizações acústicas do sinal que lhe está associado? As soluções concretas, seguidas neste sistema, para os problemas que estas duas questões levantam são apresentados seguidamente.

Topologia da cadeia

Todas as unidades acústicas elementares, à excepção do *silêncio*, são neste sistema modeladas por um HMM cuja topologia está representada na figura 2.14.

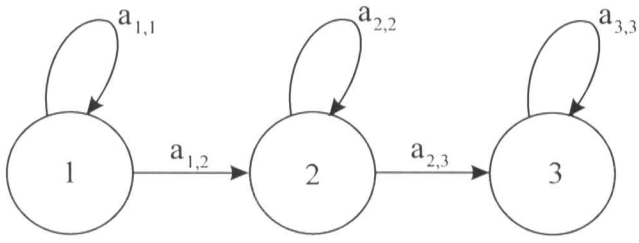

Fig. 2.14 - Cadeia de Markov utilizada na modelação fonética.

A cadeia de Markov apresenta uma estrutura muito simples, com apenas três estados sequenciais sem permitir a transição directa do estado inicial para o final. A intenção subjacente a esta escolha é, obviamente, utilizar os estados 1, 2 e 3 para observar, respectivamente, os períodos inicial, intermédio e final da unidade fonética. Esta topologia assume uma duração mínima dos fonemas correspondente a 3 vectores de características, restrição que se verifica ser razoável.

As probabilidades de transição assinaladas na representação esquemática são treinadas de acordo com a equação 2.26, isto é, definindo uma cadeia de Markov de primeira ordem. A hipótese de utilizar modelos de Markov de ordem mais elevada não foi sequer ponderada, apesar dos resultados publicados que atestam vantagens na sua utilização [76], devido à complexidade significativamente maior que acarretam.

Em relação ao *silêncio*, foi assumido que esta classe acústica não apresenta qualquer estrutura temporal, pelo que se utilizou um modelo constituído por apenas um estado, tal como mostra a figura 2.15.

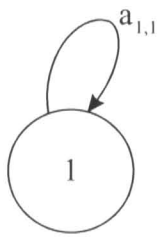

Fig. 2.15 - "Cadeia" de Markov utilizada na modelação do silêncio.

Como é natural, têm sido investigadas diferentes topologias com o objectivo de aumentar a capacidade de modelação dos HMM.

Estudos publicados [62,63,64] registam a obtenção de resultados positivos através da utilização de procedimentos automáticos, ou semi-automáticos, capazes de optimizar a topologia dos HMM com base em determinados critérios objectivos.

Uma abordagem diferente, assente no conhecimento fonético da língua e em métodos empíricos, tem sido seguida em diversos sistemas, baseados também em unidades elementares definidas ao nível do fonema, conduzindo a cadeias de Markov relativamente complexas capazes de modelar com bastante precisão realizações acústicas alternativas. Um exemplo particularmente interessante, por ter sido desenvolvido para uma aplicação idêntica a esta e utilizando também a mesma base de dados, são os modelos acústicos considerados no sistema SPHINX [59].

É frequentemente assinalada a incapacidade dos HMM *standard* modelarem correctamente a duração dos períodos de quase-estacionaridade que se verificam no sinal da fala. Com efeito, admitindo que cada um desses períodos está associado a um estado do HMM, a sua duração fica afectada pela probabilidade de permanência nesse estado, a qual, de acordo com a equação 2.30, cai exponencialmente com o tempo. No entanto, a distribuição estatística real da duração desses períodos não obedece geralmente a essa forma funcional.

Nas abordagens baseadas nos chamados modelos escondidos semi-Markov [65] este problema é resolvido substituindo as probabilidades de transição entre estados pela modelação explícita das distribuições estatísticas de permanência em cada estado. Contudo, esta técnica apresenta uma grande complexidade [66,74], em particular quando não se recorre à parametrização desses modelos de permanência, pelo que é frequentemente preterida.

De entre as diversas outras soluções que têm sido apresentadas para solucionar este problema [71,72,73,75,126,133,134], uma das que se tem revelado mais eficaz, sendo também muito simples, é baseada na alteração da topologia da cadeia de Markov original. Cada estado do modelo original é expandido numa cadeia de Markov geralmente com estrutura semelhante à apresentada na figura 2.14, embora possa ter um número diferente de estados, os quais partilham a densidade probabilística de

observação. Apesar de cada novo estado continuar a apresentar um modelo de permanência exponencial, o conjunto de estados expandidos pode modelar uma distribuição que aproxima melhor a que se verifica na realidade [67].

Esta solução apresenta também a vantagem de forçar um minorante adequado para o período de permanência num estado expandido, pois, tal como é frequentemente notado [19], as probabilidades de transição têm uma influência reduzida comparativamente às densidades de probabilidade de observação. Por exemplo, se se pretender seleccionar o estado j^* que num instante genérico apresenta a maior probabilidade conjunta de ocupação e observação de um vector de características, conhecendo o estado anterior, ou seja, $j^* = \underset{j}{argmax}\left\{p\left(\boldsymbol{x}_t, q_t = j | q_{t-1} = i\right)\right\}$, o principal factor é geralmente $b_j\left(\boldsymbol{x}_t\right)$, pois a sua gama dinâmica é bastante maior do que a das probabilidades $a_{i,j}$.

Apesar de alguma redução das taxas de erro que previsivelmente ocorreria em resultado da introdução desta simples alteração na topologia do HMM, a qual não acarretaria alterações significativas nos procedimentos de treino dos modelos e de descodificação, foi decidido manter uma topologia o mais elementar possível, utilizando os HMM do tipo Bakis com apenas três estados tal como apresentado na figura 2.14.

Densidade da probabilidade de observação

Os HMM que modelam as unidades acústicas deste sistema são do tipo semi-contínuo (SCHMM) [86] e utilizam misturas baseadas em funções de Gauss com a matriz das covariâncias diagonal.

Assim, a distribuição estatística dos pontos definidos no espaço de representação acústica do sinal e correspondentes à realização de todas as classes de sons é parametrizada com base num conjunto de funções gausseanas partilhado por todos os estados na modelação da densidade probabilística de observação. Neste sistema está então definido um conjunto de pares $\left(\mu, \sigma\right)$, geralmente designado *codebook*:

$$\left\{\left(\mu_m, \sigma_m\right),\ m = 1,\ldots,M\right\},\quad com\ \ M = 4096, \tag{2.42}$$

$$\mu_m = \left(\mu_m^1, \mu_m^2, \ldots, \mu_m^D\right),$$

$$\sigma_m = \left(\sigma_m^1, \sigma_m^2, \ldots, \sigma_m^D\right),$$

$$D = 39$$

consistindo nos vectores das médias, μ, e das variâncias, σ, que caracterizam o respectivo conjunto de funções de Gauss:

$$\left\{N_m(\boldsymbol{x}),\ m=1,\ldots,M\right\},\quad com\quad \boldsymbol{x}\in\Re^D, \tag{2.43}$$

$$N_m(\boldsymbol{x}) = k_m exp\left\{-\frac{1}{2}\sum_{i=1}^{D}\frac{\left(x^i-\mu_m^i\right)^2}{\left(\sigma_m^i\right)^2}\right\},$$

$$k_m = (2\pi)^{-D/2}\left|\prod_{i=1}^{D}\left(\sigma_m^i\right)^2\right|^{-1/2}$$

A função de Gauss é geralmente utilizada, dada a sua comprovada consistência como forma funcional básica para modelar localmente as reais distribuições acústicas [82]. Contudo, alguns sistemas utilizam outras funções, por exemplo a função de Laplace [83,87]. Refira-se que na sua forma mais geral a função de Gauss utiliza uma matriz das covariâncias, Σ_m, ao invés do vector das variâncias:

$$N_m(\boldsymbol{x}) = \frac{1}{(2\pi)^{D/2}|\Sigma_m|^{1/2}}e^{-\frac{1}{2}(\boldsymbol{x}-\mu_m)\Sigma_m^{-1}(\boldsymbol{x}-\mu_m)^t} \tag{2.44}$$

No entanto, considerando que a correlação entre as componentes do vector de características, $\boldsymbol{x}$, é muito pequena (§2.5.2.6), é possível aproximar os elementos não diagonais das matrizes covariâncias por zero, isto é, $\sigma_m^{i,j}=0,\ \forall i\neq j$, resultando portanto o conjunto de gausseanas definido em (2.43).

Em cada estado de qualquer dos SCHMM, a função densidade de probabilidade de observação é obtida combinando linearmente as funções de Gauss definidas pelo *codebook*:

$$b_j(\boldsymbol{x}_t) = \sum_{m=1}^{M} c'_{j,m}N_m(\boldsymbol{x}_t) \tag{2.45}$$

Verifica-se que a maior parte dos estados[1] utiliza com um peso não desprezável só uma pequena fracção do *codebook*. Em média, cada estado utiliza efectivamente menos de 10% do total de 4096 funções de Gauss. Apenas existe um número restrito de estados mais abrangentes como, por exemplo, o modelo do *silêncio*, o qual utiliza aproximadamente 25% do *codebook*. Assim, o número de termos do somatório na equação (2.45) pode ser muito reduzido considerando apenas as funções de Gauss

[1] Existem 154 estados: 51(fonemas)X3+1(silêncio).

necessárias na modelação da verosimilhança da observação de $\boldsymbol{x}_t$ nesse estado. Isso pode ser feito definindo em cada estado apontadores, $\chi_{j,m}$, para as funções de Gauss necessárias e pesando-as através dos respectivos coeficientes, $c_{j,m}$, da mistura gausseana, os quais se encontram normalizados. Desse modo, a função densidade de probabilidade de observação em cada estado passa a escrever-se:

$$b_j(\boldsymbol{x}_t) = \sum_{m=1}^{M_j} c_{j,m} N_{\chi_{j,m}}(\boldsymbol{x}_t), \qquad \text{com} \quad \sum_{m=1}^{M_j} c_{j,m} = 1 \tag{2.46}$$

Têm sido experimentados diferentes métodos para determinar a dimensão das misturas gausseanas, por exemplo [159] analisando a contribuição de cada estado nos erros de reconhecimento e então aumentando a dimensão das misturas nos estados cujos respectivos segmentos do sinal são frequentemente "invadidos" por outros estados e, pelo contrário, reduzindo as misturas nos estados "invasores".

Em seguida são resumidamente abordados, essencialmente na perspectiva deste sistema e baseando-se na comparação com outras variantes dos HMM *standard* também muito utilizadas, alguns dos aspectos importantes que conduziram à decisão de utilizar os SCHMM [85,86].

Uma dessas variantes consiste nos chamados HMM contínuos (CHMM) [6], que se caracterizam, em cada estado, pela modelação paramétrica da densidade probabilística de observação através de uma mistura de funções básicas, geralmente funções de Gauss, exclusivas desse estado, isto é, $b_j(\boldsymbol{x}_t) = \sum_{m=1}^{M_j} c_{j,m} N_{j,m}(\boldsymbol{x}_t)$.

A outra variante é formada pelos HMM discretos (DHMM) [6], os quais se baseiam na partição do espaço acústico num conjunto de sub-domínios disjuntos. Cada sub-domínio é representado por um vector *centróide*, υ_k, cujo conjunto é geralmente designado *codebook*. O vector de características que se pretende observar, $\boldsymbol{x}_t$, é primeiro associado ao respectivo sub-domínio, ou seja, através do processo de quantificação vectorial é-lhe atribuído o índice do *centróide* mais próximo segundo uma determinada medida de distância, $k = \arg\min_k \{dist(\boldsymbol{x}_t, \upsilon_k)\}$. Finalmente, a verosimilhança de observar $\boldsymbol{x}_t$ no estado j toma o valor do peso do sub-domínio k nesse estado, isto é, $b_j(\boldsymbol{x}_t) = c_{j,k}$.

Comparativamente aos CHMM, os SCHMM [88] partilham as distribuições gausseanas usadas na modelação das densidades de probabilidade de observação permitindo assim, em princípio, reduzir o número de parâmetros livres que é necessário treinar e a complexidade computacional do sistema sem penalizar significativamente o desempenho do reconhecedor [120,123]. No entanto, a

concretização deste princípio está dependente da dimensão da base de dados disponível para treinar o sistema. Por exemplo, tem-se verificado [90,83,88,122] alguma superioridade no desempenho dos sistemas baseados nos CHMM[1] quando o conjunto de dados de treino é suficiente para obter modelos acústicos precisos e robustos. Pelo contrário, quando o treino robusto dos modelos é um problema crítico, como sucede frequentemente, os SCHMM permitem geralmente melhores resultados [89,124].

Relativamente a este sistema, o número de unidades acústicas elementares é reduzido, pelo que é de admitir que a utilização dos CHMM não acarretaria dificuldades de treino utilizando a base de dados *Resource Managment*. Contudo, é previsível que a potencial vantagem dos CHMM relativamente aos SCHMM, assente na modelação mais precisa das distribuições estatísticas em cada estado, não se traduziria na prática por qualquer redução significativa das taxas de erro, pois os modelos acústicos usados neste sistema são relativamente difusos (§2.6.2) devido à não utilização de modelação com contexto.

Se o sistema utilizasse fonemas modelados com informação de contexto então o número de unidades acústicas elementares seria muito maior e portanto, pelas razões já apontadas, os CHMM poderiam apresentar algumas dificuldades de treino. Ainda considerando esta hipótese, a utilização dos CHMM obrigaria ao uso de misturas relativamente pequenas em cada estado, cujas consequências seriam agravadas no caso das funções de Gauss apresentarem matriz das covariâncias diagonal. Note-se que os SCHMM, mesmo dispensando a matriz das covariâncias completa, poderiam aproximar mais facilmente a distribuição real associada a cada estado ao utilizarem misturas muito maiores em virtude de partilharem as funções de Gauss.

Comparando com os DHMM [88], têm sido apontados diversos argumentos favoráveis aos SCHMM. Um dos mais relevantes reside no erro associado ao processo de quantificação vectorial nos DHMM, pois a um determinado vector de características é atribuída uma verosimilhança única independentemente da sua localização no interior de determinado sub-domínio. É possível reduzir o erro de quantificação aumentando a dimensão do *codebook*, mas esta medida só pode ser aplicada até certo limite pois acarreta maiores custos de computação. Uma solução que se tem revelado mais satisfatória na redução desta distorção consiste na utilização de múltiplos *codebooks*, cada qual com dimensão suportável mas suficiente para cobrir com razoável precisão o respectivo sub-espaço acústico [59]. Essa solução foi aplicada com sucesso num sistema anteriormente desenvolvido para reconhecer palavras isoladas. Contudo, esta solução também apresenta inconvenientes, nomeadamente uma maior exigência de espaço de memória e, sobretudo, o facto das distribuições estatísticas nesses sub-espaços serem geralmente consideradas independentes entre si, de maneira a facilitar o processamento, apesar desta assunção não se verificar na realidade.

[1] Facto referido num número significativo de publicações consultadas tratando este assunto.

Uma outra desvantagem dos DHMM resulta do facto do *codebook* ser geralmente estabelecido no início, não sendo submetido a qualquer processo de adaptação durante o treino dos restantes parâmetros do sistema, ao contrário do que sucede nos SCHMM.

Outra vantagem dos SCHMM relativamente aos DHMM está relacionada com a natural suavidade dos modelos baseados na parametrização funcional do espaço acústico. Pelo contrário, nos DHMM as probabilidades de observação são basicamente determinadas por histogramas, os quais exigem que em alguns sistemas se utilizem técnicas específicas para suavizar estas distribuições discretas e dessa maneira obter modelos acústicos com melhor capacidade de generalização.

Um argumento favorável aos DHMM consiste no facto de dispensarem a avaliação de funções de Gauss, ou outras formas funcionais, tarefa que é responsável por grande parte do esforço de computação [83] nos SCHMM e nos CHMM. No caso dos DHMM, uma parte substancial do *tempo de CPU* verifica-se no processo de quantificação vectorial[1] [93,94], cuja complexidade cresce directamente com a dimensão do *codebook* e do espaço de representação acústica.

Uma perspectiva dos SCHMM que é interessante realçar é a de que eles constituem uma abordagem híbrida relativamente aos CHMM e aos DHMM, por um lado mantendo a parametrização funcional do espaço acústico e, por outro lado, introduzindo alguma partilha de parâmetros através de um *codebook* comum.

Estabelecida a magnitude da tarefa e a base de dados para treino do sistema, os SCHMM podem constituir uma excelente solução de compromisso entre a precisão e a robustez dos modelos. Embora, tal como se referiu, estas duas vertentes não se apresentem como problemas críticos neste sistema, a opção pelos SCHMM foi tomada, em última análise, devido à sua comprovada capacidade de apresentarem um bom desempenho na generalidade dos sistemas de reconhecimento automática da fala com características semelhantes a este e, além disso, oferecerem excelentes perspectivas para trabalho futuro.

2.7 Modelação linguística

2.7.1 Introdução

Tal como foi anteriormente referido (§2.4), a descodificação consiste basicamente num processo de busca, através do espaço das soluções, orientado pela informação contida nos modelos acústico e linguístico.

---

[1] Com base na tecnologia microelectrónica dos circuitos reconfiguráveis do tipo *field programmable gate array* foi uma desenvolvida, em colaboração com um grupo de trabalho da FEUP, uma unidade de processamento dedicada à operação de quantificação vectorial.

O modelo linguístico está neste sistema definido aos níveis sintáctico e lexical (§1.3.4). Assim, através de uma gramática sintáctica estão estabelecidas restrições às sucessões de palavras associadas a esta tarefa. Existe também um léxico que define essas palavras com base num conjunto de unidades elementares da fala estabelecidas ao nível fonético e representadas internamente através dos respectivos modelos acústicos.

Estas duas componentes deste reconhecedor, a gramática e o léxico, são apresentadas na exposição que se segue.

2.7.2 Gramática

Na tarefa de reconhecimento que este sistema executa é utilizada uma gramática *word-pair*, anexa à base de dados *Resource Managment*, que determina os pares de palavras sucessivas possíveis.

Em termos formais, para cada palavra ω pertencente ao vocabulário:

$$\Gamma_\omega = \left\{\omega^1,\ \omega^2,\ \ldots,\omega^{991}\right\} \tag{2.47}$$

a gramática define o conjunto $S_\omega \subset \Gamma_\omega$ das palavras que podem imediatamente seguir-se-lhe. Define também o conjunto $S_I \subset \Gamma_\omega$ das palavras que podem iniciar qualquer frase, assim como o conjunto $S_F \subset \Gamma_\omega$ daquelas que podem terminá-la. Assim, para a frase W_I^K ser sintacticamente correcta é necessário e suficiente que:

$$W_I^K = \left\{\omega_I,\ldots,\omega_i,\omega_{i+1},\ldots,\omega_K\right\},\quad \begin{aligned} &\omega_1 \in S_I,\\ &\omega_{i+1} \in S_{\omega_i},\quad 1 \le i \le K-1\\ &\omega_K \in S_F \end{aligned} \tag{2.48}$$

A dimensão média dos conjuntos S_* é aproximadamente 60, uma estimativa do valor da perplexidade desta tarefa de reconhecimento. Sem a utilização desta gramática, para reconhecer uma palavra o sistema precisaria de extrair do sinal acústico aproximadamente 10 *bits* de informação, enquanto que utilizando-a são apenas necessários em média cerca de 6 *bits*.

Apesar da perplexidade ser por norma um bom indicador da dificuldade da tarefa de reconhecimento, é interessante tentar verificar se relativamente às palavras do vocabulário acusticamente mais confusas, aquelas que mais beneficiariam com as restrições sintácticas, a gramática é realmente eficaz.

Assim, e assumindo também que existe uma correlação significativa entre o número de fonemas de uma palavra e a sua perplexidade, foi feito um pequeno estudo relativamente a esta tarefa em concreto. Na figura 2.16 estão registados para cada uma das 991 palavras do vocabulário, em abcissas, o número de fonemas na respectiva transcrição fonética e, em ordenadas, o número de palavras imediatamente subsequentes permitidas pela gramática. Apesar dos valores em ordenadas se referirem apenas às palavras subsequentes, não explicitando portanto qualquer informação relativamente ao número de possíveis palavras imediatamente anteriores, aceitam-se esses valores como indicadores do grau de restrição sintáctica relativamente a cada palavra em particular.

O aspecto que interessa aqui salientar é o facto de uma grande parte das palavras que admitem directamente muitas sucessoras, nomeadamente as que estão na região abrangida pela curva elíptica, são compostas por um reduzido número de fonemas. Esta situação é , em princípio, bastante desfavorável na medida em que as palavras mais breves serão passíveis de maior confusão acústica pois aderem mais facilmente a uma sucessão de vectores de características incorrecta. Além disso, apresentam localmente uma grande "perplexidade" sintáctica, ou seja, a sua descodificação é pouco condicionada pela gramática.

Por exemplo, a palavra *THE*, constituída por apenas dois fonemas, é a que permite maior número de sucessoras imediatas. É curioso, mas certamente não é surpreendente, que esta palavra é responsável por uma parte muito significativa dos erros de inserção, supressão e substituição de palavras.

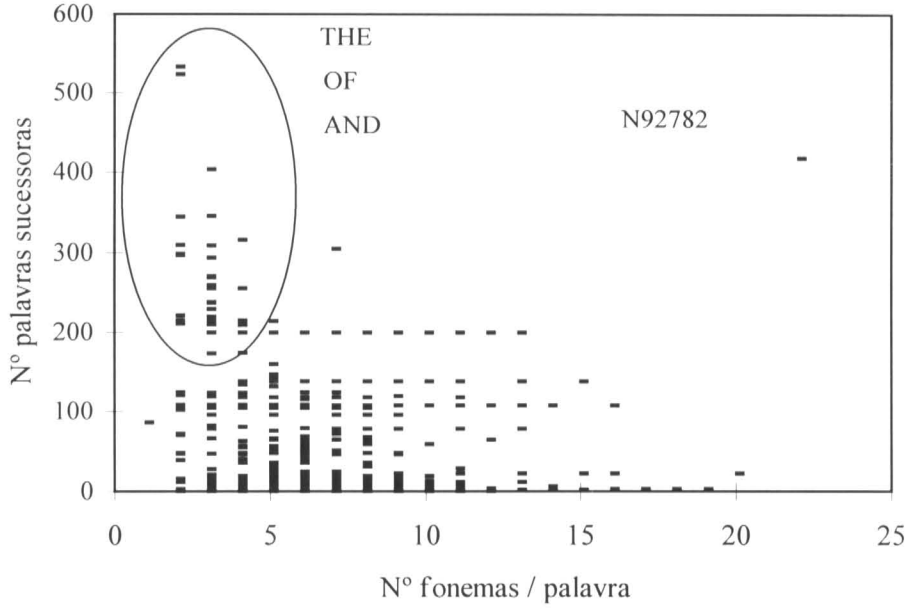

Fig. 2.16 - "Medida fonética" e "perplexidade" das palavras do vocabulário.

Apesar das insuficiências desta gramática, como o exemplo acabado de apontar, não tratando este trabalho do problema da modelação linguística em sistemas de reconhecimento automático da fala foi considerado que ela estabele uma tarefa com o grau de dificuldade adequado aos objectivos estabelecidos.

Por outro lado, por circunstâncias várias relacionadas com a actividade do laboratório onde este sistema foi desenvolvido, especialmente devido à maior facilidade em confrontar directamente os resultados com outros sistemas, revelou-se conveniente usar a gramática *word-pair*.

Refira-se, no entanto, que inicialmente se considerou também a hipótese de utilizar uma gramática estocástica [112] do tipo *bi-gram*[1], também disponível para esta base de dados e determinando a redução da perplexidade da tarefa para um valor próximo de 20.

2.7.3 Léxico

O léxico utilizado neste sistema define cada palavra do vocabulário com base nos 51 símbolos do alfabeto fonético apresentado anteriormente (§2.6.2.2). Para cada palavra encontra-se definida uma única transcrição fonética[2].

$$\omega^i \leftarrow /u_{i1}/\,/u_{i2}/\,\ldots\,/u_{iM_i}/\,, \quad \omega^i \in \Gamma_\omega \tag{2.49}$$

Assim, dados os modelos acústicos correspondentes às unidades fonéticas, este léxico determina um modelo acústico único para cada palavra. É portanto necessário que o léxico estabeleça para cada palavra uma transcrição fonética que permita acomodar as pronunciações mais significativas mas sem tornar os modelos acústicos demasiado imprecisos.

Tal como se referiu anteriormente (§1.3.6.6), alguns sistemas utilizam léxicos mais complexos integrando, de forma explícita ou através de regras fonológicas, múltiplas pronunciações na tentativa de resolver este problema particular de variabilidade do sinal. Neste sistema, essas abordagens foram preteridas por duas razões essenciais: por um lado, a utilização desses léxicos acarreta maior complexidade do sistema, o que não se justificaria face aos objectivos propostos neste trabalho; além disso, a generalidade das técnicas dedicadas a este problema não apresentam de maneira consistente uma melhoria muito significativa dos resultados [115] (§1.3.6.6).

---

[1] No desenvolvimento de um reconhecedor de sequências de dígitos haviam sido utilizadas gramáticas deste tipo.

[2] O léxico é apresentado no Anexo I.

Assim, optou-se por utilizar um léxico disponibilizado pelo laboratório ELIS, que havia sido optimizado nesta aplicação através de um reconhecedor semelhante.

2.8 Descodificação

2.8.1 Introdução

Sendo este reconhecedor baseado numa abordagem estatística integrada, o módulo de descodificação opera sobre uma única estrutura de informação que combina todas as fontes de conhecimento associadas a esta tarefa. Além disso, o reconhecimento de um dado sinal da fala é efectuado através de um algoritmo que pesquisa o espaço de soluções confinado pelas restrições estabelecidas aos níveis acústico e linguístico, determinando a sucessão de palavras que globalmente satisfaz um determinado critério estatístico de decisão [143].

Apesar deste conceito geral, a realização do módulo de descodificação é muito condicionada pelas características da aplicação. Dependendo da estrutura do modelo linguístico e também da complexidade do modelo acústico, diferentes estratégias de representação do conhecimento e de pesquisa da solução são mais ou menos adequadas.

Embora este sistema não se possa considerar, face à tecnologia actual, particularmente exigente, ainda assim o módulo de descodificação merece naturalmente alguma atenção. Na parte inicial desta secção é descrita a arquitectura da estrutura de informação associada aos modelos linguístico e acústico utilizados neste sistema. Em seguida é apresentada a organização do procedimento de pesquisa utilizado.

2.8.2 Estrutura integrada de informação

Neste sistema, a gramática, o léxico e os modelos acústicos encontram-se compilados numa estrutura de informação passível de uma representação gráfica similar à apresentada na figura 1.4 (§1.3.6.2)[1].

Iniciando a análise dessa estrutura ao nível mais elevado, correspondente à sintaxe da tarefa, verifica-se que a sua topologia apresenta 991 nós[2], cada um deles

[1] Frequentemente, estas fontes de conhecimento encontram-se compiladas numa rede finita de estados que ao nível gramatical, contrariamente à formalização aqui utilizada, apresenta uma estrutura de nós interligados por arcos de transição correspondentes à realização das palavras.

[2] Para facilitar a exposição e uma vez que isso não acarreta qualquer ambiguidade danosa utilizar-se-á a mesma notação, ω, para estes nós e para as palavras.

associado a uma das palavras do vocabulário. Existe ainda um nó inicial, ω^I, e um nó final, ω^F.

Todos estes nós estão interligados de acordo com a gramática estabelecida para esta tarefa (§2.7.2), ou seja, estão definidos os nós que podem suceder a ω^I, ou a qualquer outro nó ω^i, ou anteceder ω^F, através dos conjuntos, respectivamente, S_I, S_ω e S_F. A cada uma dessas conexões está associada uma probabilidade cujo valor é o inverso do número de nós candidatos a sucessores do nó de origem. Em alternativa poder-se-ia considerar um valor único, por exemplo unitário, para todas as transições entre palavras. Nesse caso, previsivelmente as palavras com mais sucessoras seriam menos penalizadas. Refira-se, no entanto, que o desempenho do reconhecedor é relativamente insensível ao valor destas probabilidades, tal como foi já referido (§2.6.3.3) e nesta situação comprovado experimentalmente. Além disso, a determinação precisa desses valores não se justifica dada até a prática corrente, aqui também seguida, de optimizar o desempenho do sistema ajustando de maneira empírica um parâmetro comum a todos os nós e que penaliza o seu abandono.

Considere-se então um caminho genérico, C_1^K, obrigatoriamente começando em ω^I, passando por K nós e terminando necessariamente em ω^F. É importante salientar que um qualquer dos nós pode ser visitado mais do que uma vez, embora não de maneira consecutiva pois esta gramática não o permite. Assim, seja:

$$C_1^K = \left\{\omega^I, \omega_1, \omega_2, \ldots, \omega_K, \omega^F\right\} \tag{2.50}$$

com as transições entre os nós afectadas pelas probabilidades:

$$\begin{aligned}
&\omega^I \xrightarrow[P=\frac{1}{\#S_I}]{} \omega_1, && \omega_1 \in S_I \\
&\omega_i \xrightarrow[P=\frac{1}{\#S_{\omega_i}}]{} \omega_{i+1}, && \omega_{i+1} \in S_{\omega_i}, \quad 1 \le i \le K-1 \\
&\omega_K \xrightarrow[P=\frac{1}{\#S_{\omega_K}}]{} \omega^F, && \omega_K \in S_F
\end{aligned} \tag{2.51}$$

Associada a C_1^K está a informação linguística que interessa extrair e que consiste na frase W_I^K. Então, está assim definida a probabilidade *a priori* de ocorrência da sucessão de palavras $W_I^K = \left\{\omega_1, \omega_2, \ldots, \omega_K\right\}$:

$$P\left(W_1^K\right) = \frac{1}{\#S_I}\prod_{i=1}^{K}\frac{1}{\#S_{\omega_i}} \tag{2.52}$$

Expandindo os nós inicial e final, ω^I e ω^F, verifica-se que cada um deles é constituído por um modelo escondido de Markov cuja função é modelar os períodos de silêncio, ou quase silêncio, que ocorrem nos extremos das frases. A sua topologia consiste num único estado, tal como representado na figura 2.15, e os seus parâmetros resultam do procedimento de treino (§2.9). No estado correspondente ao nó inicial está definida uma probabilidade de permanência $a_{I,I}$, tendo sido considerado o valor unitário relativamente ao nó final. Em ambos os nós é utilizada a mesma mistura gausseana, considerando portanto que nos extremos da frase o sinal acústico apresenta idêntica distribuição estatística.

Qualquer um dos restantes nós encerra o modelo acústico da respectiva palavra do vocabulário. Como se pode verificar através da figura 2.17, referente a uma palavra composta por M_i unidades fonéticas, esse modelo escondido de Markov resulta basicamente da composição dos modelos acústicos elementares de acordo com a informação lexical (equação 2.49). Verifica-se ainda que existe um estado opcional, representado pelo símbolo /-/, à saída do HMM. A sua função é modelar o eventual período de silêncio entre as palavras sucessivas.

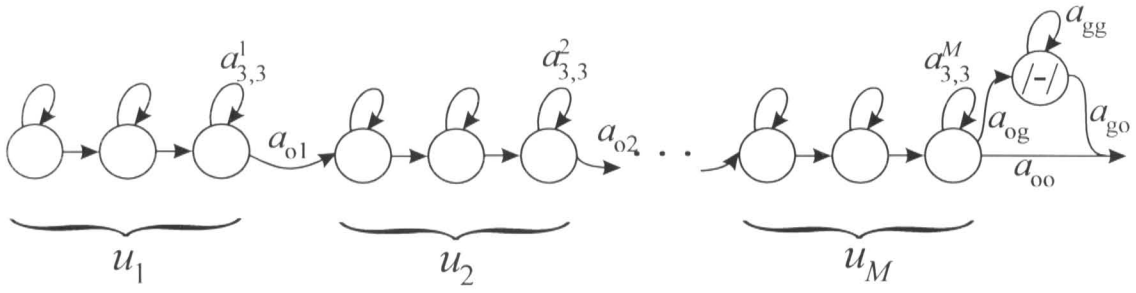

Fig. 2.17 - Modelo acústico da palavra ω^i.

Os parâmetros que caracterizam os HMM correspondentes às unidades fonéticas são adaptados durante o procedimento de treino. As probabilidades de transição entre esses HMM, no interior de cada palavra, são calculadas de maneira a verificar a condição imposta pela equação 2.27, ou seja, $a_{0i} = 1.0 - a^i_{3,3}$, $i = 1,\ldots M-1$. Em princípio, o modelo acústico seria mais rigoroso se essas probabilidades fossem treinadas dependendo de cada palavra, no entanto, tal como se justificou anteriormente, o desempenho do sistema é relativamente insensível ao seu valor. As probabilidades de transição a_{og} e a_{oo} são calculadas a partir da probabilidade de permanência no último

estado do último fonema, $a_{3,3}^{M}$, e da probabilidade, também treinada, de transitar de uma qualquer palavra para o modelo do silêncio, $P(g|\omega)$. Assim, $a_{og} = \left(1.0 - a_{3,3}^{M}\right) P(g|\omega)$ e $a_{oo} = \left(1.0 - a_{3,3}^{M}\right)\left(1.0 - P(g|\omega)\right)$.

Os parâmetros do estado que modela o período de silêncio entre as palavras, nomeadamente a_{gg}, a_{go} e os coeficientes da mistura gausseana, são comuns a todas as palavras. Além disso, a mistura gausseana é também idêntica à utilizada nos modelos de silêncio nas extremidades da frase. Refira-se que não se recorreu a um modelo mais complexo na junção de duas palavras porque previsivelmente não contribuiria para melhorar substancialmente o desempenho do sistema e os resultados obtidos com esta versão simples são plenamente satisfatórios.

Em todo o caso, está definido o modelo acústico para todas as sucessões de palavras permitidas por esta tarefa, isto é, uma estrutura de informação que permite determinar a verosimilhança, $p\left(\boldsymbol{X}_1^T \middle| W_1^K\right)$, de observar a sucessão de vectores de características, $\boldsymbol{X}_1^T$, condicionada ao conhecimento de W_1^K. No cálculo desta função estão incluídos os modelos acústicos dos nós inicial e final, ou seja, o resultado $p\left(\boldsymbol{X}_1^T \middle| W_1^K\right)$ é efectivamente obtido calculando $p\left(\boldsymbol{X}_1^T \middle| C_1^K\right)$. Note-se também que os nós que integram esta sucessão estão ligados por arcos de transição que tomam o valor da probabilidade de abandonar o último estado do respectivo modelo acústico, não considerando portanto as probabilidades associadas ao modelo linguístico.

Neste sistema, este HMM global tem aproximadamente $2\text{x}10^4$ estados e 10^5 arcos de transição. Existe uma grande partilha de parâmetros, por exemplo todos os estados são modelados utilizando apenas 154 misturas diferentes, as quais se baseiam num *codebook* com 4096 funções de Gauss. E relativamente aos arcos de transição, são apenas utilizados aproximadamente 260 valores diferentes de probabilidade de transição.

Este modelo acústico é combinado com o modelo linguístico, sintetizado na equação 2.52, formando então um modelo global que permite determinar a probabilidade conjunta de W_1^K ocorrer e do respectivo modelo acústico, considerando a extensão dos modelos de silêncio inicial e final, observar $\boldsymbol{X}_1^T$:

$$p\left(W_1^K, \boldsymbol{X}_1^T\right) = P\left(W_1^K\right) p\left(\boldsymbol{X}_1^T \middle| W_1^K\right) \tag{2.53}$$

Os modelos acústico e linguístico estão efectivamente fundidos num HMM de 1ª ordem, dado que esta gramática apresenta também dependências passíveis de representação através de uma cadeia de Markov de 1ª ordem. É apenas necessário

afectar as probabilidades do modelo linguístico aos arcos de transição entre as palavras, e entre estas e os modelos do *silêncio* inicial e final. É claro que deixa de se verificar a condição imposta pela equação 2.27, no entanto esse facto não acarreta dificuldades práticas.

2.8.3 Algoritmo de descodificação

Agora que a informação linguística e acústica associada a este sistema se encontra embutida numa estrutura integrada, dada uma sucessão de vectores de características $\boldsymbol{X}_1^T$ é necessário seleccionar a hipótese de reconhecimento correcta W_1^K segundo um determinado critério de descodificação [140].

Em concordância com os resultados discutidos anteriormente relativos ao classificador Bayesiano (§1.3.6.3), um descodificador óptimo[1] baseia-se no critério de selecção:

$$W_1^K = arg\max_W \left\{ P(W) p\left(\boldsymbol{X}_1^T | W \right) \right\} \tag{2.54}$$

A aplicação deste critério de maneira rigorosa e exaustiva, isto é, o cálculo do produto da probabilidade *a priori* pela verosimilhança para todas as sucessões de palavras possíveis, seguida da identificação do resultado máximo e a consequente selecção da respectiva frase, é apenas praticável em problemas de reduzida dimensão, pois o espaço de pesquisa cresce exponencialmente com K [85]. Contudo, existem algoritmos de descodificação baseados neste critério óptimo[2] e ainda assim viáveis em aplicações de grande dimensão, em virtude da combinação de técnicas que optimizam a geração de hipóteses e reduzem, por aproximação, o espaço de busca.

Neste sistema é utilizado o chamado descodificador Viterbi, o qual aproxima o critério (2.54) substituindo o cálculo exacto da verosimilhança pela função $p_q(\ )$:

$$W_1^K = arg\max_W \left\{ P(W) p_q\left(\boldsymbol{X}_1^T | W \right) \right\} \tag{2.55}$$

Seja Q_W^T o conjunto de todos os caminhos de comprimento T atravessando o HMM global, com início e final em ω^I e ω^F, respectivamente, e atravessando pela ordem indicada os modelos acústicos das palavras definidas na sucessão W. A função

[1] Frequentemente designado, em Inglês, por *Maximum a posteriori* (MAP) *decoder* ou *Forward decoder*.

[2] É o caso do algoritmo conhecido, em Inglês, por *stack decoder*.

$p_q(\ )$ apresenta um resultado aproximado da verosimilhança total (§2.6.3.2), considerando apenas a sucessão de estados, q pertencente a Q_W^T, que apresenta a maior verosimilhança:

$$p_q\left(\boldsymbol{X}_1^T|W\right) = \max_{q \in Q_W^T} p\left(\boldsymbol{X}_1^T, q|W\right) \tag{2.56}$$

Fica assim claro que o descodificador Viterbi apenas poderia respeitar o critério óptimo se o objectivo fosse determinar a sucessão de estados, e não a sucessão de palavras. No entanto, tem-se verificado experimentalmente que na generalidade dos sistemas a verosimilhança total é efectivamente dominada, em maior ou menor grau, por uma única sucessão de estados. Existem resultados publicados confrontando directamente os descodificadores óptimo e Viterbi, concluindo que a aproximação associada a este último acarreta uma degradação dos resultados que se pode considerar insignificante, em alguns sistemas [135], ou, noutros reconhecedores [136], não desprezável.

É oportuno fazer aqui uma breve referência a uma, de entre as diversas variantes do descodificador Viterbi que têm sido introduzidas, neste caso apresentando uma solução híbrida relativamente ao descodificador óptimo, adicionando as probabilidades nos caminhos que se fundem no interior das palavras e apenas considerando o máximo nos caminhos que se fundem na fronteira entre palavras [137].

Um argumento referido por diversas vezes [2,85] aponta a maior consistência, pelo menos no plano teórico, que o descodificador óptimo apresenta quando o sistema é treinado, com base no critério de máxima verosimilhaça, através de um algoritmo[1] que soma, em cada estado, as verosimilhanças associadas a todos os caminhos possíveis contendo esse estado (§2.9). Curiosamente, o descodificador Viterbi apresenta uma relação análoga com o algoritmo de treino utilizado no desenvolvimento deste sistema, ou seja, de maneira simplificada pode afirmar-se que ambos substituem a soma das verosimilhanças pela máxima verosimilhança.

Estreitamente relacionado com este aspecto está a capacidade do descodificador Viterbi tornar visível o processo estocástico associado à cadeia de Markov, fornecendo um alinhando entre a sucessão de vectores de características e os estados do modelo. Esta aptidão de segmentar o sinal ao nível dos estados e, consequentemente, das unidades fonéticas, ou das palavras, é frequentemente muito importante, tendo-o também sido ao longo deste trabalho.

Outra vantagem muito importante do descodificador Viterbi reside no facto de permitir uma acentuada redução da complexidade computacional relativamente ao descodificador óptimo. Além disso, este descodificador é particularmente adequado para aplicações em tempo real, pois quando inicia o processamento do vector de

[1] O algoritmo conhecido por *forward-backward* ou *Baum-Welsh*.

características $\boldsymbol{x}_t$, $1 \le t \le T$, já estão completamente definidas as hipóteses baseadas na mesma sucessão parcial de vectores de características $\boldsymbol{X}_1^{t-1}$ [1].

Em síntese pode dizer-se que foram as qualidades do descodificador Viterbi referidas, as quais têm contribuído para o tornar um *standard* nos sistemas de reconhecimento de fala contínua [2], aliadas a um desempenho plenamente satisfatório no âmbito deste trabalho que justificaram a opção pela sua utilização.

Existem diferentes algoritmos, baseados em técnicas de Programação Dinâmica [141] aplicadas à medida que a sucessão de vectores de características vai sendo disponibilizada, que realizam o descodificador Viterbi. Neste sistema utiliza-se uma versão baseada no conceito de "passagem de testemunhos"[2].

A informação representada em cada *testemunho* é a identificação do estado do HMM global em que se encontra, a hipótese de reconhecimento parcial que veicula[3] e a respectiva probabilidade conjunta. O procedimento é iniciado com um *testemunho* no nó inicial. A cada novo vector de características, todos os *testemunhos* existentes são propagados aos respectivos estados sucessores e é actualizado o seu conteúdo de acordo com a informação codificada no HMM global e as características acústicas da observação. Então, relativamente a cada estado sobrevive apenas o *testemunho* que apresenta a hipótese de reconhecimento parcial com maior valor de probabilidade, os restantes, no caso de existirem outros associados ao mesmo estado, são simplesmente eliminados antes de iniciar nova transmissão. No final do processo de descodificação a informação sobre a frase descodificada está contida no *testemunho* que se encontra no nó final.

A execução extensiva deste algoritmo em tempo real não é viável para aplicações de grande dimensão e constitui um problema ainda difícil mesmo em sistemas de média dimensão. Note-se que neste sistema existiriam, passadas apenas algumas poucas décimas de segundos após o início da descodificação, aproximadamente 2×10^4 *testemunhos* que a cada novo vector de características deveriam ser propagados. A maior parte deles seriam transmitidos a apenas dois estados, mas cerca de 2×10^3, os que se encontrassem no último estado de cada palavra ou no respectivo modelo de silêncio, apresentariam em média 60 estados sucessores. Logo, a cada 10 ms seria necessário realizar aproximadamente 1.6×10^5 passagens de *testemunho* e seleccionar então o melhor em cada um dos 2×10^4 estados, eliminando os restantes.

Neste sistema é utilizada uma técnica heurística que reduz significativamente o espaço de pesquisa sem penalizar o desempenho do sistema [142]. Consiste

[1] Diz-se então que o algoritmo Viterbi é *time-synchronous*, tal como, por exemplo, o algoritmo *one-stage*, utilizado num sistema anteriormente desenvolvido para o reconhecimento de dígitos ligados.
[2] Do Inglês, *token-based Viterbi decoder*.
[3] Decidiu-se utilizar esta informação, em vez do sistema de apontadores que permite recuperar a hipótese de reconhecimento através de um mecanismo de *backtracking*.

basicamente em preservar, em cada ciclo, apenas os *testemunhos* cuja probabilidade conjunta da hipótese parcial pertence a uma determinada vizinhança do valor máximo dessa probabilidade[1]. O facto do algoritmo Viterbi ser *time-synchronous* torna relativamente simples a aplicação desta técnica, pois as diversas hipóteses parciais já se encontram naturalmente normalizadas. A amplitude dessa vizinhança é determinada de maneira empírica, sendo suficientemente grande para não aumentar em demasia o risco de eliminar a solução correcta devido à utilização desta técnica mas, por outro lado, sendo bastante reduzida para acelerar efectivamente o processo de descodificação evitando o desperdício de recursos de computação em hipóteses pouco viáveis. Apesar da simplicidade deste método, baseado numa vizinhança constante ao longo do processo de descodificação, ele permite também uma adaptação bastante eficaz do número de *testemunhos* activos dependendo do grau de aderência do sinal da fala aos modelos linguístico e acústico.

Em seguida é apresentada, no essencial, a versão do algoritmo Viterbi utilizado neste sistema.

Algoritmo Viterbi

(Notação)

A(I,J): probabilidade de transição do estado I para o J, ambos pertencentes ao HMM global (nas transições entre as palavras este valor inclui as probabilidades do modelo linguístico);

B(J,t): verosimilhança do estado J observar o vector de características $\boldsymbol{x}_t \in \boldsymbol{X}_1^T$;

L(s): lista de estados activos; no início de cada iteração s=INICIO; no final de cada iteração s=FINAL;

V(I,s): probabilidade conjunta (verosimilhança acústica + probabilidade linguística) acumulada pelo testemunho no estado I;

H(I,s): lista de palavras visitadas pelo testemunho no estado I, no início da iteração;

(Iniciação da descodificação)

INICIO←1;
FINAL←2;
L(INICIO) ← estado_inicial;
V(estado_inicial, INICIO) ← B(estado_inicial,1);
H(estado_inicial, INICIO) ← {};

[1] Esta técnica de *prunning* é conhecida por *beam searching*.

```
(Iteração Viterbi para todos os termos da sucessão de vectores de características)
Para t=2 até T

    (Iniciação do t^ésimo ciclo)
    Esvazia L(FINAL);
    V(todos_estados, FINAL) ← 0.0;
    Max ← 0.0;

    (Transmissão de testemunhos)
    Para todos os estados, I, em L(INICIO)
        Para todos os estados, J, sucessores de I
            Se  V(I, INICIO) A(I,J) B(J,t) > Max - Largura_feixe
              e V(J, FINAL) < V(I, INICIO) A(I,J) B(J,t)
                Se V(J, FINAL) = 0.0
                    Acrescenta J a L(FINAL);
                V(J, FINAL) ← V(I, INICIO) A(I,J) B(J,t);
                H(J, FINAL) ← H(J, INICIO);
                Se I e J pertencem a palavras diferentes
                    Acrescenta palavra que contém J a H(J, FINAL);
                Se  V(J, FINAL) > Max
                    Max ← V(J, FINAL);

    (Elimina testemunhos pouco verosímeis)
    Para todos os estados, I, em L(FINAL)
        Se V(I, FINAL) < Max - Largura_feixe
            Apaga I de L(FINAL);

    (Troca conteúdo de variáveis)
    INICIO ↔ FINAL;

(Resultado)
Frase_descodificada ← H(estado_final,INICIO);
```

Foram introduzidas algumas alterações a este algoritmo para resolver os problemas numéricos decorrentes dos valores das probabilidades envolvidas serem pequenos. Mesmo utilizando a máxima precisão numérica possível, a probabilidade conjunta acumulada num qualquer *testemunho* rapidamente ultrapassa o limite inferior de representação[1] da máquina.

Quando se utiliza um algoritmo baseado no descodificador óptimo, este problema é frequentemente resolvido através da normalização das probabilidades conjuntas a cada novo vector de características, permitindo assim que esses valores se mantenham

[1] Passados apenas algumas dezenas de sectores ocorre o *floating-point underflow*.

na gama de representação numérica do computador[1]. Uma outra possibilidade é utilizar a transformada logarítmica, no entanto esta técnica é pouco prática nesses descodificadores pois exige procedimentos mais elaborados para o cálculo dos logaritmos das somas.

No caso do descodificador Viterbi, a representação das verosimilhanças e probabilidades conjuntas no domínio logarítmico é muito simples, tendo sido essa a solução seguida neste sistema. Assim, a equação recursiva básica (2.39) do algoritmo *"forward"* é substituída por:

$$\alpha_{j,t} = \max_i \left\{ \alpha_{i,t-1} + \log\left(a_{i,j}\right) \right\} + \log\left(b_j\left(\boldsymbol{x}_t\right)\right) \quad \text{para} \quad t=2,...,T \quad \text{e} \quad j=1,...,S \tag{2.57}$$

Também ao nível da eficiência computacional do processo de descodificação foram introduzidas diversas alterações, sendo de destacar as utilizadas no sentido de evitar a repetição do cálculo de gausseanas, dada a sua partilha intensiva e os custos elevados que acarretam.

Ainda em relação às funções de Gauss, atendendo à dimensão relativamente elevada do espaço de representação acústica foram efectuadas algumas experiências com o objectivo de ensaiar um estudo sobre a viabilidade de cancelar o seu cálculo se as somas parciais dos termos da exponencial atingissem um determinado limiar mínimo. Tendo presente a equação genérica das gausseanas utilizadas neste sistema:

$$N_m\left(\boldsymbol{x}\right) = k_m exp\left\{ S_{39}\left(\boldsymbol{x}\right) \right\}, \quad com \quad S_k\left(\boldsymbol{x}\right) = \sum_{i=1}^{k} - \frac{\left(x^i - \mu_m^i\right)^2}{2\left(\sigma_m^i\right)^2} \tag{2.58}$$

experimentou-se calcular as somas parciais S_{13}, correspondente aos termos estáticos, ou S_{26}, considerando também os termos correspondentes às primeiras derivadas, e dispensar o cálculo dos termos restantes no caso dessas somas parciais apresentarem um valor tão pequeno que fosse possível considerar imediatamente $N_m\left(\boldsymbol{x}\right) \cong 0.0$. Deve ser referido que não se obtiveram resultados conclusivos sobre a eficácia deste procedimento, nomeadamente quanto ao objectivo de identificar as somas parciais mais indicadas para avaliação prévia[2]. Ainda assim, neste estudo preliminar verificaram-se resultados interessantes submetendo à primeiro avaliação uma soma parcial com o termo estático da energia e apenas mais dois outros termos, os MFCC de menor ordem. Em grande medida esse resultado decorre do facto da energia apresentar

[1] Este foi o método utilizado em sistemas desenvolvidos anteriormente para reconhecimento de palavras isoladas.

[2] Este estudo não foi prosseguido sobretudo porque exigiria um esforço não justificável face aos objectivos prioritários do trabalho.

uma distribuição estatística nitidamente bi-modal[1], ao contrário do que sucede com a generalidade dos MFCC e respectivas derivadas, os quais apresentam uma distribuição aproximadamente Normal, com a clara excepção do primeiro coeficiente. Por exemplo, considerando apenas esses três termos, a distância entre o vector de características associado aos sectores que ocorrem nos períodos de silêncio[2] e a grande parte das gausseanas é já muito grande, podendo num número significatico de casos ser dispensável o cálculo dos restantes 36 termos da exponencial.

2.9 Treino do sistema

2.9.1 Introdução

Estando definida a estrutura paramétrica do sistema, impõe-se agora a determinação dos parâmetros livres.

Note-se que neste sistema apenas os parâmetros acústicos estão ainda por fixar, não sendo portanto necessário treinar o modelo linguístico. Esta é, desde logo, uma circunstância muito favorável, basta ter presente a frequente dificuldade em dispor de uma base de dados acústico-linguística com dimensão suficiente para treinar conjuntamente os modelos linguístico e acústico dos reconhecedores de fala contínua. Então, geralmente recorre-se ao seu treino de maneira independente (§1.3.6.8), o que é reconhecido [19] conduzir a uma solução sub-óptima.

Assim, nesta secção são tratados diferentes aspectos relacionados com o treino do modelo acústico deste sistema. Inicialmente é apresentado o critério da máxima verosimilhança, subjacente ao processo de treino seguido, e são salientados alguns tópicos sobre critérios alternativos. É então feita a apresentação da estratégia de treino seguida e são descritos os respectivos procedimentos, nomeadamente: a iniciação dos parâmetros; a sua adaptação através de um procedimento baseado no algoritmo Viterbi; a utilização de técnicas para expandir ou reduzir o conjunto dos parâmetros livres. Finalmente são apresentados os resultados obtidos durante o processo de treino.

2.9.2 Critério de treino

Os parâmetros livres do modelo acústico deste sistema são adaptados através de um procedimento de treino definido com base numa aproximação ao critério da máxima verosimilhança (§1.3.6.8).

---

[1] Tal como no histograma apresentado na figura 2.8, onde imediatamente se identificam os períodos de quase-silêncio.

[2] São em grande número nas frases utilizadas no treino e no teste, pois não se utilizou qualquer procedimento de *end-point detection*.

Seja T o conjunto de treino, constituído por N_{T} exemplos $\left(\boldsymbol{X}^n, W_c^n\right)$, tal como definido na equação 1.7. Então, o objectivo do treino é determinar o conjunto dos parâmetros livres θ que maximiza a verosimilhança do modelo acústico correspondente a W_c^n observar $\boldsymbol{X}^n$, estendida a todos os exemplos de T :

$$\theta = \arg\max_{\theta} \prod_{n=1}^{N_{\mathsf{T}}} p_\theta\left(\boldsymbol{X}^n | W_c^n\right) \tag{2.59}$$

Note-se que relativamente a este sistema esta equação é equivalente à equação 1.9, uma vez que os modelos acústico e linguístico são considerados independentes e este último já se encontra fixo.

Uma propriedade do critério da máxima verosimilhança é a sua eficiência assimptótica na estimação dos parâmetros [147,148], isto é, se a forma funcional do modelo paramétrico $p_\theta(\boldsymbol{X}|W)$ é correcta e se N_{T} é suficientemente grande, então essa estimativa de θ conduz a um modelo paramétrico que se aproxima da distribuição real $p(\boldsymbol{X}|W)$. No entanto, na prática não é possível satisfazer a chamada assunção da correcção do modelo dada a evidente complexidade das dependências estatísticas que se verificam no sinal da fala e, por outro lado, a magnitude das restrições impostas ao modelo acústico, tornando-o exequível mas limitando a sua capacidade de modelação.

Uma característica do critério da máxima verosimilhança que pode potenciar os efeitos da incorrecção da forma funcional do modelo é o facto da estimação dos parâmetros ser globalmente dominada pelos exemplos de treino que são típicos em cada classe. Admitindo que se verificava a correcção do modelo, então a contribuição dos exemplos típicos conduziria às fronteiras de decisão adequadas na perspectiva dum classificador Bayesiano [2]. No entanto, a "inevitável" incorrecção do modelo paramétrico e a falta de influência dos exemplos mais próximos das fronteiras na adaptação dos parâmetros, conduz à definição sub-óptima dessas fronteiras. Embora incorrendo em alguma falta de rigor formal, poder-se-á dizer que de acordo com este critério as fronteiras são definidas de maneira a representar bem os exemplos típicos de cada classe mas não a representar mal os exemplos das outras classes [2].

Os aspectos acabados de referir estão assim estreitamente relacionados com a falta de capacidade discriminativa do critério da máxima verosimilhança. Esta deficiência pode ser formalmente apresentada nos termos que se seguem [7].

De acordo com o critério da máxima probabilidade *a posteriori* [148,150] (equação 1.8):

$$\theta = \arg\max_{\theta} \prod_{n=1}^{N_{\mathsf{T}}} P_\theta\left(W_c^n | \boldsymbol{X}^n\right) \tag{2.60}$$

Recorrendo à regra de Bayes obtém-se a equação equivalente:

$$\theta = \arg\max_{\theta} \prod_{n=1}^{N_T} \frac{p_\theta\left(\boldsymbol{X}^n | W_c^n\right) P\left(W_c^n\right)}{p_\theta\left(\boldsymbol{X}^n\right)} \tag{2.61}$$

Repare-se que a notação utilizada considera já a situação particular deste sistema, em que a probabilidade *a priori* é independente do conjunto dos parâmetros a treinar. Quanto à densidade probabilística do modelo global[1] observar $\boldsymbol{X}^n$, $p_\theta\left(\boldsymbol{X}^n\right)$, depende do conjunto dos parâmetros acústicos, sendo o seu valor o resultado acumulado das probabilidades conjuntas correspondentes a todas as frases permitidas pelo modelo linguístico. Logo, a equação 2.61 pode escrever-se:

$$\theta = \arg\max_{\theta} \prod_{n=1}^{N_T} \frac{p_\theta\left(\boldsymbol{X}^n | W_c^n\right) P\left(W_c^n\right)}{\sum_k p_\theta\left(\boldsymbol{X}^n | W_k^n\right) P\left(W_k^n\right)} \tag{2.62}$$

ou, através da equação equivalente:

$$\theta = \arg\max_{\theta} \prod_{n=1}^{N_T} \frac{p_\theta\left(\boldsymbol{X}^n | W_c^n\right) P\left(W_c^n\right)}{\sum_{k \neq c} p_\theta\left(\boldsymbol{X}^n | W_k^n\right) P\left(W_k^n\right)} \tag{2.63}$$

Este último resultado evidencia o facto do treino segundo o critério da máxima probabilidade *a posteriori* produzir modelos discriminativos, isto é, os parâmetros são adaptados de maneira a aumentar a probabilidade conjunta do modelo correcto mas também de modo a reduzir a probabilidade conjunta acumulada a todos os modelos rivais. Naturalmente, este resultado é concordante com o facto do treino baseado neste critério atribuir um maior peso aos exemplos que se encontram próximos das fronteiras de decisão.

Comparando as equações 2.63 e 2.59 imediatamente se conclui que ao contrário do critério da máxima probabilidade *a posteriori*, o critério da máxima verosimilhança não tem capacidade de estimar modelos discriminativos, pois para cada exemplo de treino não considera a contribuição "negativa" dos modelos rivais. Quanto ao procedimento de treino em concreto, isto significa que cada modelo é treinado independentemente, apenas com os exemplos respeitantes à sua classe. Deve ser no entanto notado que devido à partilha de parâmetros, que é bastante acentuada neste sistema (§2.8.2), os parâmetros de cada modelo acústico não são efectivamente treinados de maneira completamente independente. Contudo, isto não invalida as considerações feitas sobre a falta de capacidade discriminativa dos modelos determinados pelo critério da máxima verosimilhança.

[1] Ou seja, o HMM global definido em §2.8.2.

Atendendo ao que se acaba de expor, e tendo presente que se optou neste sistema por utilizar modelos acústicos pouco precisos, é previsível que o treino desses modelos com base num critério discriminativo conduziria a melhores taxas de reconhecimento. É oportuno fazer aqui uma brevíssima referência a outros critérios de treino que apresentam também propriedades discriminativas [154], desde os que se baseiam em medidas de Informação [151] - como é o caso do critério da *máxima informação mútua* [149,157] -, até critérios assentes na minimização de uma *função risco* definida com base nos parâmetros acústicos do modelo e de alguma maneira relacionada com a expectativa do erro de reconhecimento - como por exemplo no critério conhecido por *mínimo erro de classificação* [158] -.

Foram essencialmente duas as razões que determinaram a utilização do critério da máxima verosimilhança. Por um lado, tal como justificado anteriormente, ao desenvolver este sistema pretendeu-se sobretudo obter uma referência satisfatória para o trabalho posterior, não constituindo o seu nível de desempenho um objectivo especialmente crítico. Por outro lado, e este foi um argumento de enorme peso, o treino discriminativo é muito mais complexo que o treino com base no critério da máxima verosimilhança. É claro que essa dificuldade se acentua à medida que a dimensão do sistema cresce [152,155], assim como o número de classes rivais. Mas mesmo em aplicações simples este aumento da complexidade é já um aspecto a ter em consideração, tal como se havia constatado durante o desenvolvimento de um reconhecedor de dígitos isolados. O treino discriminativo desses modelos acústicos[1], com base no critério da máxima probabilidade *a posteriori*, acarretou custos computacionais substancialmente maiores que os registados quando os modelos haviam sido treinados com base no critério da máxima verosimilhança.

2.9.3 Processo de treino

2.9.3.1 Estratégia

Não existindo um método analítico para estimar o conjunto dos parâmetros livres, o treino do sistema apresenta-se como um problema de optimização definido de acordo com o critério seleccionado. Apesar da existência de algoritmos relativamente eficientes[2] para o treino com base no critério da máxima verosimilhança, este continua a ser um problema difícil dada a elevada dimensão do espaço de optimização e também o grande número de exemplos de treino associados à função objectivo. Neste sistema é necessário treinar aproximadamente 4×10^5 parâmetros escalares[3] e a base de dados utilizada no treino contém aproximadamente 10^6 vectores de características.

[1] Apenas um DHMM com 5 estados para modelar cada um dos 10 dígitos.
[2] Sobretudo se comparados com os algoritmos disponíveis para o treino discriminativo.
[3] Ainda assim um número bastante inferior ao que se verifica nos reconhecedores de fala contínua de muito grande dimensão.

É assim importante utilizar uma estratégia de treino que permita um compromisso satisfatório entre a precisão dos modelos, o número de parâmetros livres e o tempo de treino. O treino deste sistema foi efectuado de acordo com a estratégia que se apresenta em seguida, a qual, após a fase de iniciação, combina a adaptação dos parâmetros com um aumento gradual do seu número. Esse aumento decorre da utilização de procedimentos simples e eficazes que, por um lado, reduzem as misturas gausseanas ao mínimo sem deteriorar o desempenho e, por outro lado, expandem essas mesmas misturas através da cisão das funções de Gauss definidas no *codebook*, permitindo uma maior precisão.

(Iniciação)

Gera um *codebook* com 512 gausseanas;
Inicia os modelos acústicos, utilizando misturas com 512 gausseanas;

(Treino)

Enquanto dimensão_do_*codebook* ≤ 4K gausseanas

1. Enquanto a taxa do erro diminui, adapta o actual conjunto de parâmetros;
2. Para cada estado elimina os coeficientes da mistura com o valor inferior a determinada fracção, ε, do maior coeficiente da mistura;
3. Enquanto a taxa do erro diminui, adapta o actual conjunto de parâmetros (correspondente aos modelos reduzidos);
4. Cinde gausseanas, duplicando a dimensão do *codebook*;

Nas próximas secções serão discutidos, com algum detalhe, os procedimentos de iniciação do conjunto dos parâmetros acústicos e também a iteração nuclear de adaptação desses parâmetros (etapas 1 e 3 do ciclo de treino). Agora, apenas é oportuno salientar alguns outros aspectos da estratégia de treino apresentada.

Assim, em relação à etapa 2 do ciclo de treino, o objectivo é reduzir a dimensão das misturas gausseanas uma vez que se verifica a expectativa de que a maior parte dos estados utiliza efectivamente só uma pequena fracção do *codebook* (§2.6.3.3 e §2.9.3.4). Em cada estado, cuja mistura depende do conjunto de coeficientes $\{c_{j,m},\ m=1,\ldots,M_j\}$, tal como mostra a equação 2.46, são eliminados os coeficientes $c_{j,m} < \varepsilon\, max\{c_{j,m}\ \ m=1,\ldots,M_j\}$. Para valores de ε superiores a aproximadamente 10^{-3} verificou-se uma significativa degradação dos resultados, tendo sido empiricamente ajustados valores da ordem de grandeza 10^{-4} ou 10^{-5}. Terminado este procedimento, os coeficientes não eliminados foram novamente normalizados. É interessante verificar que terminada esta etapa do ciclo de treino, em nenhuma iteração foi possível

dispensar qualquer das funções de Gauss definidas no *codebook*, isto é, todas elas eram utilizadas em pelo menos uma das misturas[1].

Na etapa 4 do ciclo de treino cada uma das funções de Gauss, caracterizada pelo par (μ,σ) tal como apresentado em 2.42, dá lugar a duas novas gausseanas com igual peso na mistura. É mantido o vector das variâncias de cada uma delas e o vector médio é alterado de maneira a deslocá-las em sentidos opostos e afastando-as pela distância $2k\sigma$:

$$(\mu,\ \sigma) \rightarrow \begin{cases} (\mu - k\sigma,\ \sigma) \\ (\mu + k\sigma,\ \sigma) \end{cases} \tag{2.64}$$

Nos sucessivos ciclos de treino a dimensão do *codebook* tomou os valores 512, 1024, 2048 e finalmente 4096. Foram também efectuadas algumas iterações com a dimensão do *codebook* igual a 8K, tendo o erro evoluído de forma muito irregular e sem diminuir de maneira significativa e consistente. A partir de determinado "ponto" era mesmo perceptível uma tendência clara de aumento do erro, indiciando a ocorrência de sobre-treinamento.

Assim, com esta estratégia baseada no aumento gradual e localmente selectivo da dimensão do sistema, através das etapas 2 e 4, é possível uma progressiva melhoria da precisão dos modelos acústicos mas evitando o número excessivo de parâmetros livres. Estes aspectos são fundamentais, pois um sistema sobre-dimensionado e preso aos detalhes do conjunto de treino originaria a degradação da sua capacidade de generalização.

Contrariamente ao procedimento em geral aconselhável, que consiste na utilização de um conjunto de dados para treinar o sistema, um outro para proceder à sua avaliação durante a fase de desenvolvimento, e ainda um terceiro conjunto para avaliar o seu desempenho final, neste sistema foi apenas utilizado o conjunto de teste (§2.3) conhecido por *February'89* para os dois processos de avaliação referidos. No essencial, a validação cruzada do desempenho do sistema[2] foi dispensada porque se considerou que um modelo acústico treinado com robustez como este não estaria significativamente "sintonizado" para o conjunto de dados *February'89* apenas pelo facto de o ter utilizado num mecanismo de decisão durante o treino. Considerou-se portanto que o mesmo conjunto de dados garantia uma estimativa bastante correcta do desempenho do sistema num conjunto de teste independente do treino. Deste modo, em relação às etapas 1 e 3, a variável taxa_do_erro refere a percentagem de palavras erradas, isto é, o total de palavras substituídas, inseridas e suprimidas, calculada sobre o conjunto de dados *February'89*.

---

[1] Seria certamente um mau indício se partes significativas do *codebook* fossem dispensadas seguindo esta estratégia, o que ocorreria se o valor de ε fosse demasiado elevado.

[2] Ou, em Inglês, *cross-validation training*.

2.9.3.2 Iniciação

Antes de aplicar o procedimento iterativo de adaptação dos parâmetros acústicos é necessário iniciá-los.

Embora o algoritmo de treino utilizado assegure o crescimento monótono da verosimilhança estendida ao conjunto de treino, não oferece qualquer garantia quanto à obtenção do máximo global dessa medida[1], nem sequer de que um "bom" máximo local seja atingido. Acresce que um eventual "bom" máximo local também não é suficiente para que o modelo acústico conduza efectivamente a bons resultados, dadas as limitações do critério de optimização (§2.9.2). Estas são, no essencial, as razões porque no desenvolvimento de alguns sistemas são efectuadas múltiplas sessões de treino a partir de diferentes iniciações, seleccionando então o sistema que apresenta os melhores resultados sobre um determinado conjunto de teste[2].

Apesar da natureza imprevisível dos aspectos referidos e também da notável capacidade das técnicas de treino dos HMM que lhes permitem "agarrar" adequadamente a informação acústica contida nos dados, mesmo partindo de um modelo relativamente deficiente, os resultados têm mostrado ser vantajoso iniciar os parâmetros livres com algum cuidado. Esta situação é tanto mais importante quanto maior é o número de parâmetros livres e portanto mais facilmente o sistema se possa enredar num deficiente processo de treino.

Com base nesta perspectiva, a iniciação deste modelo acústico decorreu em duas etapas: primeiro, estabelecendo o *codebook* inicial para cobrir o espaço acústico; depois, semeando o modelo de cada uma das unidades acústicas elementares.

Codebook

O procedimento que a seguir se apresenta foi utilizado para iniciar 512 gausseanas a partir da totalidade de dados de treino, ou seja, cerca de 10^6 vectores de características[3].

Começa por ser feita a transformação do conjunto de dados de modo a normalizar a variância global das componentes do vector de características.

Em seguida é construída uma árvore de decisão binária, com uma estrutura hierárquica compreendendo 9 níveis. O conjunto de dados é inicialmente atribuído ao

[1] Devido à elevada dimensão do espaço de optimização, a probabilidade de alcançar o máximo global será mesmo muito pequena.

[2] No desenvolvimento deste sistema isso não foi feito, além de que não seria fácil dadas as naturais limitações de recursos de computação.

[3] Anteriormente a este procedimento de inicialização do *codebook*, outras abordagens mais simples e utilizando apenas parte do conjunto de treino haviam sido experimentadas.

nó-raiz da árvore, definido no nível 0. Então, o seguinte procedimento recursivo é executado para cada nó, contendo um conjunto não vazio de vectores de características T, até atingir o nível 9:

1) aleatoriamente, 100 vectores são extraídos de T e os 2 mais afastados, $\boldsymbol{x}_A$ e $\boldsymbol{x}_B$, são identificados;

2) $\boldsymbol{x}_A$ e $\boldsymbol{x}_B$ são usados como centróides iniciais para correr diversas iterações do algoritmo *k-means*[1] sobre um máximo de 10^4 vectores aleatoriamente extraídos de T; utiliza-se a variação absoluta das componentes dos *centróides*, nas sucessivas iterações, como medida de distorção e critério de paragem do algoritmo;

3) T é repartido em 2 subconjuntos, T_A e T_B, de acordo com a distância Euclidiana dos vectores aos últimos *centróides* obtidos em 2); a média dos vectores em T_A e T_B é calculada e armazenada na estrutura de informação da árvore;

4) T_A e T_B são atribuídos aos nós sucessores.

Então, cada um dos vectores de características do conjunto de treino completo é "injectado" no nó-raiz e é encaminhado pela estrutura da árvore de acordo com a informação do *centróide* associado a cada nó, sendo finalmente atribuído a um dos 512 nós-terminal da árvore[2]. Para cada um desses nós, após a inversão da normalização dos vectores que lhe estão atribuídos são calculados os vectores média e variância, ficando assim definido o *codebook* inicial.

Probabilidades de transição e coeficientes das misturas

Apesar dos HMM poderem ser iniciados com uma reduzida quantidade de dados, neste sistema os parâmetros específicos de cada um dos modelos foram "semeados" utilizando o conjunto de treino completo, dada a relativa facilidade em segmentar e anotar automaticamente esse conjunto.

Cada um dos coeficientes da única mistura do HMM correspondente aos períodos de silêncio (§2.6.3.3) é iniciado contando, para todos os vectores de características que lhe estão associados de acordo com a segmentação automática, o número de vezes que a respectiva gausseana do *codebook* é a mais próxima de entre todas. Realizada essa contagem para todos os coeficientes, procede-se à sua normalização.

Para os restantes modelos acústicos, a sucessão de vectores de características correspondente a cada ocorrência do respectivo fonema é segmentada em 3 partes com comprimento semelhante, que são então atribuídas aos estados inicial, intermédio e final do HMM. Em cada estado, os coeficientes das misturas são iniciados seguindo um procedimento idêntico ao utilizado no modelo do silêncio. Relativamente às

[1] Neste caso, poder-se-ia dizer o algoritmo de *clustering 2-means*; note-se que este algoritmo opera aqui no espaço acústico já normalizado e utiliza a distância Euclidiana.

[2] Apesar da grande variabilidade do número de vectores de características nos nós terminais da árvore, não se verificou o esvaziamento de qualquer deles.

probabilidades de transição, são também iniciadas com as contagens normalizadas das transições entre estados, segundo esta segmentação linear.

2.9.3.3 Iteração nuclear de treino

Nas etapas 1 e 3 da estratégia de treino definida (§2.9.3.1), os parâmetros acústicos são iterativamente adaptados até se verificar a convergência de uma determinada medida do desempenho do sistema, baseada no número de palavras erradamente reconhecidas[1]. Nas iterações iniciais, enquanto a dimensão do sistema é relativamente reduzida, a evolução dessa taxa de erro é bastante suave mas com crescimento do número de parâmetros acústicos torna-se irregular. Assim, a decisão de paragem deste processo iterativo é estabelecida com base na filtragem da curva do erro, considerando a média aritmética de um determinado número de valores sucessivos.

No treino deste sistema a iteração nuclear de adaptação dos parâmetros baseia-se numa aproximação, análoga à adoptada no processo de descodificação (§2.8.3), ao critério da máxima verosimilhança (§2.9.2). A equação 2.59 é aproximada por:

$$\theta = \arg\max_{\theta} \prod_{n=1}^{N_T} p_{q_\theta}\left(\boldsymbol{X}^n | W_c^n\right) \tag{2.65}$$

com $p_q(\boldsymbol{X}|W)$ definida de acordo com a equação 2.56, isto é, no cálculo da verosimilhança associada a cada exemplo de treino é apenas considerado o melhor caminho pertencente ao modelo acústico correspondente a W.

O principal motivo para esta aproximação reside no facto de permitir um algoritmo de treino computacionalmente bastante mais simples. Contrariamente ao algoritmo Baum-Welch[2], que exige o cálculo das probabilidades *"forward"* e também *backward* [6], com esta aproximação é apenas necessário calcular as probabilidades *"forward"* (§2.6.3.2). Além disso, como não é necessário somar as verosimilhanças correspondentes a diferentes caminhos de estados no modelo acústico, é muito simples utilizar a transformação logarítmica evitando assim o processo de re-escalamento das verosimilhanças (§2.8.3).

Tem sido reconhecido que o algoritmo Baum-Welch ao basear-se no critério de treino tal como definido na equação 2.59, considerando portanto o espalhamento de cada vector de características por múltiplos estados, conduz em princípio a modelos

[1] O parâmetro *WER* definido em §2.9.3.4.

[2] Inicialmente tinha-se adaptado o código do algoritmo *Baum-Welch,* ou *forward-backward*, que havia sido anteriormente desenvolvido para treinar os CHMM num reconhecedor de palavras isoladas, sendo no entanto essa solução abandonada devido à bastante maior complexidade computacional.

mais suaves e possivelmente com maior capacidade de generalização [2]. No entanto, tem-se também verificado experimentalmente que os algoritmos Baum-Wech e Viterbi[1] estimam modelos acústicos próximos, acontecendo o mesmo relativamente à verosimilhança estendida ao conjunto de treino. De acordo com um resultado teórico [160] é assegurada a proximidade desta verosimilhança total se a dimensão dos vectores de características é elevado.

A satisfação do critério de treino correspondente à equação 2.65 pode em princípio ser alcançada treinando isoladamente os HMM correspondentes às unidades acústicas elementares, ou seja, adaptando de cada vez os parâmetros de um modelo com base nos segmentos extraídos do conjunto de treino e que supostamente lhe correspondem. No entanto, este procedimento exige a prévia segmentação e anotação da base de dados de treino, tarefa que pode ser simplesmente dispensada. No desenvolvimento deste sistema é utilizada uma outra abordagem, conhecida por treino embebido, em que para cada exemplo o procedimento de treino é aplicado simultaneamente a todos os HMM correspondentes às unidades acústicas elementares que, de acordo com a transcrição ortográfica e o léxico, estão associadas a essa frase.

Assim, em cada iteração de adaptação dos parâmetros e para cada exemplo do conjunto de treino, procede-se inicialmente à composição do modelo acústico correspondente à respectiva frase. Para além dos HMM correspondentes às unidades acústicas associadas à frase através da transcrição ortográfica e do léxico, são inseridos modelos de silêncio obrigatórios nas extremidades da frase e opcionais entre as palavras. Numa perspectiva baseada no chamado HMM global do sistema (§2.8.2) ,é fixado um único caminho de palavras correspondente à frase W_c^n.

Em seguida é executado o algoritmo Viterbi (§2.8.3) para alinhar a sucessão de vectores de características com os estados dos HMM que integram o modelo acústico composto para W_c^n. Note-se que contrariamente à situação que se verifica no módulo de descodificação, em que o espaço de pesquisa é muito grande e portanto é necessário recorrer a técnicas para a sua redução, aqui o algoritmo Viterbi pode ser aplicado plenamente.

São então actualizadas as variáveis que permitem, após o varrimento completo do conjunto de treino a cada iteração, re-estimar os parâmetros acústicos de acordo com as fórmulas que a seguir se apresentam. A não ser que tenha sido já alcançado um ponto crítico, aos novos parâmetros acústicos corresponderá uma maior verosimilhança $p_q(\boldsymbol{X}|W)$ estendida ao conjunto de treino [68].

As probabilidades de transição entre estados são re-estimadas aproximando o cálculo que decorre da definição (§2.6.3.2), $a_{i,j} = P(q_t = i, q_{t+1} = j) / P(q_t = i)$, pelo

[1] É usada esta designação devido ao alinhamento Viterbi do sinal com os estados dos HMM, procedimento que é basilar neste algoritmo de treino.

resultado da normalização do número de vezes[1] que, de acordo com a segmentação Viterbi de cada exemplo, as transições são verificadas no conjunto de treino:

$$a_{i,j} = \frac{\sum_{n=1}^{N_T} \sum_{t=1}^{T''-1} \delta(q_t = i, q_{t+1} = j)}{\sum_{n=1}^{N_T} \sum_{t=1}^{T''-1} \delta(q_t = i)} \tag{2.66}$$

De acordo com a equação 2.46 (§2.6.3.3) o peso relativo da componente $c_{j,m} N_{\alpha_{j,m}}(\boldsymbol{x})$ da mistura $b_j(\boldsymbol{x})$, na observação de $\boldsymbol{x}$, é dada pelo quociente:

$$\gamma_{j,m}(\boldsymbol{x}) = \frac{c_{j,m} N_{\alpha_{j,m}}(\boldsymbol{x})}{b_j(\boldsymbol{x})} \tag{2.67}$$

Assim, os coeficientes das misturas gausseanas são re-estimados a partir da fórmula recursiva:

$$c_{j,m} = \frac{\sum_{n=1}^{N_T} \sum_{t=1}^{T''} \delta(q_t = j) \gamma_{j,m}(\boldsymbol{x}_t)}{\sum_{n=1}^{N_T} \sum_{t=1}^{T''} \delta(q_t = j)} \tag{2.68}$$

O *codebook* $\{(\mu_m, \sigma_m),\ m = 1,\ldots,4096\}$ (§2.6.3.3) é treinado[2] através das fórmulas:

$$\mu_m = \frac{\sum_{n=1}^{N_T} \sum_{t=1}^{T''} \sum_{j} \delta(q_t = j) \gamma_{j,m}(\boldsymbol{x}_t) \boldsymbol{x}_t}{\sum_{n=1}^{N_T} \sum_{t=1}^{T''} \sum_{j} \delta(q_t = j) \gamma_{j,m}(\boldsymbol{x}_t)} \tag{2.69}$$

[1] Na fórmula que se apresenta é utilizada a função $\delta(proposiçao)$ com, $\delta(\mathsf{V}) = 1$ e $\delta(\mathsf{F}) = 0$.

[2] Devido a um "descuido" na sessão inicial de treino do modelo acústico só eram adaptados os parâmetros $a_{i,j}$ e $c_{j,m}$, sendo o *codebook* apenas iniciado; comparando os resultados dessas iterações de treino com os resultados obtidos após a correcção desse lapso, constatou-se que a redução do erro era superior a 30%$_{rel}$.

$$\sigma_m^2 = \frac{\sum_{n=1}^{N_T}\sum_{t=1}^{T^n}\sum_j \delta(q_t = j)\gamma_{j,m}(\boldsymbol{x}_t)(\boldsymbol{x}_t - \mu_m)^2}{\sum_{n=1}^{N_T}\sum_{t=1}^{T^n}\sum_j \delta(q_t = j)\gamma_{j,m}(\boldsymbol{x}_t)} \tag{2.70}$$

2.9.3.4 Resultados

Nesta secção são apresentados resultados correspondentes à avaliação do desempenho do reconhecedor, durante a fase de treino do modelo acústico, usando o conjunto de teste *February'89* (§2.3).

Para determinar as medidas de erro ao nível da palavra foi desenvolvido um algoritmo para alinhamento das frases reconhecidas com as respectivas frases correctas. São então contadas as palavras erradas, que aqui se considera ser o número total das palavras substituídas, inseridas e suprimidas[1]. Obtém-se assim a seguinte taxa de erro[2]:

$$WER = \frac{total\ palavras\ erradas}{total\ palavras}\ (\%) \tag{2.71}$$

Refira-se que o número *total de palavras* no conjunto de dados de teste *February'89* é 2484. O número *total de frases* é 300, sendo a percentagem de frases correctas considerado também como um indicador do desempenho do sistema:

$$SA = \frac{total\ frases\ correctas}{total\ frases}(\%) \tag{2.72}$$

A notação utilizada para referir uma configuração particular do modelo acústico consiste no par de números $(N_1;N_2)_I$, em que N_1 é o número médio de gausseanas utilizada por cada mistura e N_2 é a dimensão do *codebook*. Assim, N_1 pode ser reduzido na etapa 2 da estratégia de treino (§2.9.3.1) e é duplicado na etapa 4, altura em que duplica também o valor de N_2. A este par pode ser apenso o índice I referindo a iteração de treino correspondente.

[1] Na avaliação de alguns sistemas são utilizadas medidas ligeiramente diferentes, mas que no entanto são similares no essencial.

[2] Utilizada como critério de paragem na estratégia de treino definida (§2.9.3.1).

O sistema foi iniciado com a configuração $(512;512)_0$, com as medidas *WER* e *SA* apresentando o valor, respectivamente, 24,48% e 32,00%. Esses resultados estão mostrados nas figuras 2.18 e 2.19. Foram executadas 3 iterações de adaptação dos parâmetros, com *WER* atingindo o valor 14,05%. Foram então eliminados os 256 menores coeficientes de cada mistura gausseana, resultando um sistema $(256;512)_3$ com *WER* igual a 14,01%. Note-se que este foi um procedimento diferente do apresentado na estratégia de treino (§2.9.3.1), no entanto, nesta fase do treino considerou-se ser uma boa solução começar por reduzir as misturas a metade. Foram executadas mais 3 iterações de adaptação dos parâmetros, tendo *WER* atingindo o valor 12,08%. Utilizando um valor de ε igual a 10^{-5} as misturas foram novamente reduzidas[1], resultando um sistema com a configuração $(116;512)_6$. A figura 2.20 mostra o histograma da distribuição da dimensão das misturas. A taxa do erro, *WER*, alcançou o valor 12,04%. Dado que *WER* não diminuíu significativamente correndo mais iterações de adaptação com esta configuração, o *codebook* existente na 6ª iteração foi duplicado (2.64), resultando o sistema $(232;1K)_6$ com *WER* e *SA* iguais a, respectivamente, 12,00% e 57,33%.

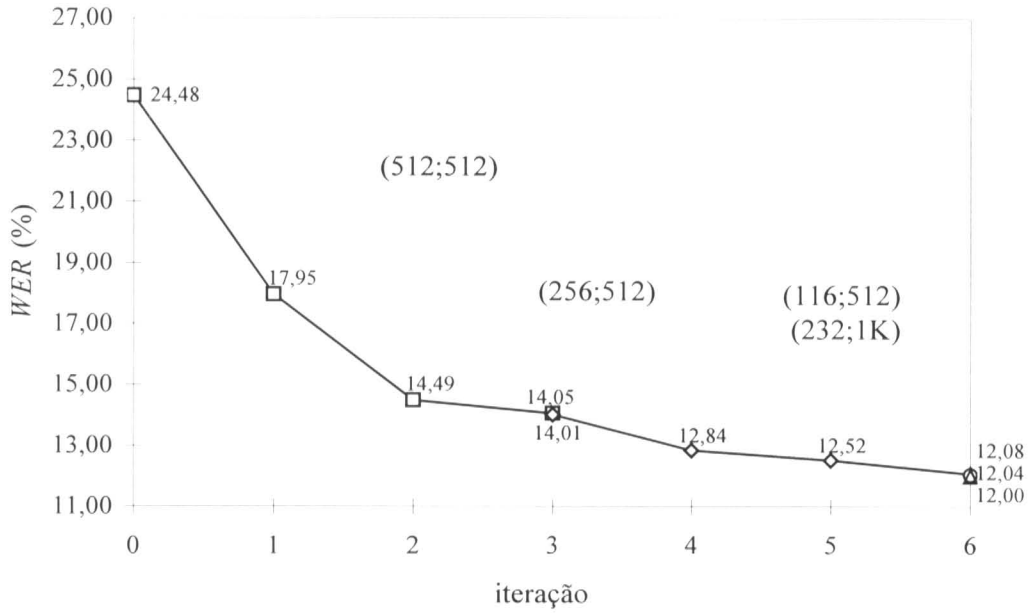

Fig. 2.18 - Evolução, durante as 6 iterações inicias, da taxa de palavras erradas.

[1] De acordo com a etapa 2 do ciclo de treino da estratégia definida (§2.9.3.1).

BIBLIOTECA F.E. PORTO

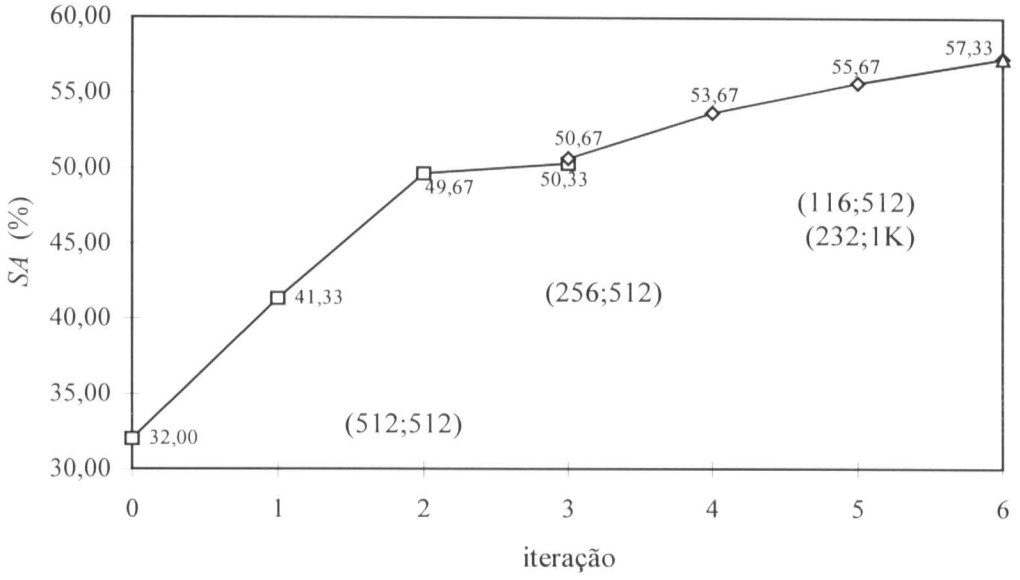

Fig. 2.19 - Evolução, durante as 6 iterações iniciais, da taxa de frases correctas.

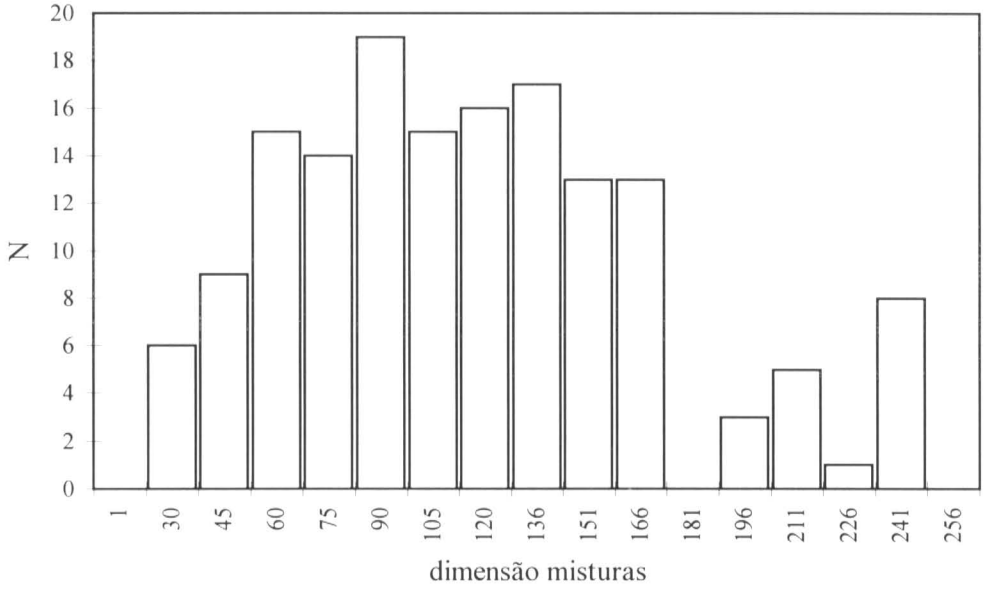

Fig. 2.20 - Histograma da dimensão das misturas do sistema (116;512).

A partir do sistema $(232,1K)_6$, 7 iterações de adaptação dos parâmetros foram executadas. As figuras 2.21 e 2.22 mostram a evolução do desempenho do sistema. Então, com o parâmetro ε igual a 10^{-5}, as misturas gausseanas foram reduzidas resultando o sistema $(159,1K)_{13}$ apresentando *WER* igual a 9,46%. A figura 2.23 mostra o histograma da distribuição da dimensão das gausseanas. Uma vez que a taxa de erro *WER* não decresceu significativamente correndo mas iterações com o sistema

(159,1K), o *codebook* existente na 13ª iteração foi duplicado resultando o sistema $(318;2K)_{13}$, com *WER* mantendo o valor de 9,46%.

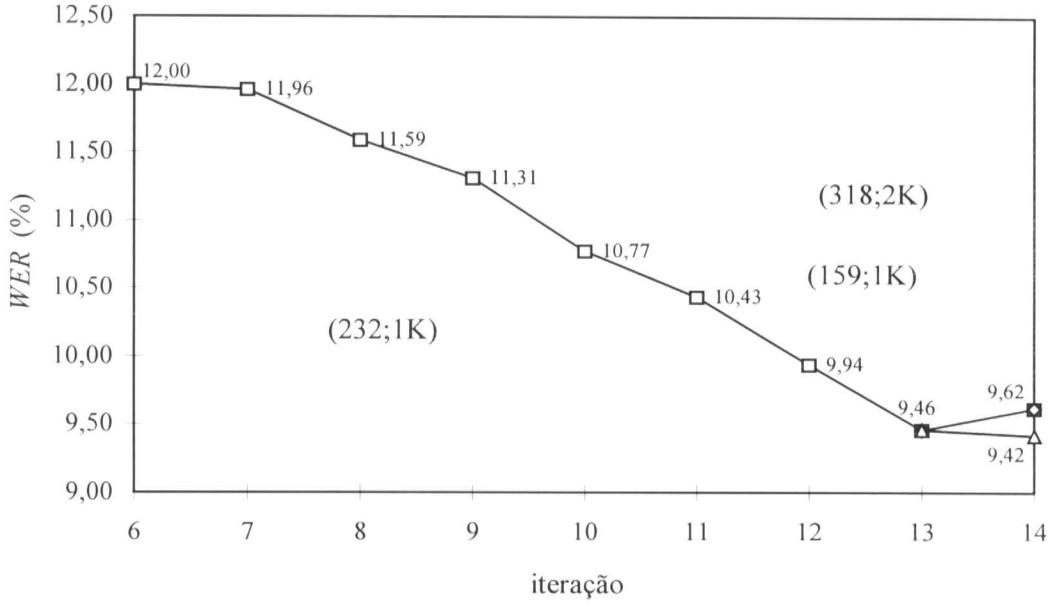

Fig. 2.21 - Evolução da taxa de palavras erradas.

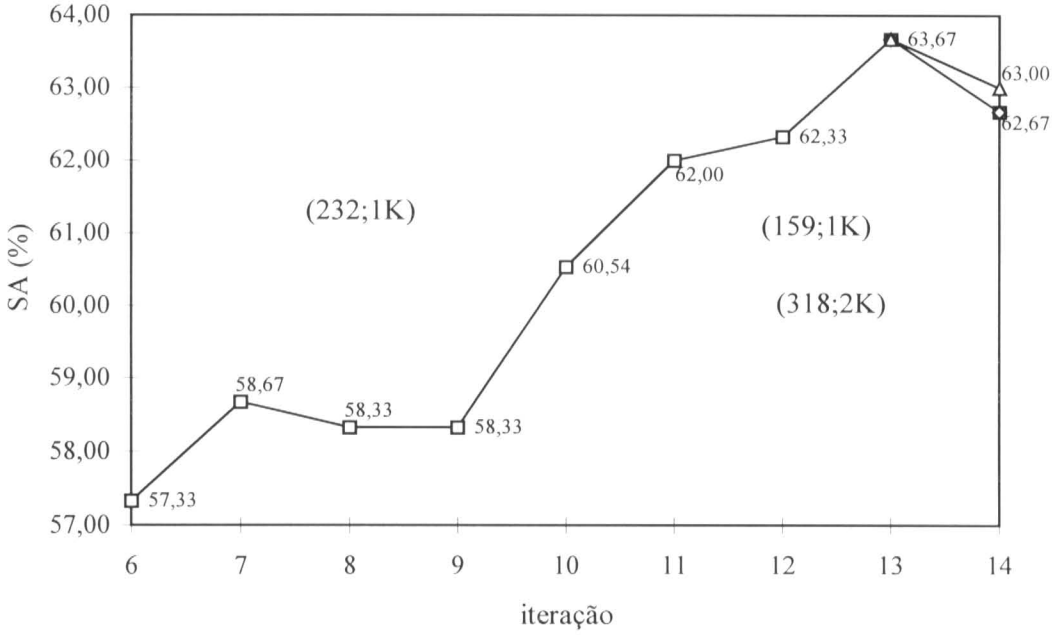

Fig. 2.22 - Evolução da taxa de frases correctas.

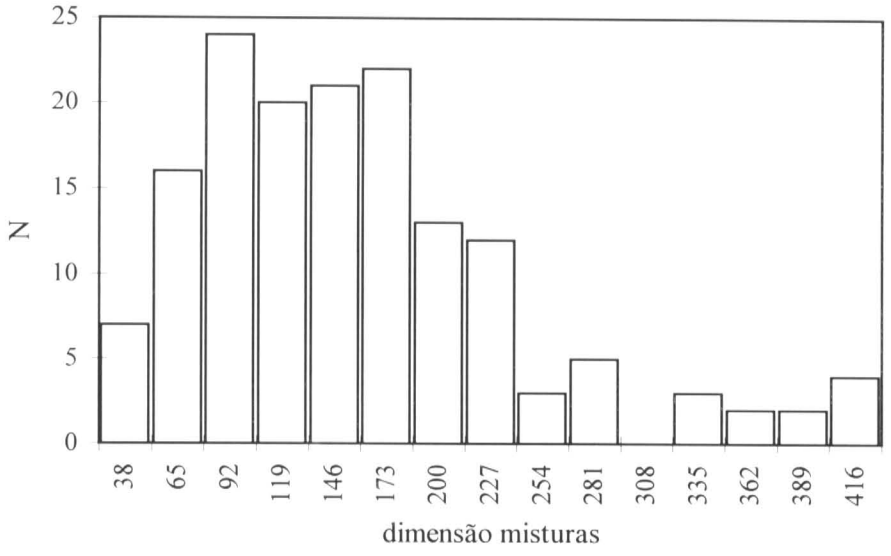

Fig. 2.23 - Histograma da dimensão das misturas do sistema (159,1K).

A partir do sistema $(318{,}2K)_{13}$, 10 iterações de adaptação dos parâmetros foram executadas. O desempenho do sistema evoluiu como mostram as figuras 2.24 e 2.25. Com o parâmetro ε igual a 10^{-5}, as misturas gausseanas na iteração 23 foram reduzidas resultando o sistema $(205{,}2K)_{23}$ com *WER* igual a 7,53%.

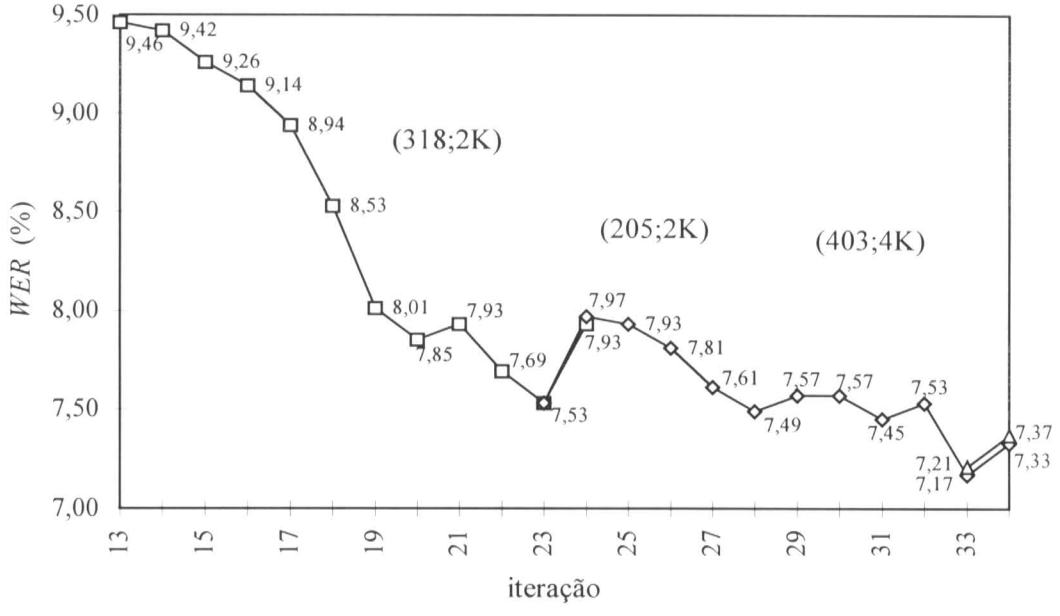

Fig. 2.24 - Evolução da taxa de palavras erradas.

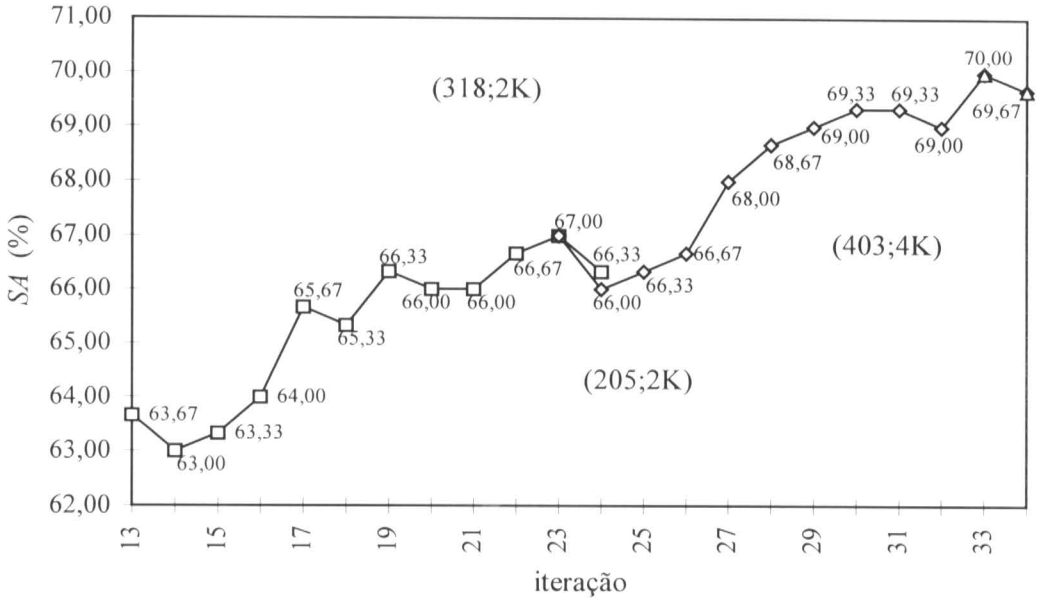

Fig. 2.25 - Evolução da taxa de frases correctas.

A figura 2.26 mostra o histograma da distribuição da dimensão das gausseanas. O treino foi retomado executando 10 iterações a partir dos modelos existentes na iteração 23, tendo *WER* alcançado o valor 7,17%. Então, o *codebook* foi duplicado resultando o sistema $(403;4K)_{33}$, com *WER* apresentando o valor de 7,21%.

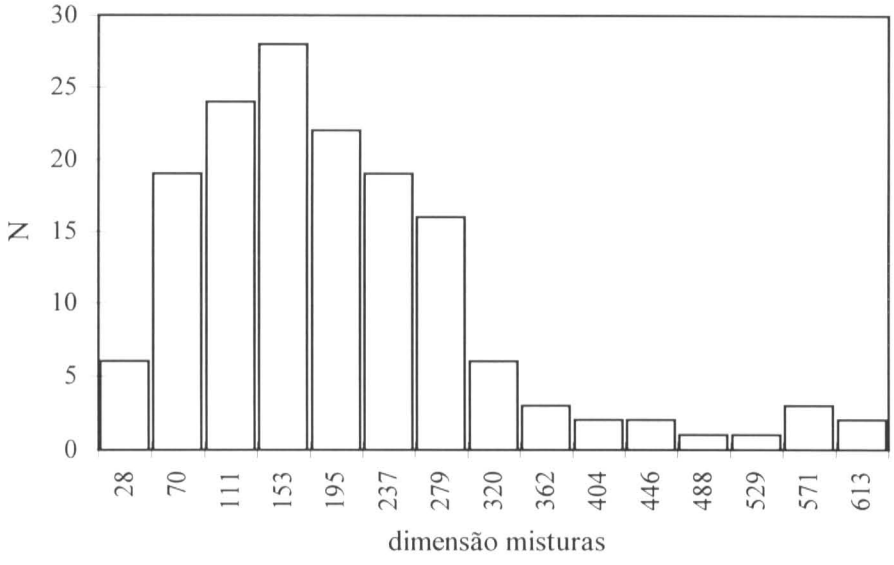

Fig. 2.26 - Histograma da dimensão das misturas do sistema (205,2K).

A partir do sistema $(403,4K)_{33}$, foram executadas 3 iterações de adaptação dos parâmetros. As figuras 2.27 e 2.28 mostram a evolução do desempenho do sistema. Na 36ª iteração, *WER* alcançou o valor 6,76%. Com ε igual a 10^{-5} as misturas gausseanas na iteração 36 foram reduzidas resultando o sistema $(353,4K)_{36}$ cujo *WER* se manteve igual a 6,76%.

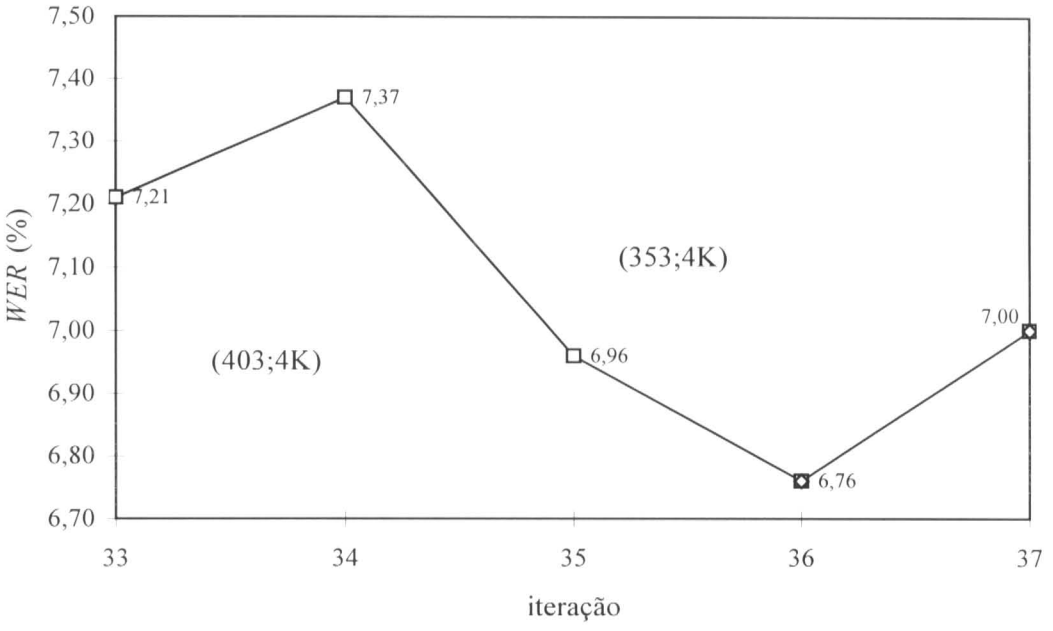

Fig. 2.27 - Evolução da taxa de palavras erradas.

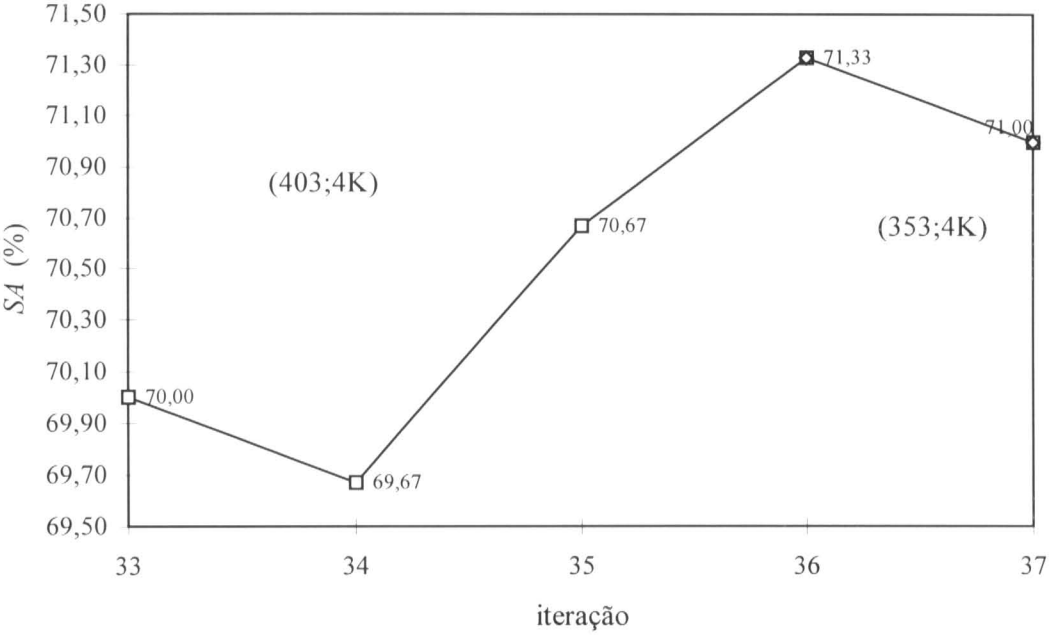

Fig. 2.28 - Evolução da taxa de frases correctas.

A figura 2.29 mostra o histograma da distribuição da dimensão das gausseanas. O prolongamento do treino não melhorou significativamente o desempenho do sistema.

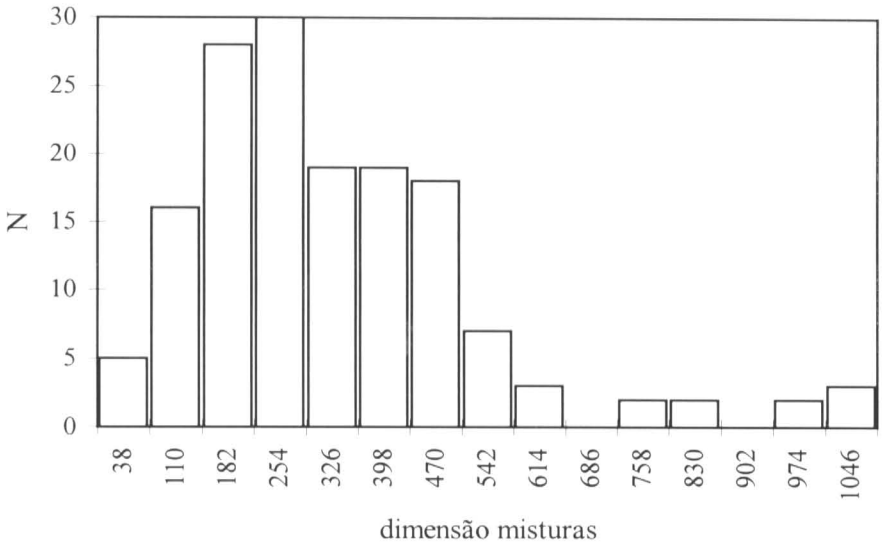

Fig. 2.29 - Histograma da dimensão das misturas do sistema (353,4K).

Duplicando novamente o *codebook*, agora com 8K gausseanas, e correndo 16 iterações de adaptação dos parâmetros verificou-se uma redução do *WER* para o valor mínimo 6,60%, na iteração 41, após reduzir novamente as misturas, com a configuração do sistema (494,8K). O número de parâmetros livres revelou-se ser já demasiado grande para ser treinado de maneira estável com o conjunto de treino disponível. Isso mesmo pôde ser constatado pela acentuada irregularidade das curvas *WER* e *SA*.

Assim, os parâmetros acústicos correspondentes ao sistema $(353,4K)_{36}$ foram tomados como o resultado do processo de treino. A figura 2.30 mostra a utilização do *codebook* correspondente a este sistema sendo possível verificar-se, por exemplo, que aproximadamente 60 funções de Gauss são utilizadas em apenas uma das misturas, ou que a maior parte das gausseanas é utilizada por cerca de 1 dezena de misturas diferentes.

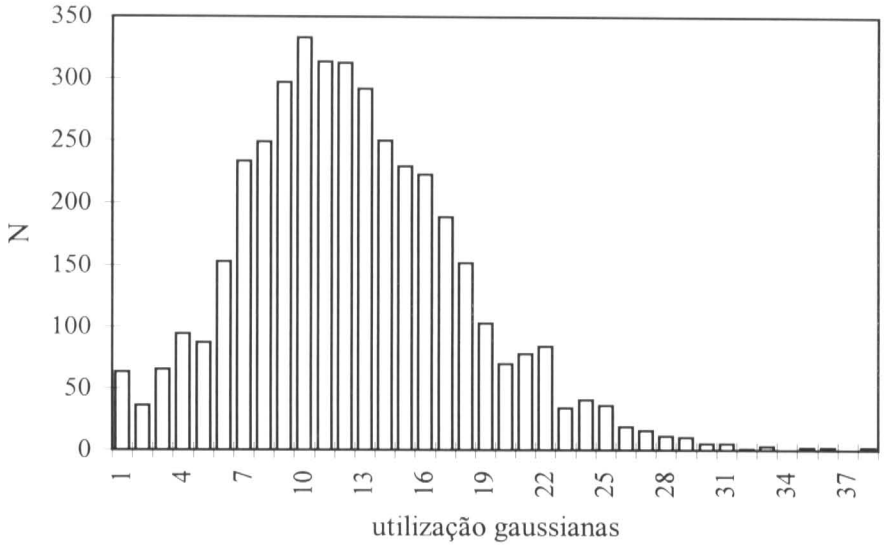

Fig. 2.30 - Histograma da utilização do *codebook* pelo sistema (353,4K).

As figuras 2.31 e 2.32 mostram a evolução completa das medidas *WER* e *SA* durante o treino do sistema.

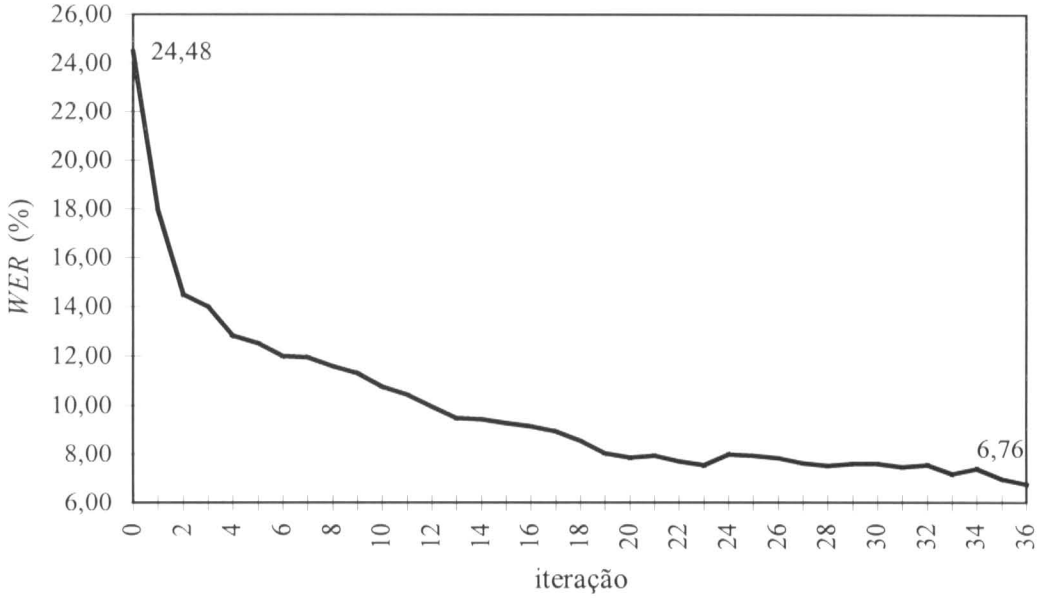

Fig. 2.31 - Evolução da taxa de palavras erradas ao longo do treino.

A figura 2.33 mostra a evolução do logaritmo da verosimilhança do modelo acústico observar o conjunto de treino. Na curva são evidentes os períodos de

acentuado acréscimo da verosimilhança que se verificam logo após a operação de duplicação do *codebook*, ao que se segue um período de gradual "saturação"[1].

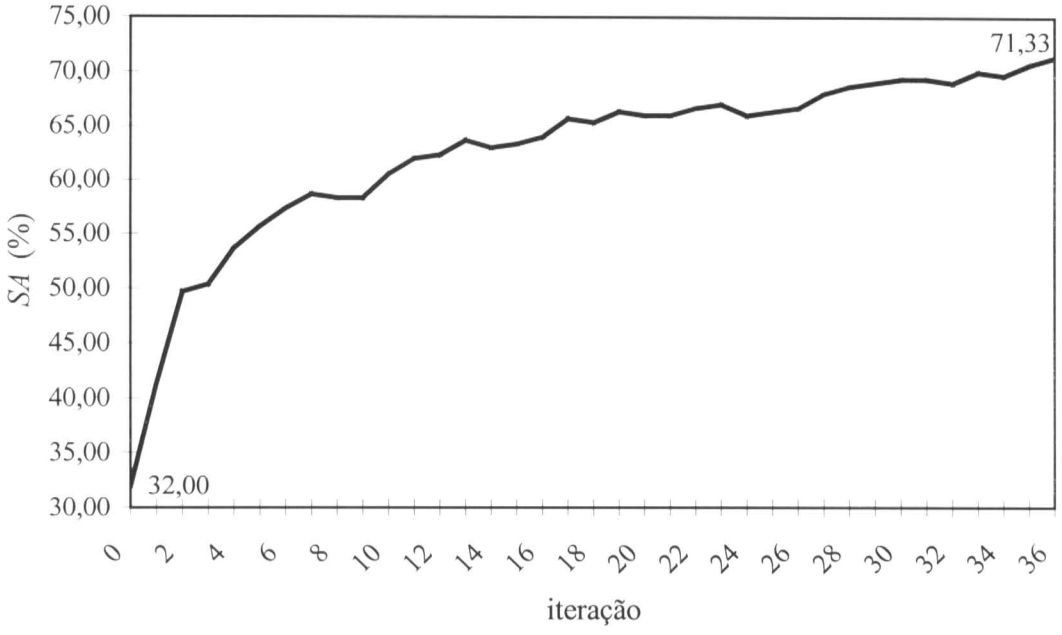

Fig. 2.32 - Evolução da taxa de frases correctas.

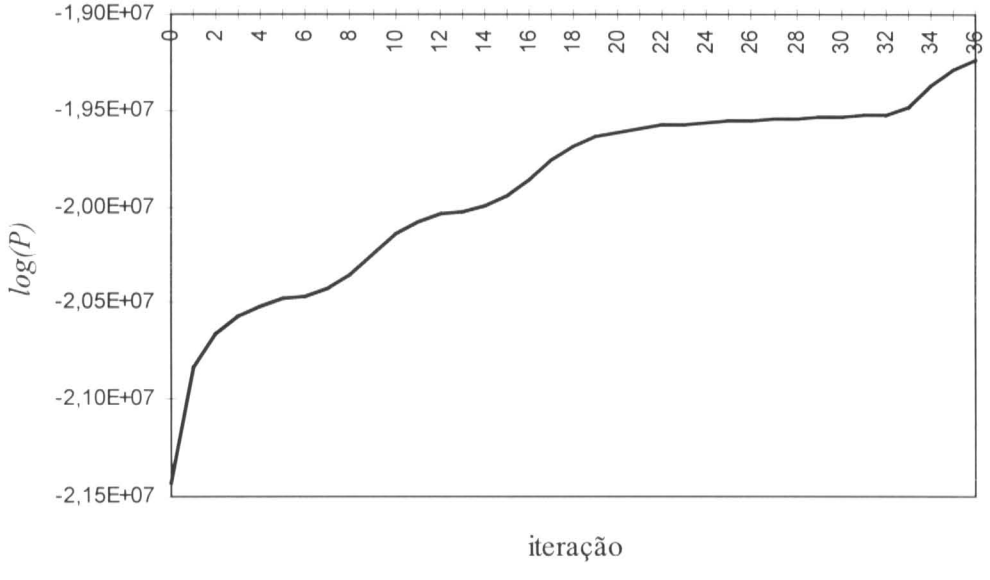

Fig. 2.33 - Evolução da log-verosimilhança acumulada sobre o conjunto de treino.

[1] Ou mesmo de redução da verosimilhança, como é evidente quando o sistema é reduzido da configuração (512;512) para (256;512).

As figuras 2.34 e 2.35 mostram a evolução das medidas de erro correspondentes às palavras substituídas (*WER_Sub*), inseridas (*WER_Ins*) e suprimidas (*WER_Sup*):

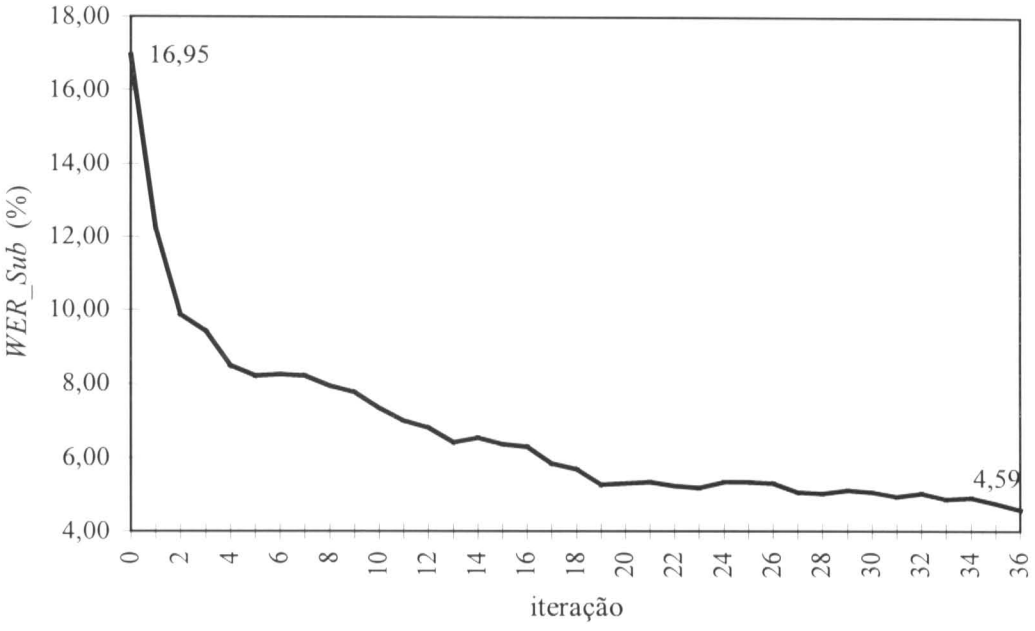

Fig. 2.34 - Evolução da taxa de erro correspondente às palavras substituídas.

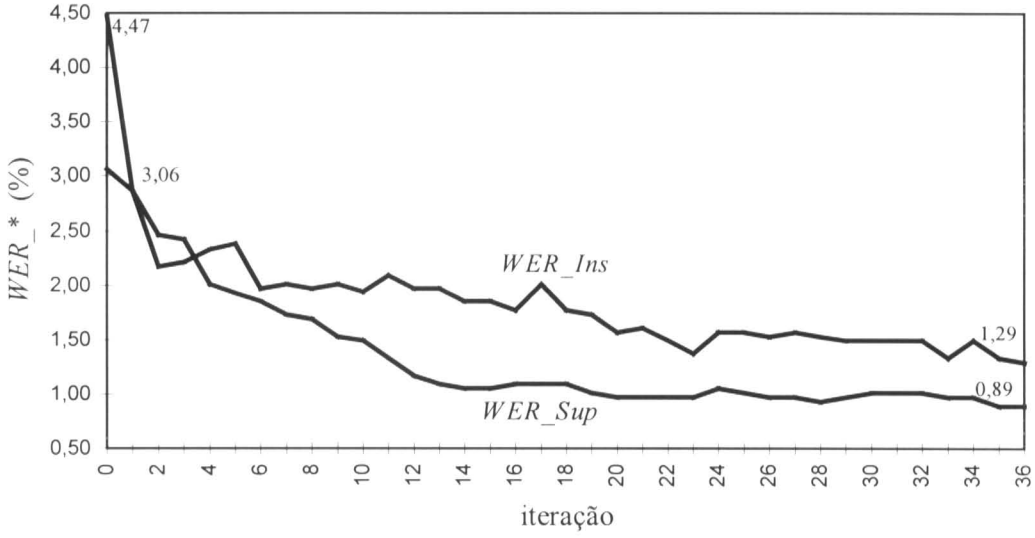

Fig. 2.35 - Evolução do erro correspondente às palavras inseridas e suprimidas.

Como se pode verificar o número de palavras inseridas e suprimidas é bastante próximo, sendo no entanto possível aproximar ainda mais estas medidas utilizando no

algoritmo de descodificação um parâmetro, *WP*[1], que penaliza a inserção ou a supresão de palavras. Essa constante é geralmente adicionada ao logaritmo da verosimilhança à saída de cada palavra, de modo que a relação *WER_Ins*/*WER_Sup* cresce ou decresce dependendo de *WP* ser, repectivamente, suficientemente positivo ou negativo. Um critério geralmente considerado mais importante que este, não necessariamente concordante[2], consiste na optimização de *WP* com o objectivo de minimizar a taxa de erro *WER*.

No descodificador foi introduzido um outro parâmetro empírico, α, utilizado para estabelecer o peso relativo do modelo acústico. Embora também relativamente a α não exista uma justificação teórica consistente, a sua utilização é generalizada dados os resultados práticos que atestam o seu mérito. Uma vez que se utiliza a transformação logarítmica, esse parâmetro é aplicado como um factor resultando a expressão $\alpha \log\left(p(\boldsymbol{X}|W)\right)$.

No desenvolvimento deste sistema os parâmetros *WP* e α foram optimizados simultaneamente. A tabela 2.10 confronta os resultados obtidos sem a sua utilização no descodificador, ou seja, considerando os seus valores neutros, com os resultados obtidos após o procedimento de optimização com o critério de minimizar *WER*. Como se pode verificar obteve-se uma melhoria da taxa de erro de aproximadamente $4{,}1\%_{\text{rel}}$, que embora seja um valor menor do que o esperado é já significativo.

	WER	*WER_Sub*	*WER_Ins*	*WER_Sup*	*SA*
WP=0.00; α=1.00	6,76%	4,59%	1,29%	0,89%	71,33%
WP=0.45; α=0.60	6,48%	4,43%	0,97%	1,09%	71,67%

Tab. 2.10 - Desempenho do sistema $(353{,}4\text{K})_{36}$ antes e após a optimização de *WP* e α.

[1] Do Inglês, *word penalty*.

[2] Com efeito, também neste sistema se verificou que a optimização correspondente a cada um deste dois objectivos conduz a parâmetros *WP* diferentes.

Capítulo 3

Sistema baseado em múltiplas características do sinal

Neste capítulo é exposta uma abordagem original ao problema do reconhecimento automático da fala contínua, baseada no processamento simultâneo de diferentes características do sinal segundo um formalismo estatístico conhecido. Paralelamente à apresentação das soluções experimentadas e aplicadas no desenvolvimento de um sistema baseado em múltiplas características[1], são também introduzidos e discutidos alguns aspectos fundamentais nesta área específica de investigação. Finalmente, são apresentados e analisados os resultados obtidos numa tarefa de reconhecimento de fala contínua.

3.1 Introdução

A combinação de múltiplas fontes de informação, aos diferentes níveis, constitui uma área de investigação importante no desenvolvimento de reconhecedores de elevado desempenho.

Ao nível da modelação acústica, as técnicas de *multi-streaming* têm vindo a adquirir um assinalável relevo. Nos reconhecedores *multi-stream* [190] são executadas múltiplas análises do sinal da fala, e eventualmente de sinais correlacionados, gerando vários *streams* que representam informação discriminante diferente. Esses *streams* são inicialmente processados de maneira independente e paralela, sendo numa fase posterior recombinados num processo integrado de decisão.

Uma das principais motivações para estes métodos reside na dificuldade, ou até mesmo inviabilidade, em fundir eficientemente numa única representação a informação discriminante disponível no sinal da fala. Um exemplo particularmente esclarecedor verifica-se em alguns sistemas que combinam características extraídas através da análise espectral do sinal em janelas de curta duração, com características que representam dependências estatísticas de mais longo alcance, à escala silábica.

---

[1] No sentido de tornar a exposição mais simples, utilizar-se-á frequentemente o termo *stream* com o significado "sucessão de vectores de características" e os termos *single-stream* e *multi-stream* para referir os sistemas baseados, respectivamente, em apenas um único *stream*, ou em vários.

Mas mesmo se a resolução temporal dos diversos *streams* é idêntica, uma abordagem assente no seu processamento independente durante os estádios iniciais do reconhecimento pode ser vantajosa. Por exemplo, os modelos acústicos *multi-stream* suportam facilmente a ocorrência de períodos assíncronos de quase-estacionaridade nos diversos *streams*, enquanto nos sistemas convencionais as componentes da única representação do sinal são geralmente forçadas ao processamento síncrono.

A constatação de que diferentes *streams* podem ser corrompidos em grau substancialmente diferente por determinadas perturbações no canal acústico proporciona uma excelente oportunidade para a utilização desta abordagem na área do reconhecimento robusto da fala. É o caso, por exemplo, das técnicas baseadas na decomposição do espectro do sinal em bandas que são então associadas a *streams* diferentes[1], ou a inclusão de *streams* com informação visual no processo de reconhecimento.

Os aspectos acabados de referir esboçam alguns dos méritos mais relevantes dos sistemas *multi-stream* e apontam algumas linhas de trabalho bastante seguras no caso do objectivo principal ser simplesmente alcançar taxas de erro favoráveis a esta tecnologia, relativamente recente, no confronto directo com os sistemas *single-stream*.

No entanto, neste trabalho foi seguida uma direcção que tanto quanto é do conhecimento do autor não tem sido objecto preferencial de investigação. Pretendeu-se investigar as potencialidades deste formalismo estatístico quando se combinam representações do sinal acústico que encerram informação diferente mas que não são necessariamente seleccionadas de acordo com critérios particularmente favoráveis a esta tecnologia. Um aspecto importante deste trabalho consiste precisamente na proposta de utilização dos sistemas *multi-stream* transpondo as aplicações típicas desta tecnologia.

Foi estabelecido um objectivo concreto consistindo no desenvolvimento de um reconhecedor *multi-stream* que alcançasse melhor desempenho que um sistema *single-stream* baseado em características *standard*, os coeficientes mel-cepstrais (MFCC). No reconhecedor *multi-stream* as características MFCC são combinados com um *stream* baseado nos coeficientes lineares cepstrais (LFCC). Embora este desafio se apresentasse à partida como bastante difícil, atendendo à significativa correlação entre ambas as características, o sistema desenvolvido permitiu a redução da taxa de palavras erradas em aproximadamente $10\%_{rel}$.

Tendo este sistema obtido resultados positivos com uma combinação de características que não se afigura particularmente favorável, é legítimo esperar com elevada confiança que as soluções técnicas propostas, ou outras nelas inspiradas, apresentem um desempenho ainda melhor se as condições forem mais favoráveis.

---

[1]Com base na pesquisa de documentação levada a cabo durante este trabalho, pode-se afirmar com razoável confiança que uma grande parte das publicações na área do *multi-streaming* trata o caso particular dos chamados sistemas multi-banda.

Durante o desenvolvimento de um sistema *multi-stream* surgem tipicamente algumas questões que são essenciais neste tipo de abordagens. Quais os *streams* a utilizar? Como recombiná-los? Como executar de maneira eficiente a descodificação do sinal? Estas questões são tratadas nas secções seguintes, com particular atenção ao sistema *multi-stream* desenvolvido neste trabalho para uma tarefa de reconhecimento idêntica à do sistema *single-stream* apresentado no capítulo anterior.

3.2 Estrutura do reconhecedor

A estrutura do reconhecedor *multi-stream*, no nível conceptual correspondente ao diagrama de blocos representado na figura 3.1, é semelhante à do sistema *single-stream* (§2.4) desenvolvido também neste trabalho.

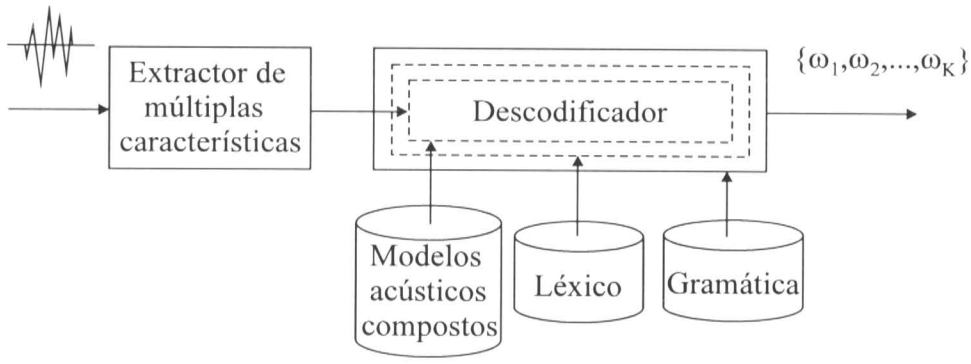

Fig. 3.1 - Diagrama de blocos do reconhecedor *multi-stream*.

No entanto, apenas os módulos correspondentes ao léxico e à gramática (§2.7) se mantêm inalterados. Em ambos os sistemas, as restrições linguísticas integradas na estrutura de informação (§2.8) correspondem à mesma gramática *word-pair* e ao mesmo léxico de pronunciação única.

No módulo extractor de características deste sistema são efectuadas paralelamente três análises distintas do sinal acústico. As sucessões de vectores de características resultantes de dois desses procedimentos de análise baseiam-se na representação cepstral do sinal correspondente a resoluções espectrais diferentes. Além destes *streams* é também extraída uma estimativa da média da frequência fundamental do sinal correspondente aos sons vocálicos.

Relativamente ao descodificador, tal como no sistema *single-stream*, baseia-se no algoritmo de pesquisa Viterbi (§2.8), tendo sofrido apenas as alterações necessárias para poder operar com modelos acústicos elementares que suportam múltiplos *streams* de características.

Ao nível acústico, o reconhecimento baseia-se no processamento paralelo dos dois *streams* com o objectivo de calcular, para cada hipótese de reconhecimento, uma estimativa da verosimilhança da sua observação conjunta. Tal como mostra a figura 3.2, esquematizando as operações realizadas no modelo acústico de uma unidade simbólica genérica, M, os *streams* são processados de maneira independente até ser obtida a verosimilhança das observações associadas a cada um deles, $p\left(\overset{1}{\boldsymbol{X}}|M_1\right)$ e $p\left(\overset{2}{\boldsymbol{X}}|M_2\right)$. Estes valores são então operados pelo modelo de recombinação[1], definido em função da classe associada à unidade simbólica e dinamicamente adaptado ao valor da média da frequência fundamental, o qual calcula o resultado pretendido.

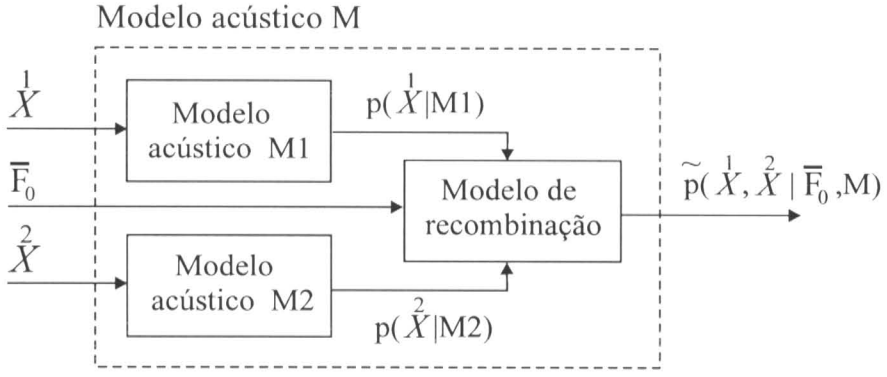

Fig. 3.2 - Modelo acústico *composto*, correspondente à unidade simbólica M.

Para cada uma das 52 unidades elementares definidas ao nível fonético neste sistema, existe um modelo acústico, baseado na tecnologia dos SCHMM, que emula o processo esquematizado na figura 3.2.

3.3 Análise do sinal

3.3.1 Introdução

Nos sistemas *multi-stream*, a informação discriminante extraída do sinal da fala, ou de outro sinal correlacionado, é repartida e possivelmente também compartilhada por vários *streams* de características. Deve ser realçado que esta metodologia se verifica independentemente de ser ou não viável condensar as várias representações

[1] Diz-se "recombinação" porque os *streams* de informação, associados às verosimilhanças parcelares e extraídas do mesmo sinal da fala, são novamente combinadas.

numa única entidade matemática, como seja uma sucessão de vectores de características pois, tal como já se referiu, essa solução pode apresentar vantagens significativas.

Ao desenvolver um sistema *multi-stream* para uma determinada aplicação, uma questão essencial reside, naturalmente, na escolha das representações mais adequadas para o efeito.

Um exemplo notável, face aos resultados que apresenta e também por se basear em fontes de informação com natureza tão diversa, consiste em determinados sistemas que emulam a capacidade humana de percepção audiovisual. São em geral desenvolvidos para tarefas de reconhecimento caracterizadas por condições acústicas extremamente adversas e modelos linguísticos relativamente simples. Tal como seria previsível, nessas condições tem-se registado um aumento de eficácia assinalável complementando a análise acústica *standard* com informação visual, correlacionada com o sinal da fala, obtida analisando a evolução de alguns traços relevantes da expressão facial do falante [180,181,182].

Com efeito, dadas as potencialidades dos sistemas *multi-stream* face ao problema da robustez no reconhecimento automático da fala, baseadas no princípio simples de que as perturbações do canal podem deixar relativamente incólumes determinados parâmetros de representação apesar de afectar os restantes, é natural que este tenha constituído desde logo uma das motivações essenciais destas abordagens. Um bom exemplo consiste nos chamados reconhecedores multi-banda [163,191], cujos *streams* estão associados à representação espectral do sinal em diferentes bandas. Para além de argumentos sólidos que lhe têm sido atribuídos, baseados em conhecimento na área da psico-acústica [47], esses sistemas têm obtido resultados assinaláveis [162,164] quando o sinal da fala se encontra contaminado por ruído aditivo e localizado em algumas bandas espectrais.

Em alguns sistemas é extraída e combinada a informação associada a unidades acústicas que subsistem em diferentes escalas temporais, gerando assim múltiplos *streams* vocacionados para representar dependências estatísticas que ocorrem com diferente alcance no sinal da fala. Nestes sistemas, designados multi-escala [45], são de destacar, também pelo significativo número de resultados que têm sido publicados registando o seu assinalável desempenho, os sistemas que combinam características à escala sub-fonética com outras à escala silábica [168,192].

O reconhecedor desenvolvido neste trabalho utiliza características resultantes de técnicas de análise do sinal que são correntes em sistemas *single-stream*, nomeadamente, um dos *streams* de características baseia-se nos MFCC e o outro nos LFCC.

Antes de tratar os aspectos particulares relacionados com os *streams* seleccionados é interessante estabelecer algumas noções gerais, recorrendo a conceitos

da Teoria da Informação, sobre a utilização de múltiplas representações do sinal no problema do reconhecimento automático da fala.

3.3.2 Conceitos de Teoria da Informação

Considere-se que são extraídos N *streams* de características do sinal da fala, $s(k)$, correspondente à frase, W, que se pretende reconhecer:

$$s(k) \rightarrow \overset{1}{\boldsymbol{X}},\ \overset{2}{\boldsymbol{X}},\ ...,\ \overset{N}{\boldsymbol{X}} \tag{3.1}$$

Embora se possam admitir outras estruturas matemáticas associadas a cada uma das representações[1], as sucessões de vectores de características constituem a opção dominante nos sistemas *multi-stream* actuais:

$$\overset{n}{\boldsymbol{X}} = \left\{ \overset{n}{\boldsymbol{x}}_1,\ \overset{n}{\boldsymbol{x}}_2,\ ...,\ \overset{n}{\boldsymbol{x}}_{T_n} \right\}, \qquad \overset{n}{\boldsymbol{x}}_i \in \Re^{D_n}, \qquad n = 1,...,N \tag{3.2}$$

O sinal da fala é assim transformado em diversas trajectórias, cada qual definida num espaço de representação diferente. A resolução temporal dessas trajectórias, assim como a dimensão do espaço onde estão definidas, é também eventualmente diferente.

Em todo o caso, o aspecto que interessa agora realçar reside na exigência de que a informação veiculada por cada *stream* apresente, por si só, alguma capacidade discriminativa no processo de decisão. Significa isto que deve ser não nula a informação mútua média[2] entre os processos estocásticos associados à geração da frase e à de qualquer dos *streams*, ou seja:

$$\overline{\mathrm{M}}\left(W; \overset{n}{\boldsymbol{X}} \right) > 0, \qquad n = 1,...,N \tag{3.3}$$

[1] Tal como se verifica, por exemplo, em alguns sistemas que utilizam representações acústicas do sinal baseadas em sucessões de matrizes.

[2] Definida como $\overline{\mathrm{M}}(\boldsymbol{X};\boldsymbol{Y}) = \mathrm{E}\left\{ \log_2 \dfrac{\mathrm{P}(\boldsymbol{X},\boldsymbol{Y})}{\mathrm{P}(\boldsymbol{X})\mathrm{P}(\boldsymbol{Y})} \right\}$.

Dado que $\overline{\mathrm{M}}\left(W;\overset{n}{\boldsymbol{X}}\right)=\mathrm{H}(W)-\mathrm{H}\left(W|\overset{n}{\boldsymbol{X}}\right)$, idealmente a informação mútua[1] seria igual à medida da entropia[2] *apriori* associada a W, e portanto $\mathrm{H}\left(W|\overset{n}{\boldsymbol{X}}\right)=0$, ou seja, a informação extraída do sinal e representada em $\overset{n}{\boldsymbol{X}}$ anularia a incerteza acerca de W. Obviamente, se se verificasse esta situação ideal não faria qualquer sentido recorrer a uma abordagem *multi-stream*.

Embora a condição imposta pela inequação 3.3 seja necessária, não é por princípio suficiente para que um *stream* contribua efectivamente para melhorar o modelo acústico de um reconhecedor *multi-stream*. É também necessário que a informação que lhe está associada seja não redundante relativamente ao conjunto dos restantes *streams*. Assim, pode ser definida uma condição que engloba a anterior mas é mais restritiva:

$$\mathrm{H}\left(W|\overset{1}{\boldsymbol{X}},\overset{2}{\boldsymbol{X}},\ldots,\overset{N}{\boldsymbol{X}}\right)<\mathrm{H}\left(W|\overset{1}{\boldsymbol{X}},\overset{2}{\boldsymbol{X}},\ldots,\overset{n-1}{\boldsymbol{X}},\overset{n+1}{\boldsymbol{X}},\ldots,\overset{N}{\boldsymbol{X}}\right),\qquad n=1,\ldots,N \tag{3.4}$$

ou seja, a entropia associada a W sendo conhecidos todos os *streams* é inferior à entropia condicional quando todos os *streams* são conhecidos à excepção de $\overset{n}{\boldsymbol{X}}$.

É claro que a contribuição de um qualquer *stream* para melhorar o modelo acústico será tanto maior quanto mais afastadas forem as medidas de entropia nos dois membros da respectiva inequação 3.4. Uma vez mais, idealmente o membro da esquerda apresentaria um valor nulo, de tal modo que admitindo já estabelecidos todos os *streams* à excepção de $\overset{n}{\boldsymbol{X}}$, então a informação extraída do sinal através deste *stream* seria suficiente para anular por completo a incerteza acerca de W.

3.3.3 Características seleccionadas

Os dois *streams* utilizados neste sistema baseiam-se na mesma técnica de análise do sinal, registando-se apenas diferenças entre as respectivas resoluções espectrais. Os bancos de filtros utilizados no cálculo dos coeficientes cepstrais Mel, num caso, e lineares, no outro, apresentam um escalonamento diferente, de acordo com a representação gráfica apresentada na figura 3.3.

---

[1] Note-se que a maximização de $\mathrm{M}(W;\boldsymbol{X})$ tem sido, em alguns sistemas *multi-stream*, com o modelo linguístico fixo, utilizado como critério de treino do modelo de recombinação das verosimilhanças.

[2] Definida como $\mathrm{H}(\boldsymbol{X})=\mathrm{E}\left\{-\log_2\mathrm{P}(\boldsymbol{X})\right\}$.

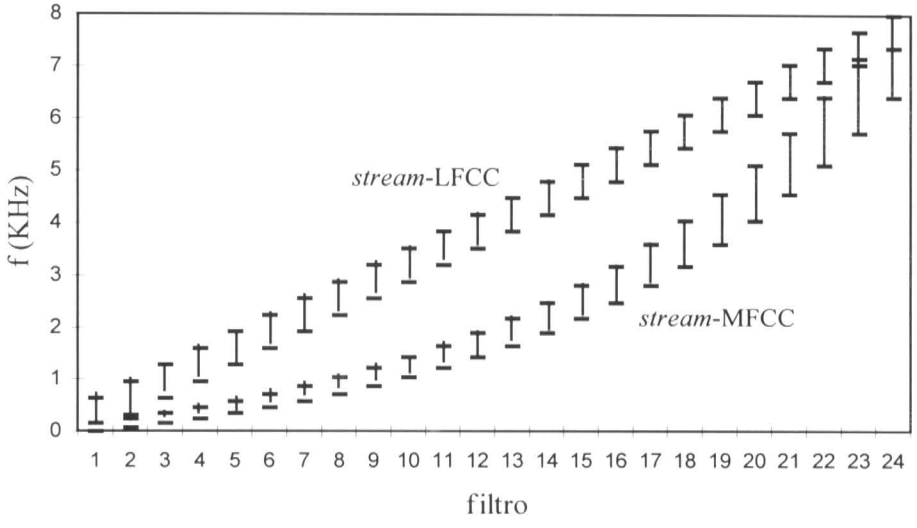

Fig. 3.3 - Escalonamento do banco de filtros correspondente aos coeficientes lineares cepstrais (*stream*-LFCC) e aos mel-cepstrais (*stream*-MFCC).

Como se observa, no caso do *stream*-LFCC o escalonamento dos filtros é linear na métrica real das frequências. No *stream*-MFCC os filtros encontram-se linearmente espaçados apenas na escala perceptiva Mel, emulando assim uma característica do sistema auditivo humano. Na escala real, estas características apresentam uma resolução espectral relativamente maior para as frequências mais baixas, o que se tem verificado permitir uma melhor representação da informação relevante para a tarefa do reconhecimento (§2.5.2.4).

Apesar destas diferenças na escala espectral, baseando-se os dois *streams* no mesmo método de análise seria previsível que pelo menos relativamente aos coeficientes cepstrais de menor ordem se verificasse, entre ambos, uma dependência estatística significativa. Assim, foi efectuado um pequeno estudo para tentar avaliar a dependência estatística linear entre os *streams*.

A figura 3.4 mostra a matriz correlação que se verifica entre as componentes do vector constituído pelos 12 coeficientes mel-cepstrais e também pelos 12 coeficientes lineares cepstrais, estimada a partir de aproximadamente 25000 sectores do sinal extraídos aleatoriamente do conjunto de treino.

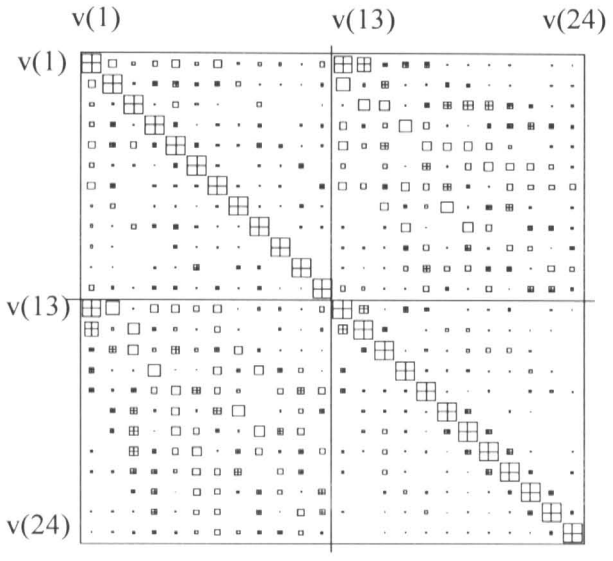

Fig. 3.4 - Matriz correlação do vector $\mathbf{v} = \left(MFCC_1, ..., MFCC_{12}, LFCC_1, ..., LFCC_{12} \right)$.

A matriz mostra que a correlação entre os MFCC, assim como entre os LFCC, é pequena, em resultado da aplicação da transformada discreta do coseno (§2.5.2.4) em ambas as análises. Além disso, e apesar destes resultados terem sido obtidos considerando indistintamente todas as classes de sons, é interessante verificar que a correlação entre os MFCC e os LFCC é em geral claramente maior. Confirma-se ainda a maior dependência linear entre os MFCC e os LFCC de menor ordem, sendo mesmo muito elevada entre os de primeira ordem.

Quanto às razões que determinaram a escolha destas características, e começando por justificar a utilização de apenas dois *streams*, isso deveu-se principalmente ao facto de esse número ser suficiente face aos objectivos gerais propostos e, além disso, facilitar um melhor aproveitamento dos recursos de computação disponíveis para este trabalho de investigação.

Um dos *streams* deste sistema seria inevitavelmente baseado na análise Mel-cepstral do sinal, dado o interesse em confrontar directamente o desempenho dos reconhecedores *multi-* e single-*stream*.

No entanto, em relação ao "segundo" *stream*, é natural que se questione se a escolha dos LFCC é a mais indicada.

Apesar do desempenho dos LFCC ser em geral substancialmente pior que o dos MFCC - o que se confirma na tarefa de reconhecimento aqui estabelecida, em que a taxa de palavras erradas obtida com o sistema *single-stream* é aproximadamente $45\%_{rel}$

mais elevada[1] -, isso não implica por si só a rejeição da opção pelos LFCC. Como é óbvio o *stream*-LFCC capta, numa medida útil, informação discriminante do sinal, pois caso contrário a taxa do erro seria muito elevada e apenas limitada pelas restrições impostas pelo modelo linguístico.

Uma questão especialmente importante está relacionada com a redundância da informação, necessária ao processo de reconhecimento, associada aos dois *streams*. Idealmente, a informação mútua média entre ambos os *streams* seria nula, o que se verificaria se eles fossem estatisticamente independentes. No entanto, tal como foi já salientado, as distribuições estatísticas correspondentes a estes dois *streams* apresentam até uma substancial dependência linear. Logo, é expectável que o aumento relativo de informação discriminante introduzida no descodificador acústico através do *stream*-LFCC não seja muito grande, o que corresponde à relativa proximidade das medidas de entropia $\mathrm{H}(W|\boldsymbol{X}_{MFCC})$ e $\mathrm{H}(W|\boldsymbol{X}_{MFCC},\boldsymbol{X}_{LFCC})$.

Apesar de se ter chegado a considerar a opção por outro *stream* de características, obtido através da análise baseada num *modelo de auditório* e disponibilizado pelo laboratório ELIS [194], que eventualmente proporcionaria uma melhor combinação com os MFCC, tomou-se esta opção sobretudo pelas razões que a seguir se apresentam.

Em primeiro lugar, existem resultados publicados atestando o melhor desempenho dos LFCC, relativamente aos MFCC, em determinadas classes de sons, em particular com um poder discriminativo maior nas consoantes oclusivas [193].

Considerando todas as classes de sons, apesar de se confirmar que os MFCC obtêm melhores resultados em sistemas independentes do falante, verifica-se também que o seu desempenho relativamente aos LFCC está relacionado com o valor típico da frequência de excitação fundamental do sinal.

Estes dois aspectos proporcionam desde logo um desafio muito interessante, sugerindo o desenvolvimento de um modelo de recombinação dos *streams* dependente da classe de sons e também capaz de se adaptar a determinada característica que pode ser extraída do sinal.

Por fim, mas não menos importante, ao considerar esta combinação de *streams* manteve-se presente a noção de que num trabalho de investigação desta natureza o objectivo principal não coincide necessariamente com a optimização de todas as partes de um determinado sistema, reside antes de mais na compreensão profunda dos fenómenos e na aquisição de mestria na utilização e no desenvolvimento de tecnologia.

[1] Para sistemas com aproximadamente igual número de parâmetros livres, os resultados obtidos com as características MFCC e com as LFCC são, respectivamente, 6,48% e 9,42%.

3.3.4 Extracção das características

3.3.4.1 Características baseadas nos coeficientes cepstrais

Neste módulo de análise são paralelamente extraídos dois *streams* de características do sinal acústico.

Um dos *streams*, baseado nos MFCC, é obtido através de um processo de análise do sinal idêntico ao utilizado no sistema *single-stream* desenvolvido neste trabalho (§2.5.2). O outro *stream*, baseado nos LFCC, é extraído através de um procedimento semelhante mas utilizando diferente resolução espectral no cálculo dos coeficientes cepstrais.

Assim, no sentido de optimizar a operacionalidade do reconhecedor, a parte inicial deste módulo de análise, correspondente ao pré-processamento do sinal e ao cálculo da transformada discreta de Fourier, é comum aos dois *streams*. De acordo com a estrutura modular apresentada na figura 2.2, são realizadas as operações de pré-ênfase do sinal (§2.5.2.2), formação de sectores, aplicação da janela de Hamming (§2.5.2.3) e, para cada sector, cálculo da FFT.

A partir do espectro de potência do sinal, os *streams* são processados separadamente. Em relação ao *stream*-MFCC, é continuada a execução do processo de análise tal como se verifica no reconhecedor *single-stream* (§2.5.2).

No *stream*-LFCC o espectro de potência é aplicado a um banco com 24 filtros cuja forma é triangular e estão linearmente escalonados, tal como mostra a figura 3.2. Os resultados disponíveis à saída do banco de filtros são então submetidos a um processo de compressão logarítmica e à transformada discreta do coseno, tal como se verifica nos MFCC (§2.5.2.4), resultando 12 LFCC. É também incluído um termo com a energia de cada sector, tal como no sistema base *single-stream*, calculando o logaritmo da soma das energias à saída do banco de filtros (§2.5.2.5). Note-se que apesar de este termo da energia fazer já parte do outro *stream*, a sua inclusão também neste facilita a comparação, no reconhecedor *single-stream*, do desempenho entre estas características e as baseadas nos MFCC. Por outro lado, embora sendo certo que a sua repetição não seria aceitável num reconhecedor definitivo, isso não apresenta inconvenientes significativos durante a fase de desenvolvimento de um protótipo para investigação, como é o caso. Finalmente, são calculadas e apensas ao vector de características as respectivas primeiras e segundas derivadas pesadas, de acordo com um método de regressão linear efectuado sobre as características nos sectores vizinhos (§2.5.2.6).

Refira-se que os coeficientes cepstrais lineares poderiam ser calculados, em alternativa ao procedimento descrito e sem recorrer ao banco de filtros, directamente a partir do espectro do sinal:

$$LFCC(i) = \frac{1}{K}\sum_{m=1}^{K/2-1} \log(S_m)\cos\left(\frac{2\pi im}{K}\right), \quad i = 1,2,\ldots,N \tag{3.5}$$

em que S_m é a $m^{\text{ésima}}$ componente da magnitude da frequência e K é a dimensão da FFT.

Assim, à saída deste módulo estão disponíveis duas representações do sinal acústico que se apresenta à entrada, cada uma delas consistindo numa sucessão de vectores de características definidos num espaço 39-dimensional e regularmente intervalados de 10ms:

$$s(k) \rightarrow \overset{1}{\boldsymbol{X}}_1^T, \ \overset{2}{\boldsymbol{X}}_1^T \tag{3.6}$$

$$\overset{n}{\boldsymbol{X}}_1^T = \left\{\overset{n}{\boldsymbol{x}}_1, \overset{n}{\boldsymbol{x}}_2, \ldots, \overset{n}{\boldsymbol{x}}_T\right\}, \quad \overset{n}{\boldsymbol{x}}_t \in \Re^{39}, \quad n = 1, 2$$

Cada vector de $\overset{1}{\boldsymbol{X}}$, designado também *stream*-MFCC, contém 12 coeficientes mel-cepstrais, um termo da energia e as respectivas primeiras e segundas "derivadas". Em relação a $\overset{2}{\boldsymbol{X}}$, designado *stream*-LFCC, cada vector contém 12 coeficientes lineares cepstrais, o termo da energia e as respectivas primeiras e segundas "derivadas".

3.3.4.2 Frequência fundamental

Este módulo de análise extrai também do sinal uma estimativa do valor médio da frequência fundamental correspondente aos sons vocálicos, informação utilizada no modelo de recombinação (§3.4.5).

Deve ser referido que o estimador da frequência fundamental, F_0, em cada sector do sinal correspondente a um som vocálico não chegou a ser construído e integrado neste sistema. Foram utilizadas as estimativas de F_0 disponibilizadas pelo Laboratório ELIS e calculadas através de um modelo de auditório aí desenvolvido [194]. Assim, e tendo obviamente presente que sendo esta solução precária é apenas aceitável num protótipo para investigação, foi simplesmente implementado um operador numérico para calcular a média aritmética de valores sucessivos de F_0 dentro de uma determinada janela móvel de operação. A definição da largura dessa janela exige alguma atenção, tal como a seguir se expõe.

De acordo com a formulação apresentada em [195], a evolução da frequência fundamental, geralmente designada curva F_0, pode ser interpretada como o resultado da resposta do mecanismo de vibração das cordas vocais à informação de natureza linguística, para-linguística e não-linguística associada ao processo de comunicação.

Os factores identificados como mais relevantes são, ao nível linguístico, a acentuação lexical e as estruturas sintácticas e de conversação. Por exemplo, tipicamente a curva F_0 apresenta picos durante as sílabas tónicas. É também conhecido que existe uma tendência estatística para o decaimento progressivo da curva nas frases ou segmentos declarativos e, pelo contrário, um aumento nas interrogativas.

Ao nível para-linguístico, o valor de F_0 pode também ser afectado quando, por exemplo, o falante pretende expressar uma determinada intenção ou atitude.

Também factores de natureza não linguística, como por exemplo, a idade, o sexo e os estados físico ou psíquico do falante afectam a frequência fundamental do sinal da fala. Um resultado bem conhecido é a forte correlação entre o valor médio de F_0 e o sexo do falante, para uma determinada faixa etária. A figura 3.5 mostra um histograma da distribuição da frequência fundamental média correspondente aos 82 falantes, 28 mulheres e 54 homens, das bases de dados *Resource Managment* e *February'89* usadas no desenvolvimento deste sistema. Refira-se que esse valor médio foi obtido, para cada falante, considerando as estimativas de F_0 disponíveis em todos os segmentos vocálicos que ocorrem nas base de dados e correspondentes a esse falante.

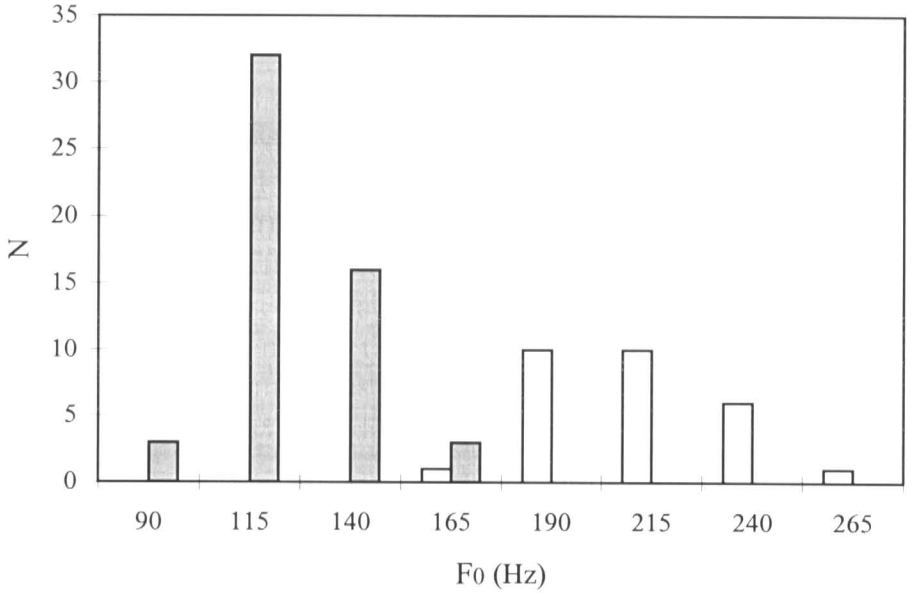

Fig. 3.5 - Histograma da distribuição de F_0 médio para um conjunto de falantes femininos (colunas a branco) e masculinos (colunas sombreadas).

Como se pode imediatamente constatar, seria possível desenhar um classificador, para as classes {mulher, homem}, com elevada precisão colocando a fronteira de decisão em aproximadamente 165 Hz. Este exemplo simples serve também para introduzir alguns aspectos importantes sobre a utilização da média de F_0 no reconhecedor *multi-stream*.

Repartindo todos os falantes do conjunto de treino em dois grupos, um formado só por mulheres e o outro só por homens, e confrontando os resultados obtidos com o reconhecedor *multi-stream*, verifica-se que o desempenho do *stream*-MFCC relativamente ao *stream*-LFCC é substancialmente melhor no grupo dos homens (§3.4.5). Refira-se que a mesma tendência, embora bastante mais acentuada, se verifica relativamente ao conjunto de teste. A verificação da existência de uma dependência estatística significativa entre o valor médio de F_0 e o desempenho relativo dos dois *streams* utilizados neste reconhecedor constitui precisamente a motivação para a utilização deste parâmetro na adaptação do modelo de recombinação. Interessa aqui realçar que o facto importante é a verificação da dependência estatística entre essas duas variáveis, sendo dispensável a existência de qualquer relação causal entre elas.

O problema que deve ser agora tratado prende-se então com a estimativa do valor médio de F_0 a partir dos valores conhecidos nos segmentos correspondentes aos sons vocálicos.

Mesmo admitindo que os factores não-linguísticos e para-linguísticos são pouco significativos no contexto da utilização de F_0 neste reconhecedor *multi-stream*, o que será até aceitável em certa medida dado que a fala nesta base de dados não é espontânea, continua a registar-se uma significativa variabilidade do valor de F_0 devido aos fenómenos de prosódia e micro-prosódia [195] ao nível da informação linguística. Fundamentalmente, reside aí a necessidade de filtrar a sucessão $\{F_0\}$ através de um operador móvel de média aritmética, utilizando um filtro digital não recursivo.

Recorrendo novamente ao exemplo do classificador de género, não é necessário efectuar qualquer experiência para poder afirmar que o seu desempenho seria pior se, ao invés de considerar a estimativa da média de F_0 tal como foi feito, fosse simplesmente considerado o valor de F_0 num segmento vocálico aleatório.

No entanto, neste reconhecedor *multi-stream* a melhor solução poderá ser utilizar apenas uma parte do material acústico localmente disponível, adaptando o modelo de recombinação dos *streams* com base na estimativa do valor médio de F_0 sobre uma janela com duração relativamente pequena. Seria eventualmente necessário contornar a dificuldade devida à existência das estimativas de F_0 apenas nos sons vocálicos. Por outro lado, esta solução apresenta desde logo uma vantagem operacional importante, pois evita o atraso no processo de reconhecimento, o qual poderia inviabilizar uma aplicação em tempo real.

Dada a complexidade em optimizar a largura da janela, no plano teórico, poderia ser seguida uma abordagem empírica baseada na avaliação dos resultados experimentais obtidos com valores diferentes deste parâmetro. No entanto, durante este trabalho de investigação foi apenas considerada uma janela com a duração fixa de 5 segundos.

Assim, à saída deste módulo de análise está também disponível um *stream* escalar correspondente à estimativa do valor médio de F_0 obtida com um filtro *averager* de ordem 500:

$$s(k) \rightarrow \overline{F}_0 \tag{3.7}$$

3.4 Classificação do sinal

3.4.1 Introdução

O objectivo do módulo de classificação de um reconhecedor *multi-stream* é encontrar a melhor interpretação, de acordo com um determinado critério estabelecido com base no conhecimento representado nos modelos linguístico e acústico, para as múltiplas representações do sinal da fala disponíveis à saída do módulo de análise.

Nos reconhecedores que são aqui tratados, a classificação assenta numa estratégia de processamento integrado e paralelo desses *streams* de informação, com o resultado determinado por restrições que se exercem aos diferentes níveis e, pelo menos em princípio, de maneira global e simultânea. Esta abordagem é substancialmente diferente da que poderia ser seguida utilizando também múltiplas fontes de informação ao nível acústico mas aplicando determinados critérios de decisão sobre várias listas de hipóteses de reconhecimento geradas, em sequência ou em paralelo, por algoritmos do tipo *N-best*[1].

Existem, naturalmente, diversas questões essenciais que se colocam no desenvolvimento do módulo de classificação. A questão primordial reside na definição do critério em que se fundamenta o processo de classificação (§3.4.2).

Uma questão subsequente consiste no estabelecimento de um formalismo estatístico (§3.4.3) que permita a aplicação desse critério de acordo também com a estratégia de reconhecimento estabelecida.

Partindo de um dado sistema *single-stream*, desenvolvido para idêntica aplicação, é apenas necessário proceder a alterações ao nível do modelo acústico. Ao nível linguístico estes reconhecedores são basicamente idênticos, com o processo de pesquisa no espaço de soluções condicionado pela gramática e pelo léxico definidos para essa tarefa e integrados na estrutura de conhecimento (§2.7; §2.8.2).

No desenvolvimento concreto do classificador emergem várias questões, desde a introdução de modelos acústicos que permitam uma maior eficácia na descodificação

[1] Serve esta nota também para referir que foi desenvolvido um descodificador *multi-stream N-best* com o objectivo de investigar a utilização de diferentes estratégias de reconhecimento baseadas em listas de hipóteses obtidas com *streams* diferentes.

(§3.4.4), a definição do processo de recombinação dos *streams* (§3.4.5), o treino dos parâmetros acústicos (§3.4.6) e diversos aspectos relacionados com a descodificação (§3.4.7).

3.4.2 Critério de classificação

Recorrendo a uma formulação análoga à que foi apresentada para o reconhecedor *single-stream* (§1.3.6.3; §2.2; §2.6.1), num classificador *multi-stream* a transformação das múltiplas observações do sinal acústico na hipótese de reconhecimento mais provável é feita com base numa função discriminante que define as fronteiras de classificação e que pode ser factorizada nos modelos linguístico, L, e acústico, M:

$$\overset{1}{\boldsymbol{X}},\ \overset{2}{\boldsymbol{X}},\ ...,\ \overset{N}{\boldsymbol{X}} \xrightarrow[L,\,M]{} W \tag{3.8}$$

De acordo com o critério de classificação Bayesiano (§1.3.6.3), o resultado desta transformação é a classe que apresenta a máxima probabilidade *a posteriori*:

$$W_c = \arg\max_{W} \left\{ P\left(W | \overset{1}{\boldsymbol{X}},\ \overset{2}{\boldsymbol{X}},\ ...,\ \overset{N}{\boldsymbol{X}} \right) \right\} \tag{3.9}$$

Atendendo à dificuldade que geralmente se verifica na modelação directa da probabilidade *a posteriori* (§1.3.6.3), para além do eventual interesse que possa existir em considerar isoladamente diferentes níveis de informação, o classificador pode ser construído com base em dois modelos estatísticos.

O modelo linguístico calcula a probabilidade *a priori* de ocorrência de uma dada hipótese de reconhecimento, isto é, $P(W)$.

O modelo acústico estima a verosimilhança da observação conjunta dos *streams* condicionada ao conhecimento da frase, ou seja:

$$p\left(\overset{1}{\boldsymbol{X}},\ \overset{2}{\boldsymbol{X}},\ ...,\ \overset{N}{\boldsymbol{X}} | W \right) \tag{3.10}$$

Assim, a partir de (3.9) e utilizando a regra de Bayes, fica definido o seguinte critério de classificação para o classificador *multi-stream*:

$$W_c = \arg\max_{W} \left\{ P(W) p\left(\overset{1}{\boldsymbol{X}}, \overset{2}{\boldsymbol{X}}, ..., \overset{N}{\boldsymbol{X}} | W \right) \right\} \tag{3.11}$$

3.4.3 Formalização do modelo acústico *multi-stream*

O formalismo estatístico utilizado nestes sistemas para modelar a verosimilhança conjunta dos *streams* condicionada ao conhecimento da frase baseia-se na recombinação das verosimilhanças calculadas de maneira independente pelos modelos acústicos correspondentes a cada *stream* [190,192]. Esse processo de recombinação ocorre no final de cada uma das unidades acústicas elementares (§2.6.2.1) que compõem a hipótese de reconhecimento.

Considere-se então que está definido um conjunto de unidades simbólicas:

$$\Gamma_u = \{u^1, u^2, ..., u^L\} \tag{3.12}$$

e que para cada uma delas existe um modelo acústico que calcula a verosimilhaça da observação correspondente a cada *stream*:

$$p\left(\overset{n}{\boldsymbol{X}} | u^l \right), \qquad l=1,2,...,L \qquad n=1,2,...,N \tag{3.13}$$

Seja dada uma hipótese válida de reconhecimento, composta por uma sucessão de unidades simbólicas:

$$W = \{u_1, u_2, ..., u_M\}, \quad \forall u_m \in \Gamma_u \tag{3.14}$$

e correspondente a um sinal acústico que se encontra representado por N *streams* de características.

Segundo este formalismo estatístico, para todos os *streams* deve verificar-se uma correspondência directa, estabelecida de acordo com uma determinada segmentação do sinal, entre a sucessão de unidades simbólicas e a sucessão de segmentos de vectores de características:

$$\overset{n}{\boldsymbol{X}} = \left\{ \overset{n}{\boldsymbol{X}}_1, \overset{n}{\boldsymbol{X}}_2, ..., \overset{n}{\boldsymbol{X}}_M \right\}, \qquad n=1,2,...,N \tag{3.15}$$

Nesta altura, relativamente a um aspecto importante relacionado com esta segmentação, interessa apenas referir que ela é gerada automaticamente durante o reconhecimento segundo um critério de maximização da verosimilhança global determinada pelo modelo acústico.

É importante sublinhar que essa segmentação é necessariamente idêntica para todos os *streams*, isto é, as transições entre os modelos correspondentes às sucessivas unidades acústicas ocorrem de maneira síncrona. Esta exigência decorre da necessidade de recombinar os *streams* precisamente no final de cada segmento.

Significa isto que no caso particular dos múltiplos *streams* de características apresentarem igual resolução temporal[1], então a segmentação poderia ser representada para todos os *streams* simplesmente utilizando uma única sucessão de índices:

$$\varsigma = \{\varsigma_0, \varsigma_1, \ldots, \varsigma_M\}, \qquad \varsigma_0 = 0, \quad \varsigma_M = T \tag{3.16}$$

correspondentes ao instante final de cada segmento da sucessão de vectores de características:

$$\overset{n}{\boldsymbol{X}}_m = \left\{ \overset{n}{\boldsymbol{x}}_{\varsigma_{m-1}+1}, \overset{n}{\boldsymbol{x}}_{\varsigma_{m-1}+2}, \ldots, \overset{n}{\boldsymbol{x}}_{\varsigma_m} \right\}, \qquad m=1,2,\ldots,M \qquad n=1,2,\ldots,N \tag{3.17}$$

Em todo o caso, para cada segmento e para cada *stream* é possível calcular a respectiva verosimilhança de observação:

$$p\left(\overset{n}{\boldsymbol{X}}_m | u_m \right), \qquad m=1,2,\ldots,M \qquad n=1,2,\ldots,N \tag{3.18}$$

Estes valores podem então ser recombinados, no final de cada segmento, através de uma função f_r que estima a verosimilhança conjunta dos múltiplos *streams* nesse segmento:

$$p\left(\overset{1}{\boldsymbol{X}}_m, \overset{2}{\boldsymbol{X}}_m, \ldots, \overset{N}{\boldsymbol{X}}_m | u_m \right) = f_r\left(p\left(\overset{n}{\boldsymbol{X}}_m | u_m \right), \quad n=1,2,\ldots,N \right), \qquad m=1,2,\ldots,M \tag{3.19}$$

Finalmente pode ser calculado o resultado pretendido (3.10) através do produto das verosimilhanças estimadas para cada unidade considerando a segmentação do sinal, ao nível dessas unidades, que permite a maximização da verosimilhança global, isto é:

[1] É o que se verifica neste sistema em particular, em que o *frame-rate* dos dois *streams* utilizados é igual.

$$p\left(\overset{1}{\boldsymbol{X}},\overset{2}{\boldsymbol{X}},\ldots,\overset{N}{\boldsymbol{X}}|W\right) = \max_{\varsigma}\left\{\prod_{m=1}^{M} p\left(\overset{1}{\boldsymbol{X}}_m,\overset{2}{\boldsymbol{X}}_m,\ldots,\overset{N}{\boldsymbol{X}}_m|u_m\right)\right\} \tag{3.20}$$

3.4.4 Modelos acústicos *compostos*

Os chamados modelos acústicos *compostos*[1] surgem [187,190] como uma solução muito eficaz para a integração do formalismo estatístico apresentado na secção anterior num reconhecedor automático da fala. Em breves palavras, pode afirmar-se que isso se deve essencialmente ao facto desses modelos permitirem a utilização de um algoritmo simples e poderoso como é o descodificador Viterbi (§2.8.3).

Na exposição que se segue são impostas algumas características particulares ao modelo *multi-stream* genérico apresentado (3.12 a 3.20), rematando na configuração correspondente ao reconhecedor construído neste trabalho.

Uma potencial vantagem do modelo *multi-stream*, explorada nos chamados sistemas *multi-model*, reside na facilidade em combinar técnicas de modelação diferentes para os diversos *streams*. Essas abordagens são motivadas pelas deficiências inerentes a qualquer tecnologia de modelação acústica, geralmente resultantes de assunções demasiado simplificadoras. Logo, a combinação de diferentes tipos de modelos, cada qual apresentando uma especial aptidão para representar determinados aspectos da informação disponível, deveria em princípio conduzir a um modelo acústico melhor. Têm sido publicados alguns resultados promissores, por exemplo combinando os HMM *standard* com os modelos de segmentos [161,185]. Contudo, esta questão excede os objectivos deste trabalho, pelo que todos os modelos acústicos elementares (3.13) do reconhecedor desenvolvido se baseiam unicamente na tecnologia dos SCHMM.

Admitindo então esta solução, haveria ainda a possibilidade de utilizar os SCHMM com diferentes topologias em cada *stream*. Por exemplo, se um determinado *stream* de características apresentasse uma estrutura temporal mais fina que os restantes deveria então corresponder-lhe um modelo com maior número de estados. Precisamente por se supor que essa situação não se verifica em grau significativo relativamente aos dois *streams* utilizados neste reconhecedor, em qualquer dos símbolos ambos os modelos acústicos elementares (3.13) baseiam-se nos SCHMM com idêntica topologia.

Note-se que esta última questão está estreitamente ligada a uma outra característica importante do modelo *multi-stream*, consistindo na possibilidade de combinar *streams* com diferentes resoluções temporais. Por exemplo, relativamente a alguns sistemas *multi-scale* de que se fez uma breve referência (§3.3.1), são obtidos

[1] Em Inglês, *composite models*, ou *multi-dimensional models*.

resultados assinaláveis combinando um *stream* com uma resolução temporal típica de aproximadamente 10 ms, modelado ao nível fonético, com um outro, modelado ao nível silábico, cujos sectores têm a duração 200 ms [45]. No entanto, pelas razões já apontadas, no reconhecedor desenvolvido neste trabalho ambos os *streams* apresentam idêntica resolução.

É então possível estabelecer uma representação gráfica [190] correspondente ao modelo *multi-stream* em que se baseia o reconhecedor desenvolvido neste trabalho. A figura 3.6 mostra um segmento dos dois *streams* de vectores de características, abrangendo a realização acústica de dois fonemas consecutivos, e também a parte correspondente da estrutura do modelo. Como se pode ver, as unidades fonéticas são modeladas, em ambos os *streams*, através dos HMM com 3 estados e uma topologia idêntica à utilizada no sistema *single-stream* (§2.6.3.2; figura 2.14). Refira-se que neste reconhecedor o modelo acústico correspondente aos períodos de silêncio se baseia, tal como no sistema *single-stream* (§2.6.3.2; figura 2.15), num único estado. A figura mostra ainda os nós de recombinação, notados por f_r, que no final de cada segmento fonético, estimam a verosimilhança conjunta dos dois *streams* através da equação 3.19.

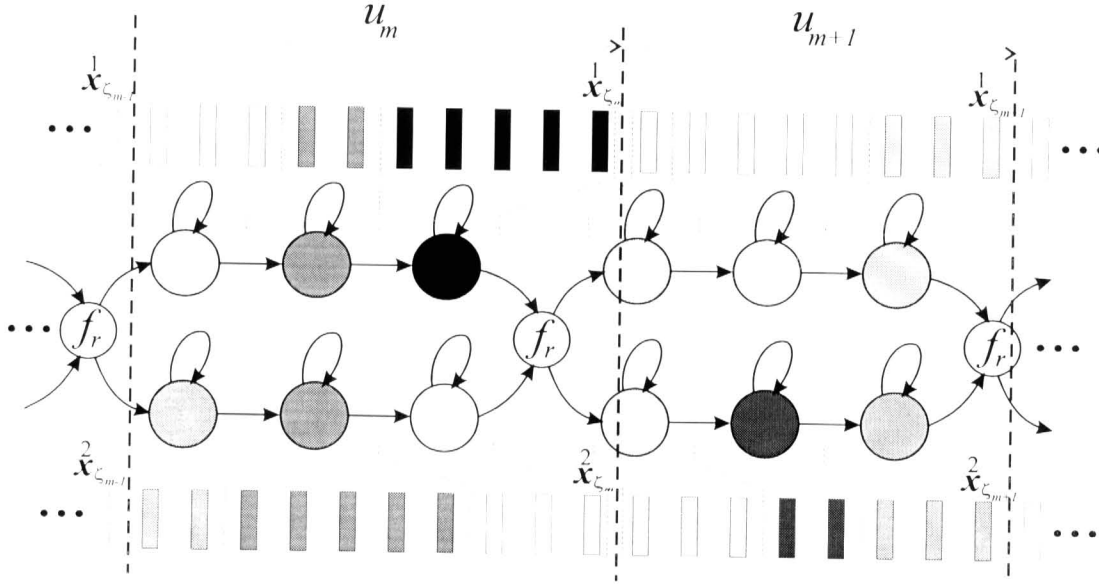

Fig. 3.6 - Representação de parte do modelo acústico *multi-stream*.

A figura mostra claramente que a segmentação dos *streams* ao nível das unidades fonéticas é forçosamente coincidente, de acordo com a formalização deste modelo (§3.4.3). É bastante complexo encontrar a segmentação óptima, o que de acordo com a equação 3.20 é um problema essencial no processo de reconhecimento. Basta notar que as segmentações naturais para os dois *streams* não são necessariamente coincidentes, por exemplo, na situação ilustrada pela figura 3.6, o vector $\overset{2}{x}_{\zeta_m}$ encontra-se atribuído ao último estado do primeiro fonema, e não ao primeiro estado do segundo fonema, o

qual modelaria melhor o seu "tom", devido ao forte "contraste local" que ocorre no *stream-1*. Além disso, admitindo uma descodificação assente no conceito de melhor caminho (§2.8.3) também ao nível da estrutura de estados no interior de cada modelo acústico, tal como a figura ilustra, os *streams* não são necessariamente síncronos.

Apesar da dificuldade deste problema da descodificação *multi-stream*, existem diversas abordagens baseadas em técnicas de Programação Dinâmica que têm permitido resolvê-lo com sucesso. Uma parte significativa dos resultados publicados refere a utilização de uma versão síncrona do algoritmo *two-level DP*[1] [190], o qual apresenta duas etapas operacionais: na primeira são calculadas as verosimilhanças de observação de segmentos do sinal pelos modelos acústicos correspondentes aos símbolos de fala; no segundo passo é pesquisada a sequência desses símbolos que apresenta uma maior verosimilhança de observação do sinal completo. A solução utilizada neste sistema assenta nos chamados modelos acústicos *compostos*[2], desenvolvidos como uma técnica de decomposição do sinal aplicada aos HMMs [187,188].

Recorrendo à situação ilustrada na figura 3.6, mais concretamente em relação à unidade elementar u_m, a figura 3.7 mostra como essa técnica permite converter o modelo acústico *multi-stream* no modelo *composto* correspondente.

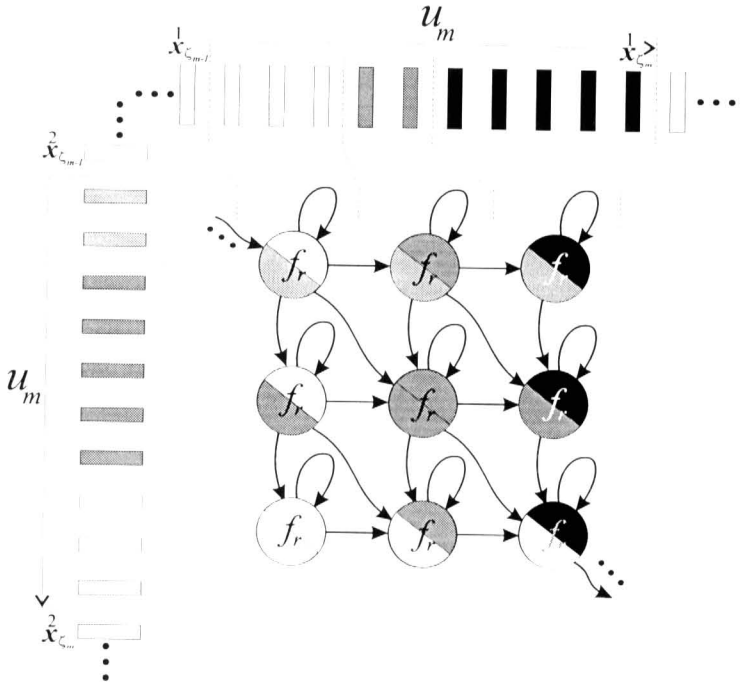

Fig. 3.7 - Modelo acústico *composto* correspondente à unidade acústica u_m (Fig. 3.6).

[1] Do Inglês, *dynamic programming*.
[2] Com base na pesquisa de informação realizada, esta solução é uma das mais correntes e aparentemente será até mesmo a mais utilizada.

Comparando os modelos verifica-se que os dois HMM paralelos originais, um para cada *stream*, são convertidos num único HMM. É claro que a topologia do novo modelo é mais complexa, de maneira a permitir todas as combinações de ocupação de estados possíveis, pois tal como foi já referido os *streams* não são necessariamente síncronos ao nível dos estados. Com efeito, as transições entre os segmentos mais estacionários dos vários *streams* de representação não ocorrem necessariamente ao mesmo tempo, o que torna as assunções de sincronismo bastante frágeis. Uma desvantagem destes modelos consiste no crescimento exponencial do número de estados com o aumento do número de *streams*. No entanto, neste caso concreto, em que apenas existem dois *streams*, esse aspecto não constitui um obstáculo significativo. Note-se que podem ser introduzidas algumas alterações no nível de sincronismo permitido entre os *streams*, eventualmente melhorando o desempenho do sistema, apenas eliminando, ou acrescentando, alguns estados do modelo *composto*. Deve ser salientado que a restrição de forçar o sincronismo entre os *streams* na transição entre as unidades elementares, fundamental na formulação inicial do modelo *multi-stream* [190], é neste trabalho relaxada, obtendo-se resultados positivos.

Considere-se então uma cadeia discreta de Markov (§2.6.3.2) bi-dimensional[1], cujo processo estocástico se representa pela variável aleatória $\left(\overset{1}{q}_t,\overset{2}{q}_t\right)$. Esta variável está definida para valores de t discretos e identifica um dos SxS estados do modelo *composto*:

$$\left(\overset{1}{q}_t,\overset{2}{q}_t\right) \in \Gamma_q, \qquad t = 1,2,\ldots,T \tag{3.21}$$

$$\Gamma_q = \{(1,1),(1,2),\ldots,(1,S),\ldots\ldots,(S,1),(S,2),\ldots,(S,S)\}$$

De maneira análoga à definição apresentada nos HMM *standard* (§2.6.3.2), ficam assim definidas as probabilidades de transição entre os estados:

$$a_{\left(\overset{1}{i},\overset{2}{i}\right),\left(\overset{1}{j},\overset{2}{j}\right)} = P\left(\overset{1}{q}_t = \overset{1}{j},\overset{2}{q}_t = \overset{2}{j} \middle| \overset{1}{q}_{t-1} = \overset{1}{i},\overset{2}{q}_{t-1} = \overset{2}{i}\right) \tag{3.22}$$

Neste sistema, o processo estocástico de transição entre os estados deste modelo bi-dimensional considera independentes os processos estocásticos de transição segundo qualquer dimensão, isto é:

$$P\left(\overset{1}{q}_t = \overset{1}{j},\overset{2}{q}_t = \overset{2}{j} \middle| \overset{1}{q}_{t-1} = \overset{1}{i},\overset{2}{q}_{t-1} = \overset{2}{i}\right) = P\left(\overset{1}{q}_t = \overset{1}{j} \middle| \overset{1}{q}_{t-1} = \overset{1}{i}\right) P\left(\overset{2}{q}_t = \overset{2}{j} \middle| \overset{2}{q}_{t-1} = \overset{2}{i}\right) \tag{3.23}$$

[1] Na sua forma mais geral a cadeia seria N-dimensional, correspondendo a *N-streams*; para simplificar a exposição e também porque o sistema desnvolvido se baseia em *2 streams*, considera-se N igual a 2.

Além disso, as probabilidades de transição entre os estados deste modelo *composto* podem ser estimadas com base nas probabilidades de transição conhecidas para o modelo acústico elementar correspondente a cada *stream*:

$$a_{\left(\begin{smallmatrix}1&2\\i,i\end{smallmatrix}\right),\left(\begin{smallmatrix}1&2\\j,j\end{smallmatrix}\right)} = \overset{1}{a}_{i,j}\,\overset{2}{a}_{i,j} \tag{3.24}$$

Note-se que estas probabilidades de transição continuam a ser inteiras, isto é, verificam uma condição equivalente à expressa na equação 2.27 para os modelos *single-stream*. Refira-se, no entanto, que apesar de alguns resultados publicados [186] atestarem uma maior influência, relativamente aos sistemas *single-stream*, das probabilidades de transição no desempenho de reconhecedores *multi-stream*, durante este trabalho verificou-se a fraca relevância destes parâmetros.

Tal como a figura 3.7 sugere, a conversão dos dois HMM num único HMM obriga a que cada novo estado modele a verosimilhança da observação conjunta dos dois *streams*. Assim, de maneira análoga à que se verifica nos HMM *standard* (§2.6.3.2), existe um processo estocástico associado a cada estado do HMM e definido através de uma função densidade de probabilidade:

$$b_{\left(\begin{smallmatrix}1&2\\j,j\end{smallmatrix}\right)}\left(\overset{1}{\boldsymbol{x}}_t,\overset{2}{\boldsymbol{x}}_t\right) = p\left(\overset{1}{\boldsymbol{x}}_t,\overset{2}{\boldsymbol{x}}_t \middle| \overset{1}{q}_t = \overset{1}{j},\overset{2}{q}_t = \overset{2}{j}\right) \tag{3.25}$$

O nó de recombinação, f_r, existente no modelo *multi-stream* original é eliminado, uma vez que agora essa operação passa a efectuar-se em cada estado. No entanto, tal como no formalismo *multi-stream* através da equação 3.19, continua a estimar-se a verosimilhança conjunta dos múltiplos *streams* a partir das verosimilhanças correspondentes a cada um deles:

$$p\left(\overset{1}{\boldsymbol{x}}_t,\overset{2}{\boldsymbol{x}}_t \middle| \overset{1}{q}_t = \overset{1}{j},\overset{2}{q}_t = \overset{2}{j}\right) = f_r\left(p\left(\overset{1}{\boldsymbol{x}}_t \middle| \overset{1}{q}_t = \overset{1}{j}\right), p\left(\overset{2}{\boldsymbol{x}}_t \middle| \overset{2}{q}_t = \overset{2}{j}\right)\right) \tag{3.26}$$

Neste sistema, o modelo de recombinação definido pela função f_r é tal que:

$$p\left(\overset{1}{\boldsymbol{x}}_t,\overset{2}{\boldsymbol{x}}_t \middle| \overset{1}{q}_t = \overset{1}{j},\overset{2}{q}_t = \overset{2}{j}\right) = p\left(\overset{1}{\boldsymbol{x}}_t \middle| \overset{1}{q}_t = \overset{1}{j}\right)^{\omega} p\left(\overset{2}{\boldsymbol{x}}_t \middle| \overset{2}{q}_t = \overset{2}{j}\right)^{1-\omega} \quad \text{com} \quad 0 \le \omega \le 1 \tag{3.27}$$

Com base no conhecimento da verosimilhança associada a cada *stream*, tal como apresentada na equação 2.31, a verosimilhança da observação conjunta dos dois *streams* é:

$$b_{\left(\overset{1}{j},\overset{2}{j}\right)}\left(\overset{1}{\boldsymbol{x}}_t,\overset{2}{\boldsymbol{x}}_t\right)=\left(\overset{1}{b}_j\left(\overset{1}{\boldsymbol{x}}_t\right)\right)^{\omega}\left(\overset{2}{b}_j\left(\overset{2}{\boldsymbol{x}}_t\right)\right)^{1-\omega} \tag{3.28}$$

É oportuno sublinhar aqui alguns aspectos importantes relacionados com o facto deste resultado se basear em funções de densidade probabilística definidas para modelar localmente as distribuições correspondentes a cada *stream*, considerando que essas mesmas distribuições são estatisticamente independentes.

Desde logo, note-se que esta assunção de independência é bastante abusiva relativamente aos *streams* utilizados neste reconhecedor, dado que se verifica mesmo uma significativa correlação entre as respectivas distribuições (§3.3.3).

É certo que o processamento independente dos *streams* nos modelos correspondentes às unidades elementares é uma característica do formalismo *multi-stream*(§3.4.3). No entanto, a motivação essencial reside no facto de essa ser uma boa solução operacional para acomodar as múltiplas representações do sinal. Por princípio, o modelo acústico seria melhor se fosse capaz de modelar as interdependências verificadas nas representações do sinal e relevantes no processo de classificação.

Relativamente aos modelos *compostos*, o ideal seria definir em cada estado uma função densidade de probabilidade que modelasse localmente a real distribuição conjunta dos *streams*, sendo portanto capaz de representar também as dependências estatísticas entre as componentes de ambos os vectores. Por outro lado, mesmo admitindo que a solução utilizada não é óptima segundo um critério baseado apenas na precisão dos modelos, a evidência experimental mostra que permite um bom compromisso entre o grau de complexidade do modelo acústico, assim como do seu desenvolvimento, e o desempenho do reconhecedor.

3.4.5 Formalização do modelo de recombinação

O objectivo mais imediato estabelecido para o desenvolvimento do modelo de recombinação foi, naturalmente, a procura de uma solução eficaz no melhoramento do desempenho deste reconhecedor perante a tarefa que serviu de suporte experimental a este trabalho. Importa realçar que os *streams* de características escolhidos antecipavam desde logo que este fosse um objectivo particularmente difícil (§3.3.3). Pretendeu-se também desenvolver um modelo simples, apresentando apenas o grau de complexidade necessário para incorporar o conhecimento fundamental à satisfação do objectivo anterior. Como facilmente se intui, frequentemente as soluções mais simples permitem também com maior consistência e abrangência a extrapolação de resultados ou conclusões.

Retomando a equação 3.28 e aplicando a transformação logarítmica obtém-se:

$$log\left(b_{\left(\begin{smallmatrix}1\\j,j\end{smallmatrix}\right)}\left(\overset{1}{\boldsymbol{x}}_t,\overset{2}{\boldsymbol{x}}_t\right)\right)=\omega\, log\left(\overset{1}{b}_j\left(\overset{1}{\boldsymbol{x}}_t\right)\right)+\left(1-\omega\right) log\left(\overset{2}{b}_j\left(\overset{2}{\boldsymbol{x}}_t\right)\right) \tag{3.29}$$

ou seja, em cada estado, o logaritmo da verosimilhança da observação conjunta dos vectores de características dos dois *streams* é uma função linear das estimativas correspondentes a cada *stream*.

Refira-se que em diversos sistemas são utilizados modelos de recombinação mais complexos, frequentemente recorrendo à capacidade dos MLP estabelecerem qualquer transformação, correspondente à função f_r na equação 3.26, não linear da entrada [191,192]. Tem sido apontada a esses modelos também a vantagem de poderem ser treinados de maneira a captar algumas dependências entre os *streams* de características [190], aligeirando assim a assunção de independência entre eles. No entanto, não existem resultados publicados e conhecidos pelo autor que atestem uma significativa e consistente supremacia desta solução.

Apesar do modelo de recombinação aqui adoptado estar limitado a uma forma funcional básica relativamente às log-verosimilhanças associadas a cada *stream*, ainda assim dispõe de grande flexibilidade para incorporar, através do coeficiente de recombinação ω, conhecimento fundamental na abordagem deste problema. Se existir informação sobre o desempenho relativo das representações acústicas, com base em conhecimento apriorístico ou a partir de alguma(s) característica(s) extraída(s) do sinal durante o processo de reconhecimento, então através de ω pode ser estabelecido o peso adequado de cada *stream* no resultado final.

Neste problema concreto verifica-se que o mérito relativo dos *streams* está significativamente relacionado com a identidade da unidade elementar, u, correspondente ao modelo acústico a que pertence o estado em questão e também com a frequência fundamental, F_0, do sinal a reconhecer:

$$\omega=\omega\left(u,\overline{F}_0\right) \tag{3.30}$$

Quanto à dependência de ω relativamente a u, é um resultado conhecido que apesar dos MFCC apresentarem um melhor desempenho global em reconhecedores independentes do falante, alguns fonemas, nomeadamente as consoantes oclusivas [193], podem ser melhor representadas pelos LFCC.

Apesar de não se poder estabelecer com rigor uma relação simples entre a precisão dos modelos acústicos e o desvio padrão médio das funções de Gauss que compõem as respectivas misturas, em princípio os modelos acústicos mais precisos correspondem a misturas com gausseanas mais "fechadas". Logo, uma vez que neste caso os modelos são estruturalmente idênticos e foram desenvolvidos também pelo mesmo processo, gausseanas mais "fechadas" corresponderão a uma representação

mais precisa. Assim, relativamente aos modelos acústicos correspondentes a ambos os *streams*, para todos os grupos de fonemas definidos neste sistema foi estimada a média dos desvios padrão das funções de Gauss utilizadas nas respectivas misturas nos três estados.

Esses resultados, apresentados na figura 3.8[1], mostram que as gausseanas-LFCC são mais selectivas[2] que as gausseanas-MFCC nos modelos correspondentes às consoantes oclusivas, fricativas e africadas e que se verifica uma situação claramente oposta em relação às vogais.

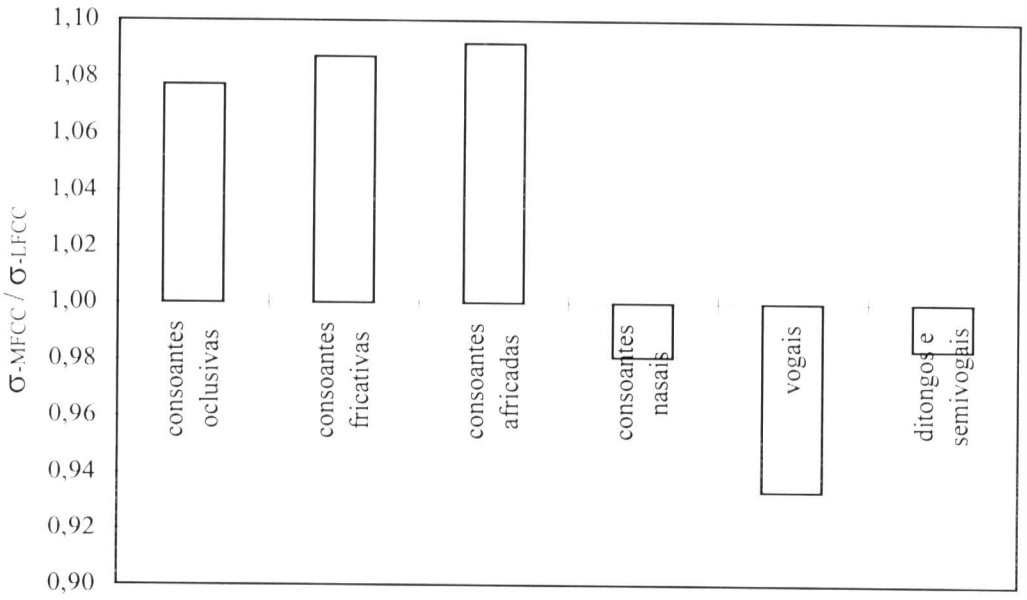

Fig. 3.8 - Relação entre o desvio médio das gausseanas dos modelos acústicos correspondentes aos *streams* MFCC e LFCC para várias classes de sons.

Comparando os resultados correspondentes, por exemplo, às consoantes oclusivas e às vogais, constata-se que eles não são surpreendentes. O espectro das consoantes oclusivas apresenta uma acentuada concentração da energia nas frequências mais elevadas, onde reside também a informação discriminante, ao contrário do que acontece nas vogais, cuja representação na região espectral correspondente às formantes de menor ordem é essencial para o reconhecimento. Por outro lado, comparando as características MFCC com as LFCC, verifica-se que as primeiras

[1] Os valores apresentados foram calculados considerando em cada mistura apenas metade das gausseanas, aquelas cujos coeficientes na mistura são mais elevados; no entanto, refira-se que os resultados são semelhantes, e permitem conclusões idênticas às que aqui se extraem, considerando diferentes partições das misturas, ou até mesmo as misturas completas.

[2] Note-se que o domínio das gausseanas está definido em $\Re^{39}$, logo o aumento do desvio médio em, por exemplo, 10% equivale a uma redução para aproximadamente 1/41 do "pico" da função.

apresentam claramente melhor resolução às baixas frequências e as segundas às frequências mais elevadas (Fig.3.3). Logo, é natural que os modelos acústicos MFCC sejam mais precisos que os LFCC na modelação das vogais e menos precisos no caso das consoantes oclusivas.

No modelo de recombinação utilizado neste sistema, a dependência de ω relativamente a u está definida de maneira independente para cada uma das 52 unidades elementares de fala (§2.6.2.2) em que se baseia o modelo acústico.

A frequência fundamental do sinal acústico é o factor dinâmico integrado no modelo de recombinação deste sistema.

Inspirado no classificador de género referido a propósito da estimação da frequência fundamental média (§3.3.4.2), foram medidas as taxas de palavras erradas (§2.9.3.4) correspondentes a dois grupos de falantes diferenciados apenas pelo valor médio da frequência fundamental, tomando como referência o valor 160 Hz. Esses resultados, registados na tabela 3.1, foram obtidos sobre o próprio conjunto de treino com dois reconhecedores *single-stream*, um deles baseado nas características MFCC e o outro nas LFCC. Note-se que esses reconhecedores apresentam uma estrutura idêntica e são treinados também de maneira idêntica (§2.9.3; §3.4.6), sendo apenas diferente a análise do sinal.

	Nº falantes	Nº palavras	WER_{MFCC} (%)	WER_{LFCC} (%)	ε_{WER} ($\%_{rel}$)
$F_0 < 160$ (Hz)	49	17197	4,55	4,93	**8,4**
$F_0 \geq 160$ (Hz)	23	8031	3,75	3,88	**3,5**
$\forall F_0$	72	25228	4,29	4,59	7,0

Tab. 3.1 - Resultados baseados na taxa de palavras erradas (WER) dos reconhecedores *single-stream* MFCC e LFCC para grupos de falantes diferenciados pelo valor de F_0.

Apesar destes resultados terem sido obtidos com o próprio conjunto de treino dos modelos, não deixam de ser um indicador significativo atendendo também ao assinalável volume dos dados, totalizando cerca de 25000 palavras.

O aspecto que interessa aqui salientar reside no facto de apesar das características MFCC proporcionarem um melhor desempenho, em relação às LFCC, para os dois grupos de falantes, a diferença entre as taxas do erro é substancialmente maior no caso dos falantes a que corresponde um valor de F_0 mais baixo. Nesses falantes o erro cresce, relativamente ao reconhecedor baseado nos MFCC, em cerca de $8{,}4\%_{rel}$, enquanto nos falantes com o valor de F_0 elevado o crescimento é de apenas $3{,}5\%_{rel}$, um valor cerca de 2,4 vezes inferior.

Ainda relativamente a estes dois grupos de falantes e utilizando os mesmos reconhecedores e conjunto de dados de teste, foi observado um outro indicador de desempenho do sistema, embora correlacionado com o anterior: a percentagem de frases correctas (§2.9.3.4). Na figura 3.9 são mostrados com base nesses resultados as percentagens de falantes, em cada um dos dois grupos, que obtêm mais frases correctas, num total de 40 frases por falante, com as características MFCC, nas colunas da esquerda, ou com as características LFCC, nas colunas da direita. As colunas ao centro correspondem aos falantes que obtêm o mesmo número de frases correctas com os dois reconhecedores.

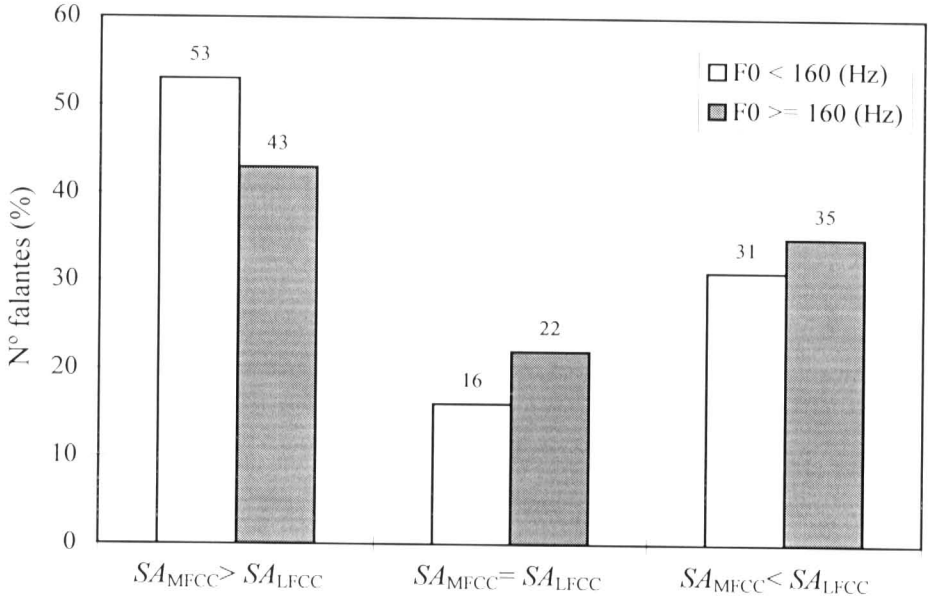

Fig. 3.9 - Resultados baseados na taxa de frases correctas (SA) dos reconhecedores *single-stream* MFCC e LFCC para grupos de falantes diferenciados pelo valor de F_0.

Note-se, por exemplo, que 53% dos falantes com o valor de F_0 mais baixo têm um maior número de frases correctas com o reconhecedor baseado nos MFCC do que com o reconhecedor utilizando os LFCC, enquanto que no outro grupo de falantes isso se verifica em apenas 43% deles.

O desempenho relativo das características MFCC e LFCC em função da frequência fundamental média do sinal tem origem nas diferenças entre essas duas representações ao nível da resolução espectral. Foi realizado algum trabalho preparatório no sentido de investigar os mecanismos envolvidos nessa relação, uma vez que esses resultados seriam certamente úteis na abordagem ao problema da modelação da dependência de ω relativamente a F_0. No entanto, uma vez que esse estudo exige um grande esforço, optou-se por seguir uma abordagem baseada no arbítrio, condicionado a determinados pressupostos, de uma forma funcional para

modelar a dependência estatística entre ω e F_0, seguindo-se o treino dos parâmetros livres dessa função segundo um determinado critério e a sua posterior avaliação.

Mesmo admitindo, por hipótese, que a dependência real de ω relativamente ao valor médio de F_0 é muito complexa e também precisa, a dificuldade em obter uma estimativa fidedigna deste valor (§3.3.4.2) impõe a utilização de uma função simples, por exemplo uma função polinomial de ordem reduzida ou, tal como se considerou inicialmente:

$$f\left(\overline{F}_0\right) = \alpha + \beta \log\left(\frac{\overline{F}_0}{160}\right) \tag{3.31}$$

Através de uma avaliação preparatória do reconhecedor num conjunto independente de teste[1], tornou-se evidente a necessidade de modelar a dependência de ω relativamente a F_0 através de uma forma funcional que permitisse uma variação mais rápida num intervalo do domínio em torno de aproximadamente 160 Hz. Optou-se então por definir $f\left(\overline{F}_0\right)$ em duas partes, de acordo com a equação 3.34.

Apesar de não se ter verificado uma influência significativa destes parâmetros no desempenho final do reconhecedor, sobretudo pelo aspecto da coerência formal[2] resolveu-se estender a recombinação pesada às probabilidades de transição (3.24) do modelo *composto*.

Está então formalizado o modelo de recombinação, integrando ambos os factores u e F_0, utilizado neste reconhecedor *multi-stream*:

$$\log\left(a_{\left(\begin{smallmatrix}1&2\\i,i\end{smallmatrix}\right),\left(\begin{smallmatrix}1&2\\j,j\end{smallmatrix}\right)}\right) = \omega\left(u,\overline{F}_0\right)\log\left(\overset{1}{a}{}^{1\,1}_{i,j}\right) + \left(1-\omega\left(u,\overline{F}_0\right)\right)\log\left(\overset{2}{a}{}^{2\,2}_{i,j}\right) \tag{3.32}$$

$$\log\left(b_{\left(\begin{smallmatrix}1&2\\j,j\end{smallmatrix}\right)}\left(\overset{1}{\boldsymbol{x}}_t,\overset{2}{\boldsymbol{x}}_t\right)\right) = \omega\left(u,\overline{F}_0\right)\log\left(\overset{1}{b}{}^{1}_{j}\left(\overset{1}{\boldsymbol{x}}_t\right)\right) + \left(1-\omega\left(u,\overline{F}_0\right)\right)\log\left(\overset{2}{b}{}^{2}_{j}\left(\overset{2}{\boldsymbol{x}}_t\right)\right) \tag{3.33}$$

com:

[1] Exactamente 1/3 das frases do conjunto de teste February'98, 10 frases extraídas de maneira salteada (1-0-0) de cada um dos 10 falantes.

[2] E obviamente porque isso não acarreta também qualquer custo adicional.

$$\omega\left(u,\overline{F}_0\right)=\begin{cases}\alpha_u, & \overline{F}_0<160\\ \beta_u+\chi_u\,log\left(\dfrac{\overline{F}_0}{160}\right), & \overline{F}_0\geq 160\end{cases},\qquad u=1,2,...,52 \tag{3.34}$$

Para definir por completo o modelo de recombinação resta determinar, com base num determinado critério, o conjunto de parâmetros $\{\alpha_u, \beta_u, \chi_u,\ u=1,2,...,52\}$ (§3.4.6.3).

3.4.6 Treino dos modelos acústicos *compostos*

3.4.6.1 Introdução

Os reconhecedores baseados no formalismo *multi-stream*, tal como em geral os sistemas que recorrem a representações múltiplas, em princípio fazem um melhor aproveitamento dos dados disponíveis para treino porque podem extrair maior quantidade de informação do sinal. Pela mesma razão, podem também suportar melhor as eventuais insuficiências de dados para treino [183].

Em relação ao tempo de treino, apesar dos modelos acústicos *compostos* serem mais complexos que os respectivos modelos *single-stream*, não exigem necessariamente mais tempo, pois é possível estimá-los simplesmente combinando os parâmetros destes últimos, os quais podem obviamente ser treinados em simultâneo desde que existam recursos de computação disponíveis.

Apesar desta abordagem ao treino ser usual, tem sido notado [186] que dessa maneira não é considerada a restrição relativa à sincronização dos *streams* nas transições entre as unidades acústicas elementares. É claro que isso pode ser feito treinando directamente os modelos *compostos*, usando um processo baseado na segmentação Viterbi essencialmente idêntico ao que foi utilizado no treino dos modelos *single-stream* desenvolvidos neste trabalho. No entanto, não são conhecidos pelo autor resultados que atestem qualquer ganho significativo no desempenho do reconhecedor usando esta técnica de treino.

A topologia dos SCHMM correspondentes aos símbolos fonéticos está representada na figura 3.10. Tal como nos reconhecedores *single-stream*, o silêncio é modelado por um único estado.

Os parâmetros dos 52 modelos *compostos* utilizados neste sistema, correspondendo ao mesmo conjunto de símbolos linguísticos usados no reconhecedor *single-stream* (§2.6.2.2), são estimados com base nos respectivos modelos (§2.6.3.3) desses reconhecedores. As probabilidades de transição do modelo *composto* são calculadas através das equações 3.32 e 3.34 a partir das respectivas probabilidades dos

modelos *single-stream* treinadas (§2.9.3; §3.4.6.2) com a equação 2.66. Os parâmetros do modelo de observação conjunta em cada estado são extraídos, de acordo com as equações 3.33 e 3.34, do *codebook* e respectivos coeficientes das misturas dos modelos *single-stream* treinados (§2.9.3; §3.4.6.2) através das equações 2.67 a 2.69. Os parâmetros, correspondentes à equação 3.34, do modelo de recombinação são neste sistema treinados com base num critério aproximado de maximização da verosimilhança (§3.4.6.3).

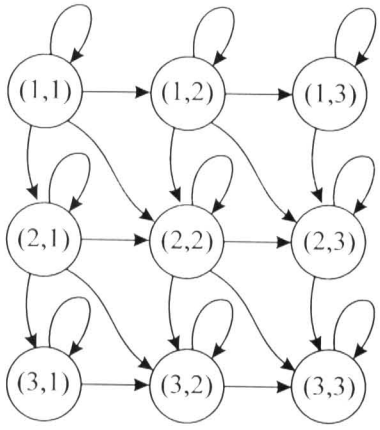

Fig. 3.10 - Modelo *composto* correspondente a um símbolo fonético.

3.4.6.2 Treino dos parâmetros dos modelos *single-stream*

Na secção §2.9 está descrito o treino do modelo acústico correspondente ao reconhecedor baseado nas características MFCC.

Nesta secção é sobretudo feita a apresentação dos resultados obtidos durante o treino do reconhecedor baseado nos LFCC. Quanto ao critério, estratégia e procedimentos de treino, tal como em relação às bases de dados de treino e teste, são idênticos aos utilizados nos modelos mel-cepstrais (§2.9.2; §2.9.3). O mesmo se verifica relativamente às medidas de avaliação do desempenho do sistema, *WER* e *SA* (§2.9.3.4), e à notação referente a uma configuração particular do modelo acústico, $(N_1;N_2)_I$.

O sistema foi iniciado com a configuração $(512;512)_0$, com as medidas *WER* e *SA* apresentando os valores, respectivamente, 31,44% e 23,67%. Foram executadas 12 iterações de adaptação dos parâmetros, com *WER* atingindo o valor 15,30%. As misturas foram reduzidas utilizando um valor de ε igual a 10^{-5}, resultando um sistema com a configuração $(107;512)_{12}$, com os indicadores *WER* e *SA* mantendo-se inalterados. Dado que *WER* não diminuíu significativamente correndo mais iterações de adaptação com esta configuração, o *codebook* foi duplicado, resultando o sistema $(214;1K)_{12}$ com *WER* e *SA* iguais a, respectivamente, 15,22% e 48,00%.

A partir do sistema $(214,1K)_{12}$, 14 iterações de adaptação dos parâmetros foram executadas atingindo *WER* o valor 12,40%. As misturas gausseanas existentes na iteração 26 foram reduzidas, com o parâmetro ε igual a 10^{-5}, resultando o sistema $(140,1K)_{26}$, sem alteração dos valores de *WER* e *SA*. O desempenho não melhorou com mais iterações sobre a configuração (140,1K). Assim, o *codebook* foi duplicado resultando o sistema $(279;2K)_{26}$, com *WER* e *SA* iguais a, respectivamente, 12,60% e 54,00%.

Os modelos acústicos correspondentes ao sistema $(279,2K)_{26}$ foram então submetidos a 16 iterações de treino até ser atingido um mínimo local de *WER* com o valor 10,43%. Com o parâmetro ε igual a 10^{-5}, as misturas foram reduzidas resultando o sistema $(190,2K)_{42}$ com *WER* e *SA* mantendo os valores anteriores. Não se verificou uma redução significativa da taxa do erro prosseguindo o treino do sistema nesta configuração. Então, o *codebook* foi duplicado resultando o sistema $(378;4K)_{42}$, com as medidas do desempenho do sistema mantendo os valores.

A partir do sistema $(378,4K)_{42}$ o treino foi continuado, tendo o *WER* alcançado o valor mínimo 9,86% na iteração 48. Com ε igual a 10^{-4}, as misturas gausseanas foram reduzidas resultando o sistema $(307,4K)_{48}$, sem alteração da taxa do erro. Não se verificou a melhoria do desempenho do sistema com a continuação do procedimento de treino usando este *codebook*, pelo que os parâmetros acústicos correspondentes ao sistema $(307,4K)_{48}$ foram tomados como o resultado do processo de treino. Os respectivos valores de *WER* e *SA* são, respectivamente, 9,86% e 61,67%.

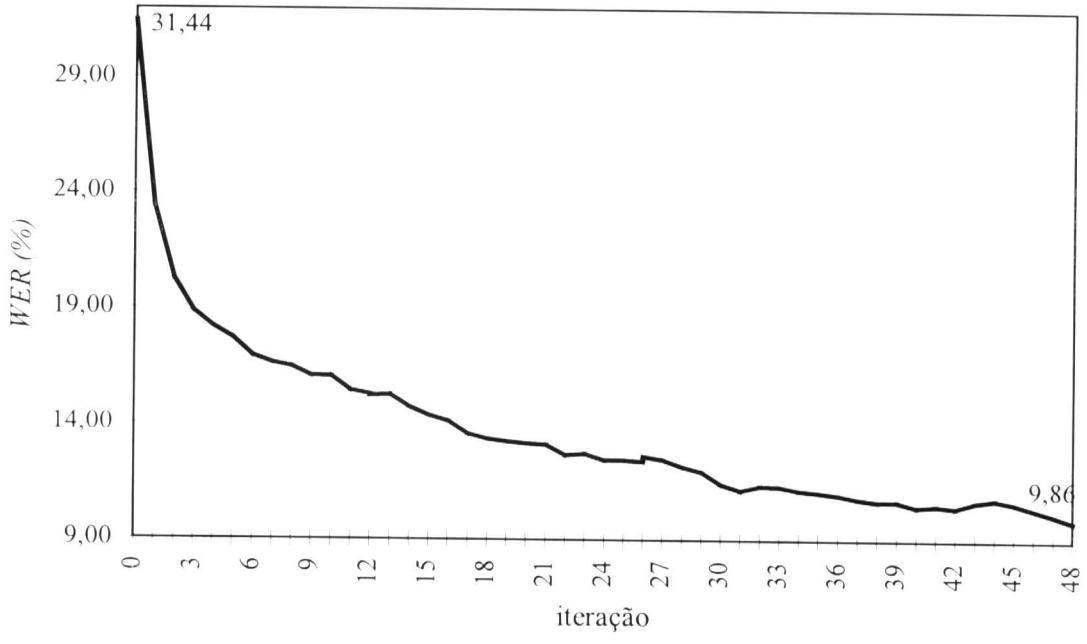

Fig. 3.11 - Evolução da taxa de palavras erradas ao longo do treino.

As figuras 3.11 a 3.14 mostram a evolução de diferentes medidas do desempenho do sistema durante a fase de treino.

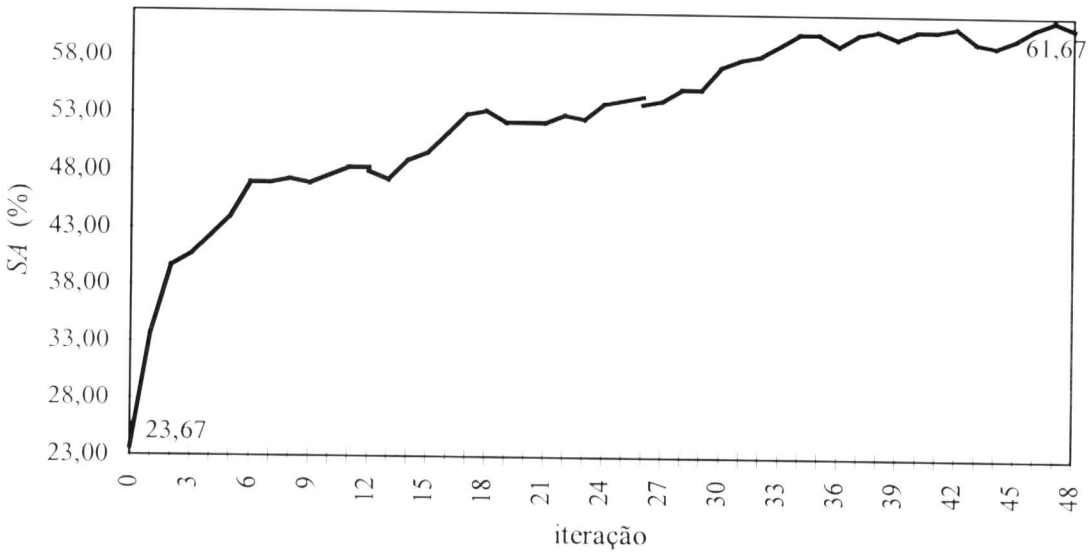

Fig. 3.12 - Evolução da taxa de frases correctas.

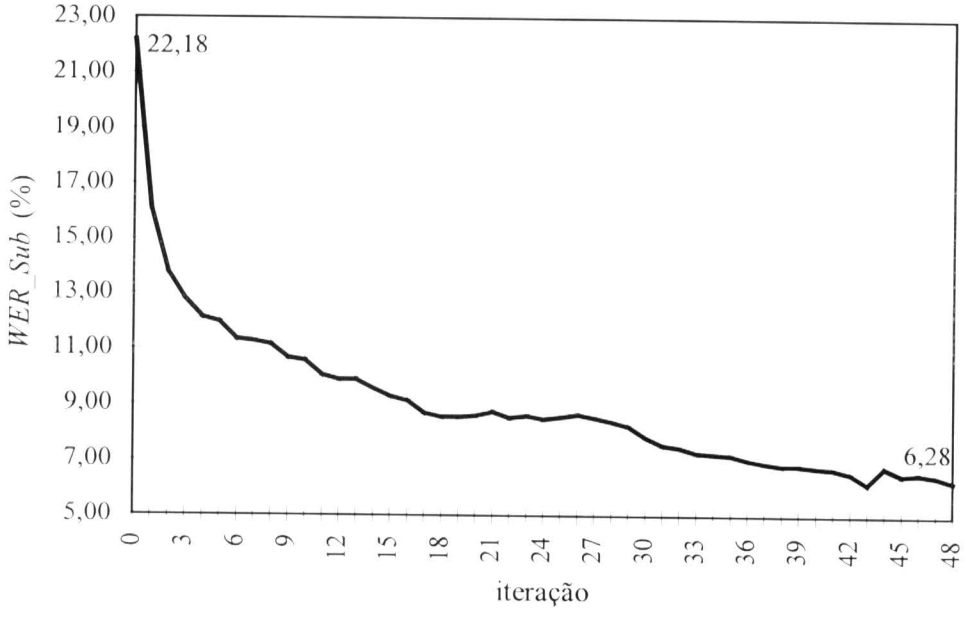

Fig. 3.13 - Evolução da taxa de erro correspondente às palavras substituídas.

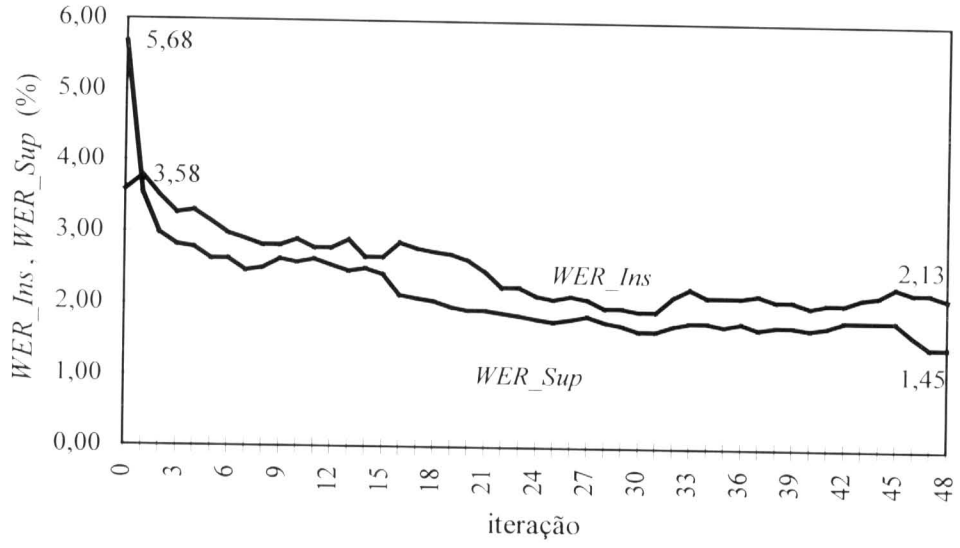

Fig. 3.14 - Evolução do erro correspondente às palavras inseridas e suprimidas.

A figura 3.15 mostra a evolução monótona do logaritmo da verosimilhança do modelo acústico na "observação" do conjunto de treino.

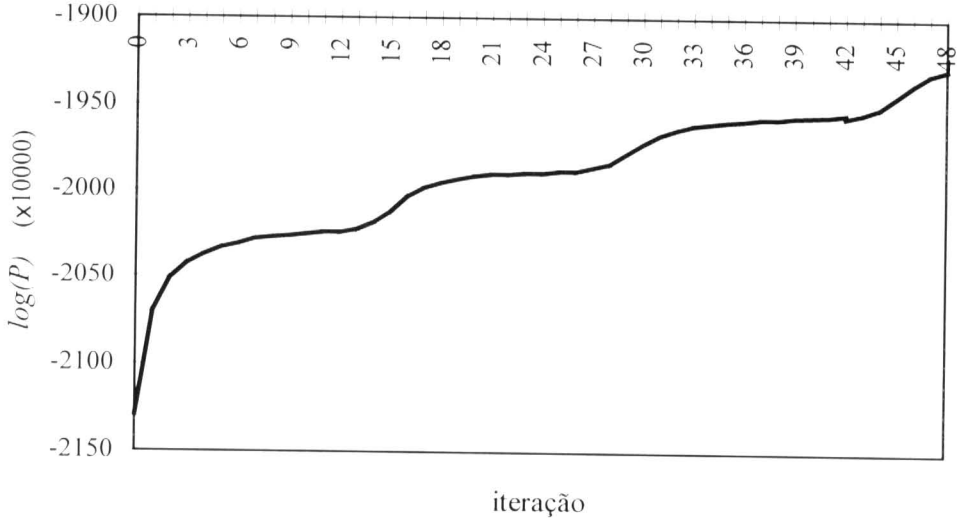

Fig. 3.15 - Evolução da log-verosimilhança acumulada sobre o conjunto de treino.

A figura 3.16 mostra a distribuição da dimensão das misturas gausseanas utilizadas na configuração final do sistema. Como se pode verificar, a grande maioria das misturas não utiliza mais do que a décima parte do *codebook*.

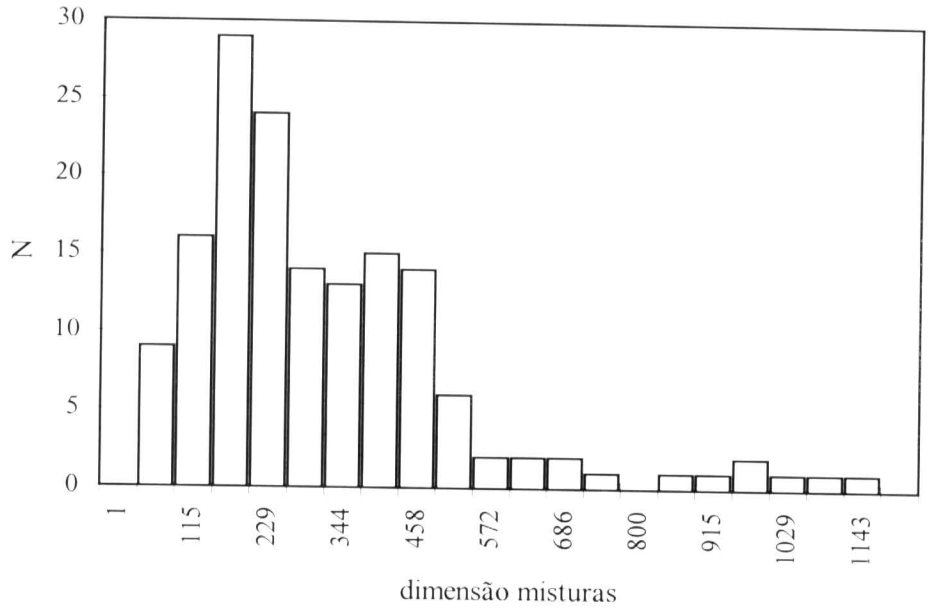

Fig. 3.16 - Histograma da dimensão das misturas do sistema (307,4K).

A figura 3.17 apresenta uma outra perspectiva da utilização do *codebook* correspondente aos modelos acústicos definitivos. Pode-se por exemplo verificar que a maior parte das funções de Gauss não é utilizada por mais do que aproximadamente a décima parte das 154 misturas diferentes necessárias para definir os modelos acústicos.

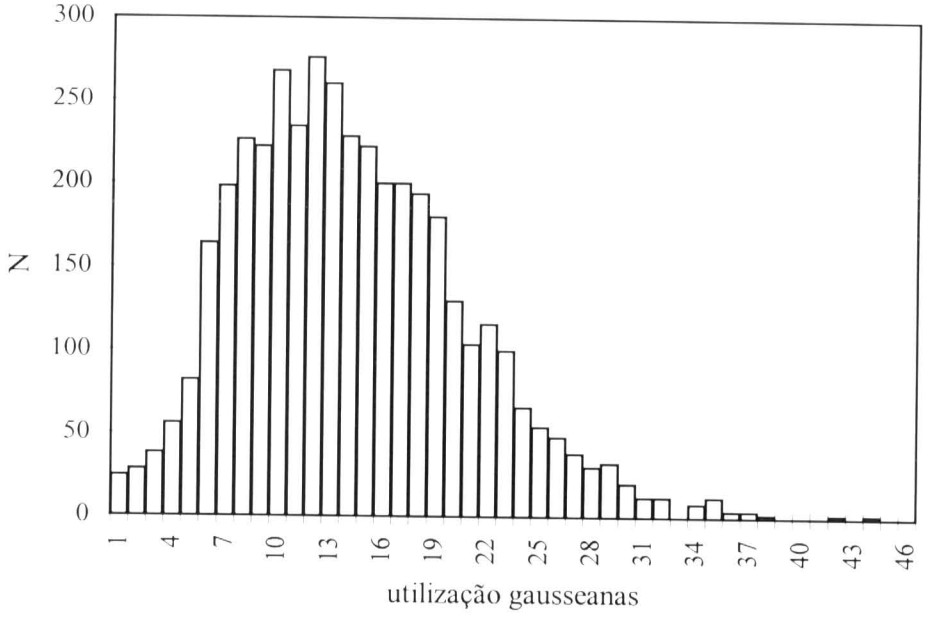

Fig. 3.17 - Histograma da utilização do *codebook* pelo sistema (307,4K).

À semelhança da operação efectuada com os modelos acústicos correspondentes à análise Mel-cepstral, também para estes modelos foram simultaneamente optimizados os parâmetros *WP* e α, ambos utilizados no algoritmo de descodificação (§2.9.3.4). A tabela 3.2 confronta os resultados obtidos sem a sua utilização no descodificador com os obtidos após o procedimento de optimização com o critério de minimizar *WER*. Como se pode verificar obteve-se uma melhoria da taxa de erro de aproximadamente $4{,}5\%_{\text{rel}}$.

	WER	*WER_Sub*	*WER_Ins*	*WER_Sup*	*SA*
WP=0.00; α=1.00	9,86%	6,28%	2,13%	1,45%	61,67%
WP=-1.50; α=0.60	9,42%	6,00%	0,60%	2,82%	59,67%

Tab. 3.2 - Desempenho do sistema $(307{,}4\text{K})_{48}$ antes e após a optimização de *WP* e α.

3.4.6.3 Treino do modelo de recombinação

Estabelecida a forma funcional 3.34 para modelar a recombinação das verosimilhanças em cada estado (§3.4.5), é então necessário fixar o conjunto dos parâmetros livres:

$$\theta = \left\{ \alpha_u,\ \beta_u,\ \chi_u,\ u = 1,2,\ldots,52 \right\} \tag{3.35}$$

de maneira a satisfazer um determinado critério.

O critério que aqui se estabelece baseia-se numa aproximação à maximização da verosimilhança da observação conjunta dos *streams* estendida a todas as frases do conjunto de treino.

O conjunto dos parâmetros deve então aproximar-se da solução da equação:

$$\theta = \arg\max_{\theta} \prod_{n=1}^{N_T} p_\theta\left(\overset{1}{\boldsymbol{X}}^n, \overset{2}{\boldsymbol{X}}^n \middle| W_c^{\,n} \right) \tag{3.36}$$

sendo T o conjunto de treino (§2.9.2) constituído por N_T exemplos $\left(\overset{1}{\boldsymbol{X}}^n, \overset{2}{\boldsymbol{X}}^n, W_c^{\,n} \right)$.

Aplicando a transformação logarítmica a esta equação e recorrendo à equação 3.20, obtém-se:

$$\theta = arg\max_{\theta} \sum_{n=1}^{N_T} \max_{S_n} \left\{ \sum_{m=1}^{M_n} log\, p\left(\overset{1}{\boldsymbol{X}}_m^n, \overset{2}{\boldsymbol{X}}_m^n | u_m^n \right) \right\} \tag{3.37}$$

Tal como no treino dos modelos acústicos correspondentes ao reconhecedor *single-stream* desenvolvido neste trabalho[1], o critério da máxima verosimilhança é aproximado considerando, em cada exemplo, apenas a contribuição do melhor caminho (§2.9.3.3) possível no modelo acústico global resultante da concatenação dos modelos *compostos* correspondentes a W. A segmentação dos *streams* de características, inicialmente efectuada apenas ao nível das unidades acústicas elementares tal como proposto na formalização do modelo de recombinação (§3.4.3), passa então a estar definida também ao nível dos estados dos modelos *compostos*.

Assim, a equação 3.37 pode ser aproximada por:

$$\theta = arg\max_{\theta} \sum_{n=1}^{N_T} log\, p_{\left(\overset{1}{q},\overset{2}{q}\right)^n}\left(\overset{1}{\boldsymbol{X}}^n, \overset{2}{\boldsymbol{X}}^n | W_c^n \right) \tag{3.38}$$

com a log-verosimilhança, num exemplo genérico com comprimento T, calculada apenas sobre o caminho:

$$\left(\overset{1}{q},\overset{2}{q}\right) = arg \max_{\left(\overset{1}{q},\overset{2}{q}\right) \in Q_W^T} \left\{ p\left(\overset{1}{\boldsymbol{X}}_1^T, \overset{2}{\boldsymbol{X}}_1^T, \left(\overset{1}{q},\overset{2}{q}\right) | W \right) \right\} \tag{3.39}$$

em que Q_W^T representa o conjunto de todos os caminhos com comprimento T e pertencentes ao modelo acústico associado a W.

A solução da equação 3.38 pode ser aproximada considerando apenas a contribuição das funções densidade de probabilidade de observação nos sucessivos estados pertencentes ao melhor caminho:

$$\theta = arg\max_{\theta} \sum_{n=1}^{N_T} \sum_{t=1}^{T_n} log\, b_{\left(\overset{1}{q}_t,\overset{2}{q}_t\right)}\left(\overset{1}{x}_t^n, \overset{2}{x}_t^n \right) \tag{3.40}$$

Esta aproximação simplifica em alguma medida o procedimento de treino e implica desprezar apenas a contribuição das probabilidades de transição na estimação dos parâmetros do modelo de recombinação, o que é plenamente aceitável face à reduzida influência dessas probabilidades no desempenho do sistema.

[1] Assim como na descodificação, em ambos os sistemas.

A partir desta equação e do modelo de recombinação, tal como se apresenta nas equações 3.33 e 3.34, obtém-se:

(3.41)

$$\theta = arg\max_{\theta} \sum_{n=1}^{N_T} \left[\sum_{\substack{t=1 \\ (\overline{F}_0<160)}}^{T_n} \alpha_{u\left(q_t^1, q_t^2\right)} \left(log\ b_{q_t^1}^1\left(x_t^{1\,n}\right) - log\ b_{q_t^2}^2\left(x_t^{2\,n}\right) \right) + \sum_{\substack{t=1 \\ (\overline{F}_0\geq 160)}}^{T_n} \left(\beta_{u\left(q_t^1, q_t^2\right)} + \chi_{u\left(q_t^1, q_t^2\right)} log\left(\frac{\overline{F}_0}{160}\right) \right) \left(log\ b_{q_t^1}^1\left(x_t^{1\,n}\right) - log\ b_{q_t^2}^2\left(x_t^{2\,n}\right) \right) \right]$$

A contribuição de cada exemplo de treino é no sentido de atribuir maior peso, em cada estado pertencente ao melhor caminho, ao *stream* que se apresenta mais verosímil. Para evitar que a contribuição das ocorrências seja dependente da sua duração, a log-verosimilhança correspondente a cada exemplo da unidade acústica elementar é normalizada, dividindo esse valor pelo número de sectores do respectivo segmento.

$$\gamma\left(x_t^{1\,n}, x_t^{2\,n}, q_t^1, q_t^2\right) = \frac{log\ b_{q_t^1}^1\left(x_t^{1\,n}\right) - log\ b_{q_t^2}^2\left(x_t^{2\,n}\right)}{comprimento\ do\ segmento} \tag{3.42}$$

Admitindo que a estimativa da frequência fundamental média em todos os sectores do sinal pertencentes à mesma frase de treino é exclusivamente superior, ou inferior, a 160Hz, uma aproximação razoável dado que esta é uma fronteira natural, então, com base nas equações 3.41 e 3.42, o objectivo do critério de treino pode ser aproximado pela maximização das funções independentes:

$$f_1(\theta_1) = \sum_{n=1}^{N_T} \sum_{\substack{t=1 \\ (\overline{F}_0<160)}}^{T_n} \alpha_{u\left(q_t^1, q_t^2\right)}\ \gamma\left(x_t^{1\,n}, x_t^{2\,n}, q_t^1, q_t^2\right) \tag{3.43}$$

definida no espaço de representação dos vectores:

$$\theta_1 = (\alpha_1, \alpha_2, \ldots, \alpha_{52}) \tag{3.44}$$

e:

$$f_2(\theta_2) = \sum_{n=1}^{N_T} \sum_{\substack{t=1 \\ (\overline{F}_0\geq 160)}}^{T_n} \left(\beta_{u\left(q_t^1, q_t^2\right)} + \chi_{u\left(q_t^1, q_t^2\right)} log\left(\frac{\overline{F}_0}{160}\right) \right) \gamma\left(x_t^{1\,n}, x_t^{2\,n}, q_t^1, q_t^2\right) \tag{3.45}$$

definida para a variável vectorial independente:

$$\theta_2 = (\beta_1,\ \chi_1, \beta_2,\ \chi_2, \ldots, \beta_{52},\ \chi_{52}) \tag{3.46}$$

A estimação dos vectores de parâmetros, θ_1 e θ_2, é então um problema de optimização das funções de campo 3.43 e 3.45, respectivamente. É aqui seguida a abordagem típica para resolver este problema, utilizando um procedimento iterativo baseado no gradiente da função objectivo.

Assim, os parâmetros são iniciados:

$$\theta_1^{(0)} = (\alpha_1^0, \alpha_2^0, \ldots,\ \alpha_{52}^0) \tag{3.47}$$

$$\theta_2^{(0)} = (\beta_1^0,\ \chi_1^0, \beta_2^0,\ \chi_2^0, \ldots, \beta_{52}^0,\ \chi_{52}^0) \tag{3.48}$$

sendo então executado o processo iterativo de adaptação:

$$\theta_i^{(k+1)} = \theta_i^{(k)} + \lambda_i\ \nabla_i\left(\theta_i^{(k)}\right), \quad k = 1,2,\ldots \qquad i = 1,\ 2 \tag{3.49}$$

Admitindo que o gradiente não varia significativamente ao longo das sucessivas iterações, o processo de optimização pode ser aproximado executando apenas a primeira iteração:

$$\theta_i = \theta_i^{(0)} + \lambda_i\ \nabla_i\left(\theta_i^{(0)}\right), \qquad i = 1,\ 2 \tag{3.50}$$

e ajustando de maneira independente o valor dos parâmetros λ_1 e λ_2 com base num determinado critério de desempenho do sistema.

Deve ser salientado o interesse desta aproximação do ponto de vista do esforço de computação, pois tipicamente a convergência deste método de optimização é lenta. Por outro lado, não se considerou ser justificável recorrer a métodos de optimização baseados em aproximações de maior ordem à hiper-superfície local definida pela função objectivo [91].

Em relação à consistência desta aproximação, apesar de não se ter tentado confirmar experimentalmente esta hipótese existe a convicção de que a segmentação dos *streams* não se altera de maneira muito substancial quando os parâmetros do modelo de recombinação variam num pequeno intervalo, tal como sucede neste caso. Admitindo que se confirma esta hipótese, então a aproximação será aceitável.

Assim, partindo dos valores inicialmente arbitrados (3.47 e 3.48), os parâmetros do modelo de recombinação são estimados através das seguintes fórmulas:

$$\alpha_u = \alpha_u^0 + \lambda_1 \frac{1}{N_{u1}} \sum_{n=1}^{N_T} \sum_{\substack{t=1 \\ (q_t^1, q_t^2) \in u \\ (\overline{F}_0 < 160)}}^{T_n} \gamma\left(x_t^{1\,n}, x_t^{2\,n}, q_t^1, q_t^2\right), \qquad u = 1,\ 2,\ \ldots,\ 52 \tag{3.51}$$

$$\beta_u = \beta_u^0 + \lambda_2 \frac{1}{N_{u2}} \sum_{n=1}^{N_T} \sum_{\substack{t=1 \\ (q_t^1, q_t^2) \in u \\ (\overline{F}_0 \geq 160)}}^{T_n} \gamma\left(x_t^{1\,n}, x_t^{2\,n}, q_t^1, q_t^2\right), \qquad u = 1,\ 2,\ \ldots,\ 52 \tag{3.52}$$

$$\chi_u = \chi_u^0 + \lambda_2 \frac{1}{N_{u2}} \sum_{n=1}^{N_T} \sum_{\substack{t=1 \\ (q_t^1, q_t^2) \in u \\ (\overline{F}_0 \geq 160)}}^{T_n} log\left(\frac{\overline{F}_0}{160}\right) \gamma\left(x_t^{1\,n}, x_t^{2\,n}, q_t^1, q_t^2\right), \quad u = 1,\ \ldots,\ 52 \tag{3.53}$$

Na equação 3.51 é introduzido o factor $1/N_{u1}$ no termo do gradiente, em que N_{u1} é o número de ocorrências de u no conjunto de treino com $F_0<160$, para evitar que a direcção do gradiente seja dominada pelos símbolos mais frequentes. Note-se que entre alguns símbolos existe um diferença de duas ordens de grandeza no número de ocorrências. Nas equações 3.52 e 3.53 é realizada uma operação idêntica mas para ocorrências correspondentes a $F_0 \geq 160$.

Na secção §3.5 são analisados os resultados obtidos com os parâmetros estimados, a partir dos respectivos pontos iniciais, com diferentes valores dos passos λ_1 e λ_2.

3.4.7 Descodificação

No essencial, o processo de descodificação neste reconhecedor é semelhante ao que ocorre no sistema *single-stream* desenvolvido também neste trabalho (§2.8).

A estrutura de informação que integra todas as fontes de conhecimento associadas a esta tarefa é idêntica ao nível do modelo linguístico, verificando-se apenas diferenças no modelo acústico. Em concreto, a única alteração consiste na substituição

do SCHMM *standard* (§2.6.3.3) correspondente ao sistema *single-stream*, em cada uma as 52 unidades acústicas elementares, pelo respectivo modelo *composto* (§3.4.4).

A utilização destes modelos permite que a descodificação possa ser efectuada com base no algoritmo Viterbi. É claro que a verosimilhança global associada a uma hipótese de reconhecimento deixa de ser calculada através da equação 3.20, de acordo com a formulação original do modelo *multi-stream*, para passar a sê-lo com base na verosimilhança calculada sobre o melhor caminho, definido na equação 3.39. No entanto, esta abordagem é frequente [180,182,183,186,192] face à evidência experimental da obtenção de resultados semelhantes aos que podem ser obtidos utilizando técnicas de descodificação mais complexas, como é o algoritmo conhecido por *two-level DP*, directamente aplicadas sobre o formalismo *multi-stream*.

O descodificador que aqui se utiliza baseia-se na versão do algoritmo Viterbi construída para o sistema *single-stream* (§2.8.3). No essencial, para além do acesso aos estados passar a corresponder à topologia dos modelos acústicos *compostos*, apenas o cálculo da verosimilhança da observação do sinal em cada estado passa a ser diferente. É necessário avaliar as duas funções que modelam localmente as verosimilhanças das observações correspondentes aos dois *streams* e, então, dependendo também da estimativa dinâmica da frequência fundamental média, calcular o resultado pretendido de acordo com as equações 3.32, 3.33 e 3.34. O protótipo construído, com o qual foram obtidos todos os resultados que aqui se apresentam, não poderia ser utilizado em *tempo real* numa tarefa como aquela em que é testado, dado que exige um atraso que pode prolongar-se até 5 segundos para estimar do valor médio da frequência fundamental do sinal.

Apesar deste descodificador operar apenas sobre dois *streams* de características, o tempo de execução é bastante superior ao que se verifica, para tarefa idêntica, num sistema *single-stream*. Refira-se que em relação a esse sistema também o número de parâmetros do modelo acústico aumenta para aproximadamente o dobro.

Não sendo fácil efectuar uma abordagem analítica ao problema concreto da complexidade computacional, dada a dificuldade em controlar ou prever factores como por exemplo o número aproximado de estados activos[1], foi experimentalmente estimada a relação entre os tempos de execução correspondentes a diferentes sistemas utilizados neste trabalho. A figura 3.18 ilustra as diferentes soluções experimentadas na conexão dos modelos *compostos* elementares no interior de cada palavra e também na topologia do modelo, no caso (D). A solução (A) corresponde ao maior nível de relaxação do sincronismo[2] e, pelo contrário, em (D) os *streams* de características são forçados a manter o alinhamento na estrutura interna do modelo *multi-stream*.

[1] Naturalmente, são também aqui utilizadas técnicas de *prunning*.

[2] Por exemplo, um dos vectores de características de um dos *streams* pode "estar" no primeiro estado de um fonema, enquanto no mesmo instante o vector de características do outro *stream* "já está" no primeiro estado do fonema seguinte.

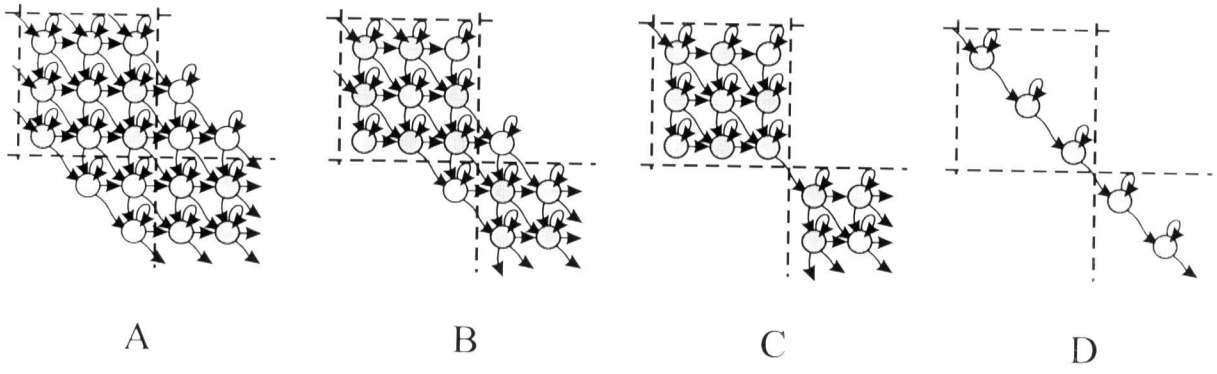

Fig. 3.18 - Conexões e topologias diferentes dos modelos *compostos*.

Na tabela 3.3 são apresentados os tempos de execução correspondentes a estas arquitecturas, tomando como referência o tempo de execução no descodificador *single-stream*[1]. Esses resultados foram obtidos com o valor do parâmetro *beam-width* (§2.8.3) próximo do necessário e suficiente para não se verificar a degradação do desempenho do reconhecedor. A tarefa que serviu de suporte a esta experiência consiste no reconhecimento de 300 frases, a que correspondem aproximadamente 10^5 vectores de características.

A	B	C	D
3,37 **x** $T_{1\text{-}str}$	3,19 **x** $T_{1\text{-}str}$	3,07 **x** $T_{1\text{-}str}$	2,80 **x** $T_{1\text{-}str}$

Tab. 3.3 - Tempos relativos de utilização de CPU, correspondentes às arquitecturas A, B, C e D (Fig. 3.18); $T_{1\text{-}str}$ corresponde ao descodificador *single-stream*.

Apesar destes resultados estarem necessariamente dependentes das soluções específicas utilizadas na concretização dos algoritmos, ainda assim são um bom indicador da complexidade computacional do descodificador *multi-stream*. Como se pode verificar, a sua execução é cerca de 3 vezes mais lenta comparativamente ao descodificador *single-stream*. Por outro lado, e apesar do número de estados dos modelos *compostos* variar bastante nas quatro arquitecturas, não se verifica uma diferença muito substancial nos tempos de execução. Em grande parte isso resulta da transição entre as palavras, precisamente onde se verifica uma grande parte do esforço de computação (§2.8.3), ser necessariamente efectuada através de um único estado em cada palavra. Um aspecto que poderá ter também alguma influência reside na eventualidade das topologias mais complexas, ao permitirem um alinhamento menos

[1] É curioso que o sistema *single-stream* baseado nos MFCC é em média aproximadamente 5% mais rápido que o baseado nos LFCC com igual *beam-width*, o que pode ser um indicador da maior precisão na modelação no primeiro sistema.

rígido, proporcionarem valores de verosimilhança mais elevados e assim reduzirem a largura do feixe de pesquisa.

3.5 Resultados

3.5.1 Comparação dos sistemas *single-stream*

Nesta secção são apresentados resultados da avaliação do desempenho do reconhecedor *multi-stream* construído, com base numa tarefa idêntica à estabelecida para os sistemas *single-stream* também desenvolvidos neste trabalho(§2.9.3.4). No entanto, é importante previamente salientar alguns aspectos do desempenho relativo dos reconhecedores *single-stream*.

Os resultados apresentados para esses reconhecedores mostram que existe uma clara degradação do seu desempenho quando se substituem as características MFCC pelas LFCC. Sobre o conjunto de treino, essa degradação é estimada em aproximadamente $7\%_{rel}$ (tabela 3.1). É importante salientar que essa diferença é bastante maior, estimada em aproximadamente $45\%_{rel}$, quando a avaliação é feita sobre o conjunto de teste *February'89* (§2.9.3.4, §3.4.6.2).

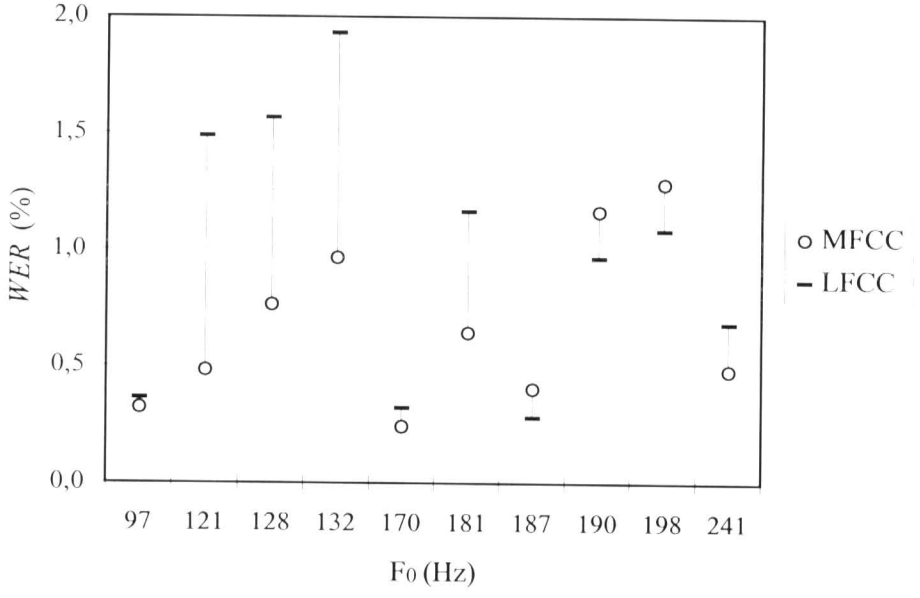

Fig. 3.19 - Taxa de erro, correspondente aos reconhecedores *single-stream* MFCC e LFCC, para todos os falantes do conjunto de teste.

A figura 3.19 mostra a taxa de palavras erradas, medida relativamente ao número total de palavras do conjunto de teste, para cada um dos 10 falantes do conjunto de

teste, quando se utiliza o reconhecedor *single-stream* baseado nas características MFCC e nas características LFCC. Em abcissas estão assinalados os valores estimados da frequência fundamental média para cada falante[1].

Por simples inspecção deste gráfico constata-se de imediato que os resultados são globalmente superiores no sistema baseado nos MFCC. É também possível verificar dois aspectos importantes: 1) essa superioridade é claramente maior para falantes pertencentes ao grupo com o valor de F_0 mais baixo; 2) as únicas excepções a essa superioridade ocorrem para falantes pertencentes ao grupo com o valor de F_0 mais elevado[2].

Dadas as limitações deste conjunto de teste, cuja representatividade no universo de falantes é certamente muito reduzida, não seriam consistentes eventuais normas, sobre a dependência entre F_0 e o desempenho relativo das características MFCC e LFCC, obtidas apenas por extrapolação destas constatações. Mas deve ser também destacado que estes resultados apontam no mesmo sentido dos verificados no conjunto de treino, suportados por um volume de dados significativo e portanto pouco susceptível de permitir conclusões fortuitas (§3.4.5).

Repartindo então os falantes em dois grupos, diferenciados apenas pelo valor médio da frequência fundamental[1], F_0, obtêm-se os resultados apresentados na tabela 3.4 relativamente ao aumento da taxa de palavras erradas calculada sobre os conjuntos de treino e teste.

	$F_0 < 160$ (Hz)	$F_0 \geq 160$ (Hz)
Conjunto de treino	$8\%_{rel}$	$3\%_{rel}$
Conjunto de teste	$111\%_{rel}$	$7\%_{rel}$

Tab. 3.4 - Variação (aumento) relativa do resultado WER_{LFCC} / WER_{MFCC}.

O crescimento do erro, do sistema baseado nos MFCC para o baseado nos LFCC, é maior no grupo de falantes com F_0<160, sendo essa variação extremamente acentuada neste conjunto de teste particular.

[1] Note-se que o eixo das abcissas não está linearmente calibrado.

[2] Note-se também que dois desses falantes apresentam taxas de erro elevadas, e portanto poderiam em princípio ser mais afectados com a utilização das características globalmente piores.

3.5.2 Recombinação básica dos *streams*

Como um passo preparatório da experimentação do reconhecedor *multi-stream*, foi avaliado o desempenho do sistema recombinando as verosimilhanças correspondentes aos *streams* de características, com base na equação e 3.33, mas considerando o parâmetro ω idêntico para todas as classes acústicas e também independente do valor da frequência fundamental do sinal. A figura 3.20 mostra a taxa de palavras erradas em função do valor do parâmetro ω, arbitrado no intervalo [0 , 1] cujos limites correspondem respectivamente aos reconhecedores *single-stream* LFCC e MFCC.

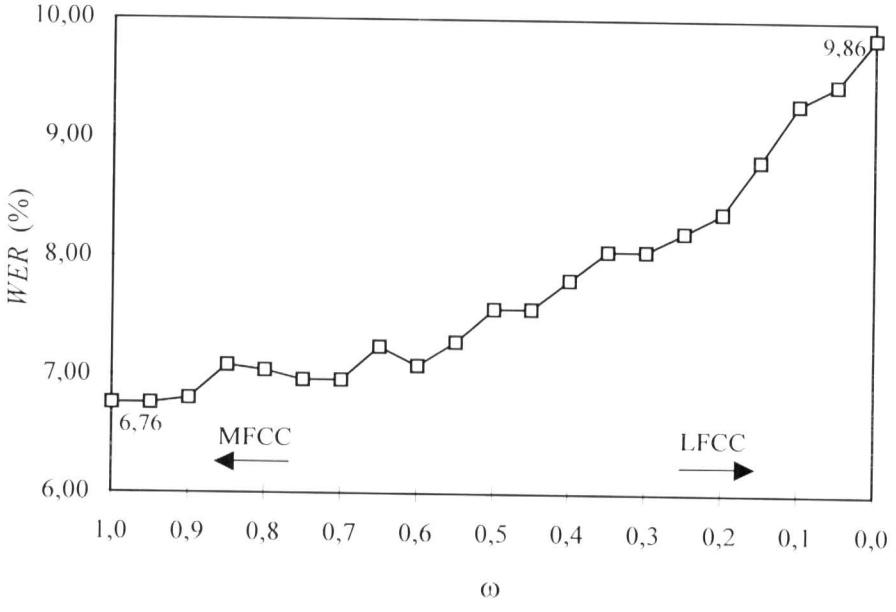

Fig. 3.20 - Taxa de erro em função do valor de ω arbitrado.

Filtrando a curva[2] verifica-se o crescimento monótono do erro com o aumento do peso do *stream* LFCC.

Repartindo agora todos os falantes do conjunto de teste nos dois grupos já referidos, obtêm-se os resultados apresentados na figura 3.21, com cada curva mostrando a taxa de palavras erradas no respectivo grupo medida relativamente ao número total de palavras, que é 2484. Refira-se que 997 palavras são pronunciadas pelo grupo de falantes F_0<160 e 1487 pelo outro.

[1] Note-se que o valor fronteira aqui considerado, 160 Hz, não foi escolhido com o propósito de "optimizar" estas conclusões particulares, o que não seria aceitável, é antes inspirado no exemplo do classificador de género baseado na frequência fundamental média (§3.3.4.2).
2 É suficiente um filtro de 4ª ordem.

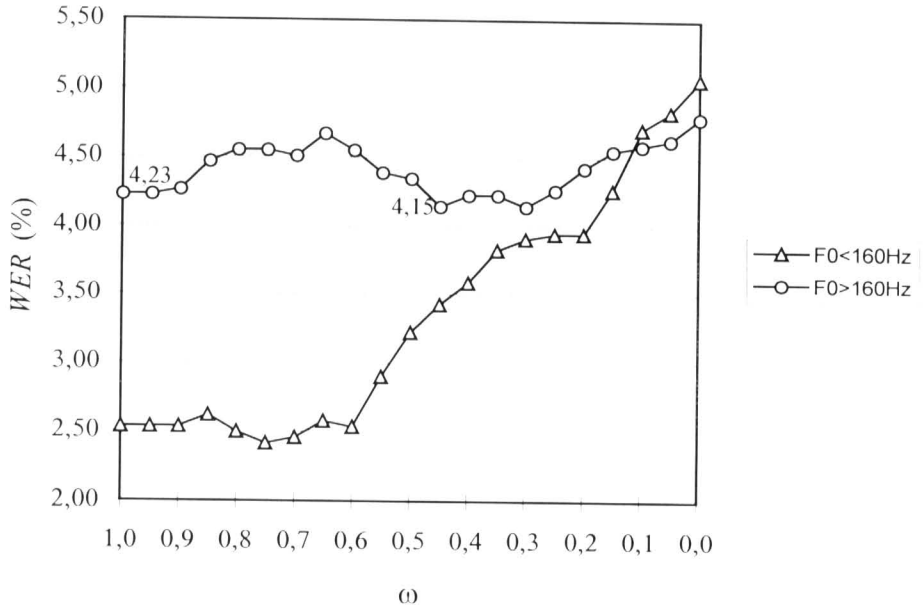

Fig. 3.21 - Taxas de erro, relativas a dois grupos de falantes, em função do valor de ω arbitrado.

Apesar do ruído inerente ao relativamente pequeno volume de dados, estes resultados, para além de concordarem com as análises já efectuadas, revelam alguns aspectos interessantes.

Em relação à curva correspondente ao grupo de falantes com F_0<160, verifica-se que para valores de ω superiores a aproximadamente 0,6 o erro se mantém praticamente constante, o que é um resultado aparentemente contraditório, sobretudo comparando com a outra curva. No entanto, em grande medida esse comportamento deve-se certamente à falta de aderência do *stream* LFCC aos respectivos modelos acústicos, ou seja, a verosimilhança conjunta da observação dos dois *streams* é fortemente dominada pelo modelo acústico MFCC. Assim, globalmente o efeito do parâmetro ω é semelhante ao do factor empírico α aplicado à log-verosimilhança e utilizado para estabelecer o peso relativo do modelo acústico (§2.9.3.4). Para valores de ω inferiores a 0,6, com a progressiva diminuição do peso das estimativas do modelo acústico MFCC, então tornam-se evidentes as deficiências do modelo acústico global baseado nas características LFCC. Deve ser notado que isso por si só não implica que para algumas classes acústicas os modelos LFCC não possam apresentar uma melhor capacidade discriminativa.

Em relação à outra curva, apesar de não se verificar uma redução significativa do erro, regista-se uma evolução claramente não monótona e com um mínimo local já próximo da configuração LFCC, o que sugere a existência de factores importantes actuando também no sentido de favorecer o desempenho do sistema baseado nesta representação.

3.5.3 Recombinação através do modelo definido

Atendendo à clara superioridade do desempenho do modelo acústico baseado nos MFCC, relativamente aos LFCC (§3.5.1), quando F_0<160, e também às aproximações introduzidas no procedimento de treino do modelo de recombinação das verosimilhanças (§3.4.6.3), nomeadamente adaptando os parâmetros apenas num único passo (3.50), o vector de parâmetros (3.47) é iniciado no ponto:

$$\theta_1^{(0)} \quad \text{com:} \quad \alpha_u^0 = 1, \quad u = 1,2,\ldots,52 \tag{3.54}$$

Apesar da diferença entre os modelos acústicos não ser tão acentuada quando F_0≥160, o vector de parâmetros (3.48) é iniciado no ponto:

$$\theta_2^{(0)} \quad \text{com:} \quad \beta_u^0 = 1 \quad \text{e} \quad \chi_u^0 = 0, \quad u = 1,2,\ldots,52 \tag{3.55}$$

resultando, portanto, os coeficientes de recombinação:

$$\omega\left(u, \overline{F}_0\right) = 1, \quad u = 1,\ 2,\ \ldots,\ 52, \quad \forall \overline{F}_0 \tag{3.56}$$

ou seja, de acordo com as equações 3.32, 3.33 e 3.34, com os vectores de parâmetros iniciais apenas o *stream* de características MFCC e o respectivo modelo acústico são utilizados.

No essencial, a tarefa que agora se coloca é analisar o desempenho do reconhecedor *multi-stream* quando ω se liberta da equação 3.56 em resultado da alteração dos vectores de parâmetros θ_1 e θ_2 segundo as direcções definidas pelos gradientes das respectivas funções objectivo, de acordo com a equação 3.50.

A figura 3.22 ilustra a contribuição dos segmentos da vogal /ae/, que ocorrem no conjunto de treino, para o cálculo das respectivas componentes dos gradientes, utilizadas posteriormente nas equações 3.51 a 3.53. Note-se que esta segmentação é obtida com o sistema configurado de acordo com a equação 3.56, isto é, considerando apenas o *stream* MFCC. Para cada ponto, (F_0, γ), em abcissas é marcada a estimativa, fornecida pelo módulo de análise do sinal (§3.3.4.2), do valor da frequência fundamental associado ao vector de características. Em ordenadas é assinalada a diferença normalizada entre as log-verosimilhanças MFCC e LFCC da observação do vector de características, de acordo com a equação 3.42.

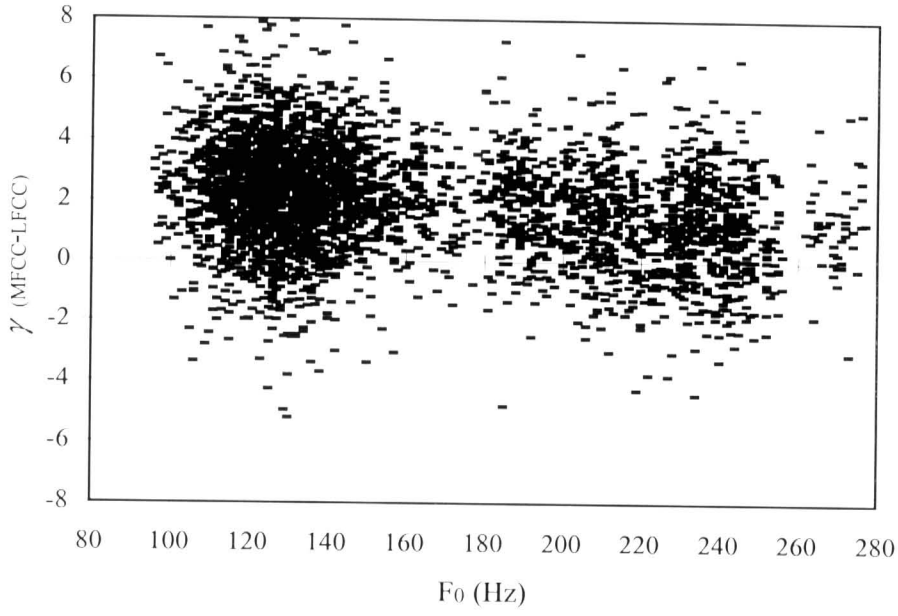

Fig. 3.22 - Distribuição dos pontos (F_0, γ) correspondente à vogal /ae/.

A figura 3.23 é semelhante à anterior, mas em relação à consoante fricativa /v/.

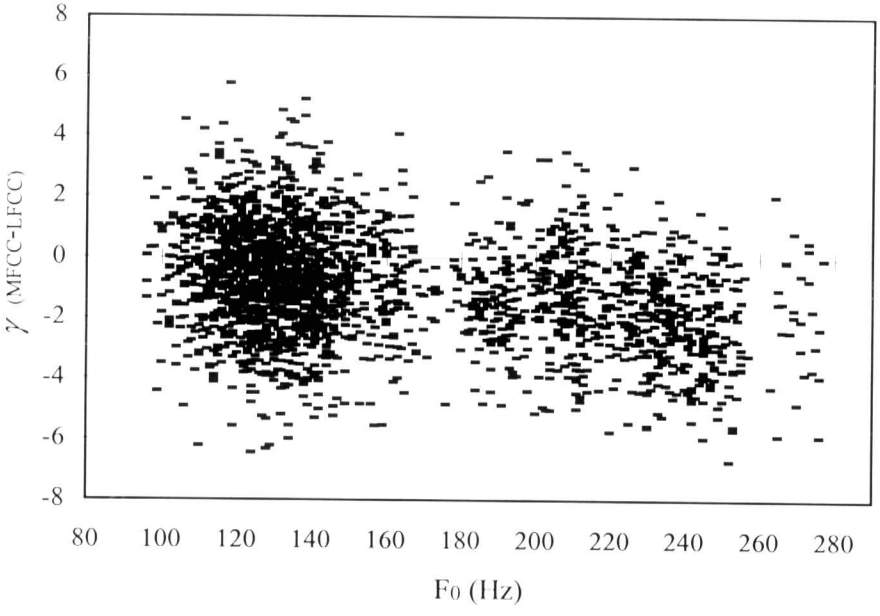

Fig. 3.23 - Distribuição dos pontos (F_0, γ) correspondente à consoante fricativa /v/.

Os gráficos 3.22 e 3.23 reflectem a existência dos dois grupos distintos de falantes já observados anteriormente, através da figura 3.5 (§3.3.4.2), no conjunto de treino.

Comparando os dois gráficos verifica-se de imediato a existência de uma relação significativa entre o valor médio de γ, estendido a todos os pontos das distribuições, e a identidade da classe acústica. Por exemplo, relativamente à vogal /ae/[1], em média a verosimilhança da observação correspondente ao *stream* MFCC é duas ordens de grandeza superior à do *stream* LFCC, tal como mostra a tabela 3.5. Também de acordo com a expectativa, uma tendência em sentido oposto verifica-se por exemplo para a consoante fricativa /v/.

Um aspecto muito importante evidenciado por cada gráfico é a diferença entre o valor médio de γ na distribuição de pontos com F_0 abaixo de, aproximadamente, 160 Hz e o mesmo valor médio calculado para os restantes pontos. Assim, e de acordo também com os resultados apresentados na tabela 3.5, para ambos os sons, a média da verosimilhança normalizada do *stream* MFCC menos a que corresponde ao *stream* LFCC é, quando o sinal apresenta $F_0<160$, uma ordem de grandeza maior relativamente à distribuição correspondente a $F_0\geq160$.

	/ae/	/v/
$\forall F_0$	1,97	-0,98
$F_0 < 160$ (Hz)	2,30	-0,63
$F_0 \geq 160$ (Hz)	1,41	-1,61

Tab. 3.5 -Valores médios de γ (equação 3.42) estimados sobre o conjunto de treino.

A figura 3.24 mostra os valores dos coeficientes de recombinação correspondentes a todas as unidades fonéticas definidas neste sistema e com o valor arbitrado de F_0 inferior a 160, calculados através das equações 3.34 e 3.50[2].

O gráfico confirma o aumento do peso, no modelo de recombinação, do *stream* de características MFCC na generalidade das vogais e a sua diminuição em grande parte das consoantes.

A figura 3.25 permite comparar os coeficientes de recombinação obtidos com o mesmo valor do passo na equação 3.50, mas correspondentes às duas partes diferentes da função 3.34, com o valor de F_0 arbitrado igual a 190 na parte $F_0\geq160$.

[1] A mesma situação verifica-se nas restantes vogais, assim como em outras classes de sons, como por exemplo as consoantes nasais.

[2] Com o passo λ_1 igual a 0,01.

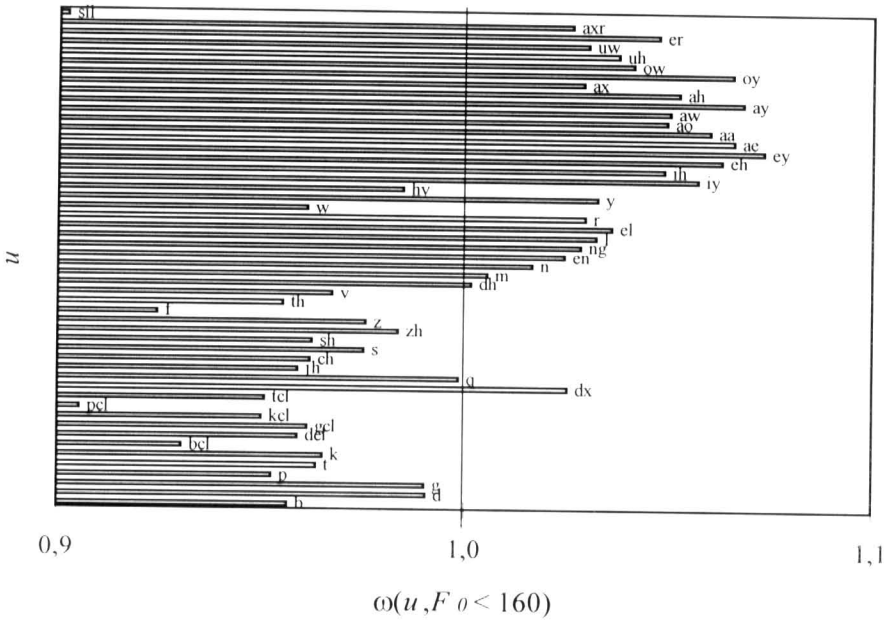

Fig. 3.24 - Coeficientes de recombinação estimados para as unidades fonéticas (ainda não limitados ao intervalo [0 , 1]).

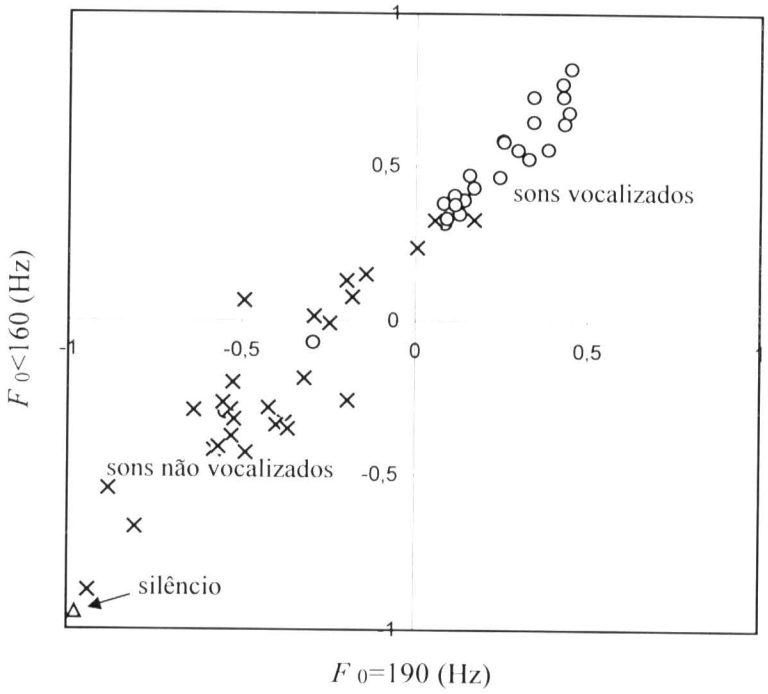

Fig. 3.25 - Coeficientes de recombinação, ω, para valores diferentes de F_0.

Todos os resultados da avaliação do desempenho do sistema que se apresentam são obtidos com o valor de ω automaticamente limitado ao intervalo [0 , 1], tendo-se

considerado não existir qualquer vantagem significativa em efectuar essa operação através de uma função limitadora com um comportamento suave, como por exemplo uma função log-sigmóide. Assim, através da figura 3.25 pode verificar-se que em quase todos os sons vocalizados (símbolo "o") o factor de recombinação é forçado a tomar o valor 1,0 qualquer que seja a estimativa do valor médio de F_0. Em relação aos sons não vocalizados (símbolo "**x**"), verifica-se que alguns deles são também forçados a serem apenas modelados pelo *stream* MFCC independentemente de F_0, mas com algumas consoantes isso apenas ocorre para valores de F_0 inferiores a 160.

De acordo com a formulação apresentada (§3.4.6.3), a optimização dos vectores de parâmetros θ_1 e θ_2 pode ser feita de maneira independente. Assim, mantendo θ_2 fixo no ponto inicial (3.55) e alterando apenas θ_1 (3.50), a taxa de palavras erradas avaliada sobre o conjunto de teste varia com o passo λ_1 de acordo com o gráfico apresentado na figura 3.26.

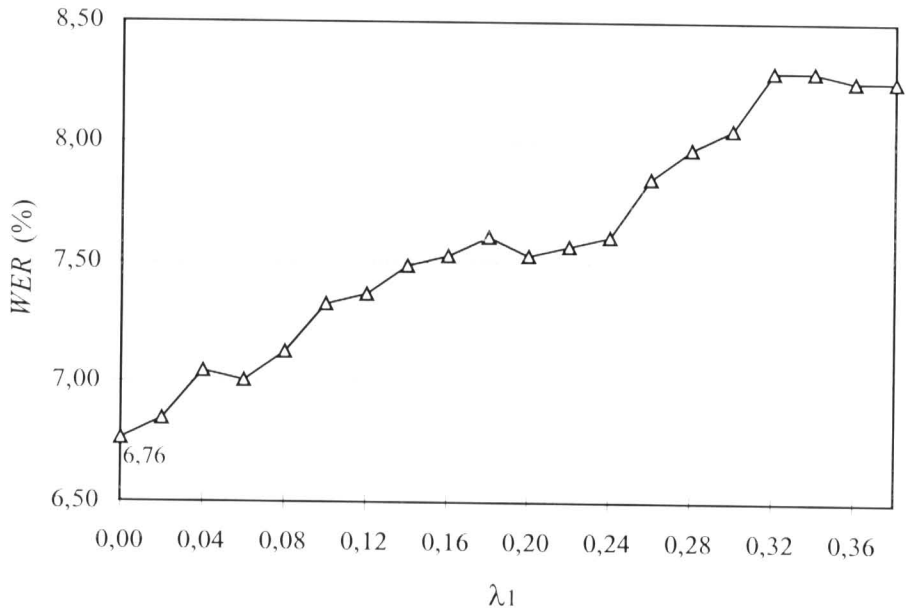

Fig. 3.26 - Taxa de palavras erradas para adaptações diferentes de θ_1 (e θ_2 inicial).

A curva apresenta alguns pontos mínimos locais mas verifica-se uma tendência[1] consistente e acentuada da degradação do desempenho do reconhecedor com o aumento do valor do parâmetro λ_1.

Comparando estes resultados com a curva F_0<160 da figura 3.21, uma vez que em relação ao outro grupo de falantes não se regista qualquer alteração significativa da taxa do erro, verifica-se que o seu aumento é agora imediato. Este facto não seria em princípio previsível uma vez que o *stream* LFCC é agora aplicado selectivamente às classes acústicas, adquirindo maior peso apenas quando os respectivos modelos

[1] Prolongando-se para os valores de λ_1 superiores ao fim de escala das abcissas deste gráfico.

apresentam maior verosimilhança de observação. Assim, a comparação das duas curvas evidencia a dificuldade em avaliar o mérito intrínseco de uma solução que está condicionada por formalismos eventualmente demasiado redutores. Por exemplo, neste caso concreto como o *stream* LFCC é dispensado na generalidade das vogais então o peso 1,0 é aí imposto, através das equações 3.32 e 3.33, ao *stream* MFCC. Assim, devido ao grande efeito da redução da taxa do erro com o factor empírico α inferior a 1,0 (§3.5.2), esta avaliação conduz necessariamente a piores resultados comparativamente aos da figura 3.21.

Em todo o caso, objectivamente conclui-se que o critério estabelecido para a adaptação do vector de parâmetros θ_1, fixando a partir do ponto inicial arbitrado a parte livre do modelo de recombinação formalizado, não permite a melhoria do desempenho do sistema sobre este conjunto de teste.

Existem, naturalmente, inúmeras abordagens alternativas que poderiam ter sido experimentadas. Por exemplo, em relação à forma funcional do modelo de recombinação, mesmo continuando a considerar o coeficiente ω invariante com o valor da frequência fundamental, para $F_0<160$, poder-se-ia estender a sua dependência ao nível dos estados, e não apenas dos modelos elementares como aqui é feito. Dada a evolução acústica do sinal nas unidades fonéticas, este modelo de recombinação poderia eventualmente conduzir a melhores resultados.

Também quanto ao critério de optimização em que se baseia a estimação dos parâmetros, poderiam ter sido tentadas outras abordagens. Optou-se por considerar uma aproximação à maximização da verosimilhança da observação conjunta dos *strems* (§3.4.6.3) baseado no princípio simples de, em cada estado, atribuir maior peso ao *stream* que tipicamente se apresenta mais verosímil no cômputo de todos os estados do modelo *composto*. No entanto, em princípio conseguir-se-iam obter melhores resultados utilizando um critério discriminativo que nas classes fonéticas mais próximas favorecesse, na medida adequada, o *stream* de características que conduzisse a uma maior separação acústica dos respectivos modelos *compostos*. Têm sido publicados resultados obtidos com diferentes critérios de adaptação dos parâmetros do modelo de recombinação baseados, por exemplo, na formulação directa do critério da máxima verosimilhança *a posteriori* [190], ou em medidas de Informação [176], ou usando estratégias de minimização do erro de classificação assentes em funções heurísticas com capacidade discriminativa [169,170,171,175,190].

Deve ser referido que foi ainda analisada a possibilidade de adaptar os parâmetros com base numa função objectivo similar às funções 3.43 e 3.45 mas aplicada sobre múltiplas hipóteses de reconhecimento[1] de maneira a introduzir alguma capacidade discriminativa no modelo através da contribuição negativa dos exemplos erradamente reconhecidos. Não se chegou a empreender esta abordagem em grande medida porque ela acarreta dificuldades operacionais substancialmente maiores que a

[1] Utilizando um programa, desenvolvido pelo autor para outra finalidade, assente no algoritmo *N-best* e no formalismo *multi-stream*.

solução aqui adoptada. Além disso, com a utilização de um critério discriminativo geralmente acentua-se a diferença de ganho no desempenho do reconhecedor quando é medido sobre os conjuntos de treino ou de teste, situação que se pode tornar particularmente aguda neste caso em que se regista uma clara desadaptação entre os modelos acústicos LFCC e este grupo de falantes do conjunto de teste.

A figura 3.27 mostra que a taxa do erro calculada independentemente para os 4 falantes do conjunto de teste cujo valor médio da frequência fundamental[1] é inferior a 160 apresenta uma tendência consistente de aumento em todos eles.

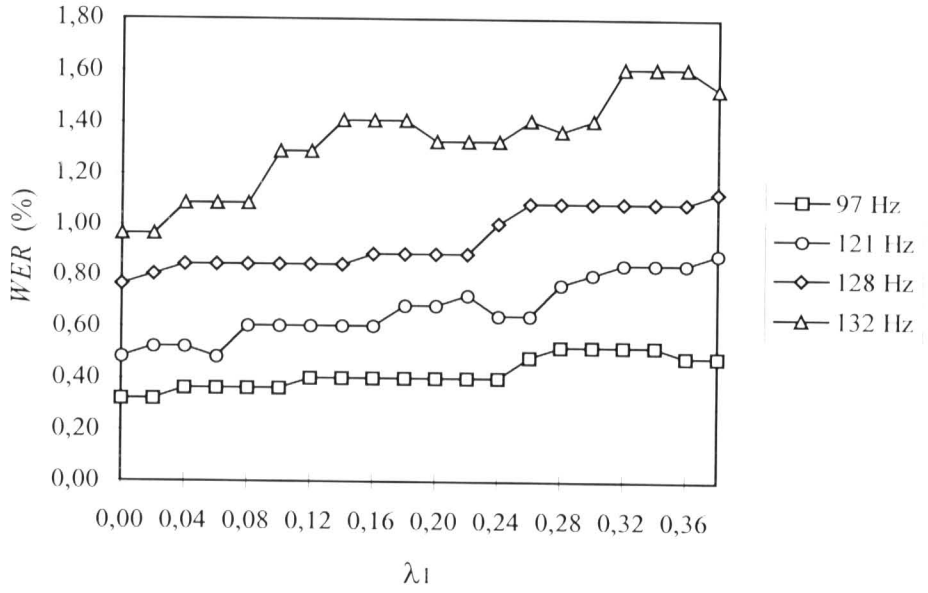

Fig. 3.27 - Taxa de palavras erradas para diferentes vectores θ_1, para o grupo dos falantes do conjunto de teste com F_0<160.

Não sendo aqui utilizado um conjunto de dados independente do conjunto de teste para fixar o parâmetro λ_1, e apesar do número muito limitado de falantes, a consistência destes resultados é importante para legitimar a fixação do valor desse parâmetro.

Assim, mantendo θ_1 fixo no ponto inicial (3.55) e alterando agora apenas θ_2 (3.50), a taxa de palavras erradas avaliada sobre o conjunto de teste varia com o passo λ_2 de acordo com o gráfico apresentado na figura 3.28.

[1] Estimado sobre todo o conjunto de treino.

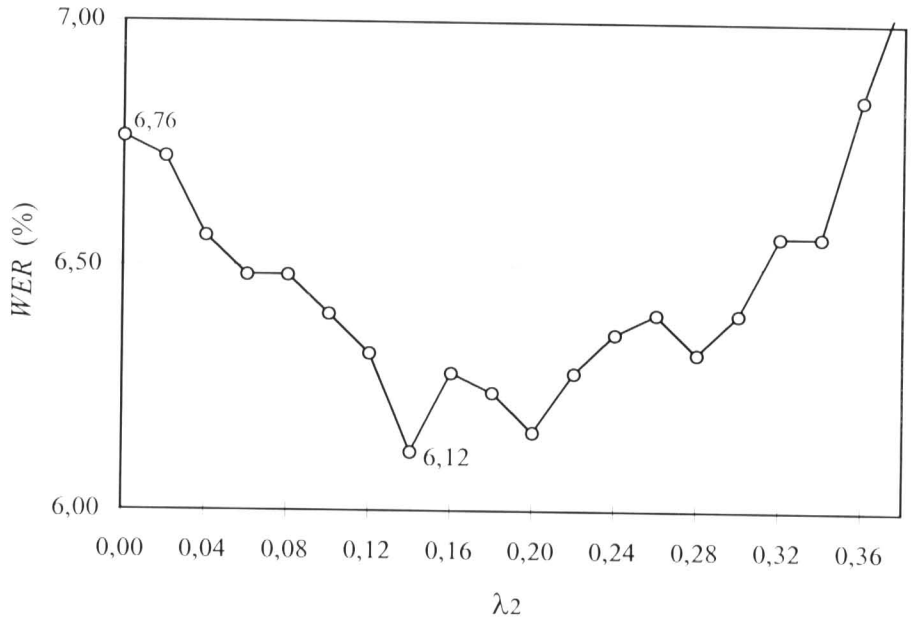

Fig. 3.28 - Taxa de palavras erradas para adaptações diferentes de θ_2 (e θ_1 inicial).

Partindo do passo nulo, o erro apresenta uma clara tendência de diminuição até se aproximar do valor 6,1, passando então a aumentar[1]. Tal como assinalado no gráfico, o valor mínimo da taxa do erro é 6,12%, o que corresponde a uma redução de 9,5%$_{rel}$.

Deve ser notado que a determinação do valor do parâmetro λ_2 num conjunto independente conduziria possivelmente a um pior resultado. Por outro lado, desde que não se verificassem factores de desadaptação acústica significativos entre os dois conjuntos de dados, essa eventual diferença deveria esbater-se com o aumento da dimensão de ambos.

A figura 3.29, apresentando a evolução da taxa do erro calculada independentemente para os 6 falantes do conjunto de teste com o valor médio da frequência fundamental superior a 160, mostra que em todos eles existe alguma gama de valores de λ_2 para a qual diminui o número de palavras erradas. Apesar do número de falantes e o volume de dados ser pequeno, é interessante confrontar este facto com a situação que se verifica no outro grupo de falantes.

[1] Prolongando-se para os valores de λ_2 superiores ao fim de escala das abcissas deste gráfico.

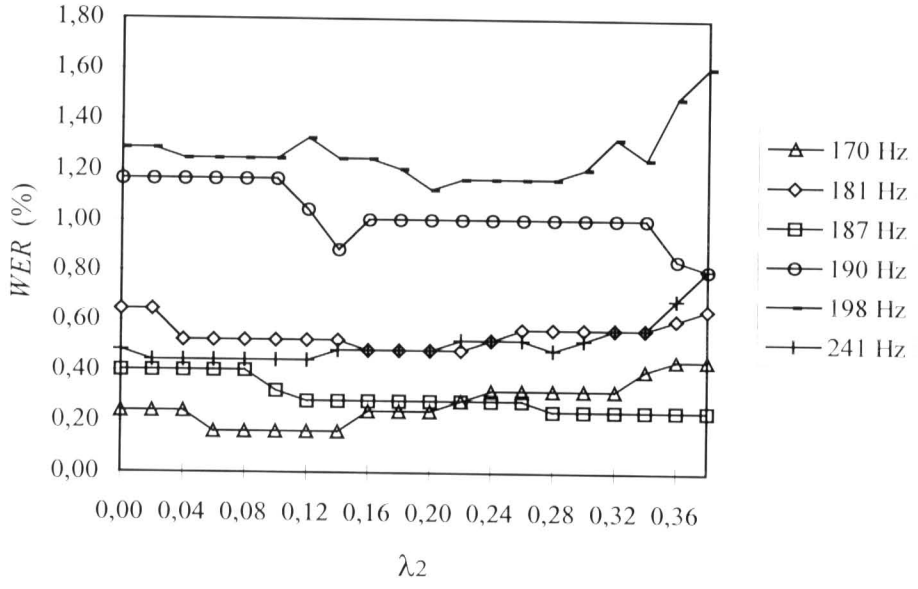

Fig. 3.29 - Taxa de palavras erradas para diferentes vectores θ_2, para o grupo dos falantes do conjunto de teste com F_0>160.

É importante referir que previamente a estes resultados havia sido experimentada uma forma funcional única para o modelo de recombinação, baseada na função 3.31 para todos os valores de F_0. Embora essa seja a função que é neste modelo também utilizada para $F_0 \geq 160$, a adaptação dos parâmetros é agora feita com base nas equações 3.52 e 3.53, enquanto que anteriormente se somavam também as contribuições dos vectores correspondentes a F_0<160. O melhor resultado então obtido foi 6,28%.

Todos os resultados apresentados nesta secção são obtidos com a conexão dos modelos *compostos* elementares no interior de cada palavra, e também dos estados no interior dos modelos, correspondente à topolologia (A) apresentada na figura 3.18 (§3.4.7). A figura 3.30 mostra os resultados obtidos para essa topologia, e também para as referenciadas por (B), (C) e (D), que correspondem a graus inferiores de relaxação de sincronismo entre os *streams*.

Neste caso particular os resultados são melhores quando a relaxação do sincronismo entre os *streams* é maior. Embora não tenha sido feito um estudo exaustivo, nas experiências onde se compararam as diferentes topologias propostas verificou-se geralmente também um melhor desempenho, embora por vezes pouco significativo, da solução (A). Verificou-se também que a solução (D), que força o alinhamento dos *streams* na estrutura interna do modelo *composto*, conduz em geral a resultados piores. Este resultado parece confirmar a importância do facto, muitas vezes assinalado, dos períodos de estacionaridade dos vários *streams* não serem necessariamente simultâneos. Foi ainda ensaiado um estudo sobre a possibilidade de

fazer depender, da identidade das unidades de fala, o nível de sincronismo imposto [184], não tendo no entanto sido obtidos resultados conclusivos.

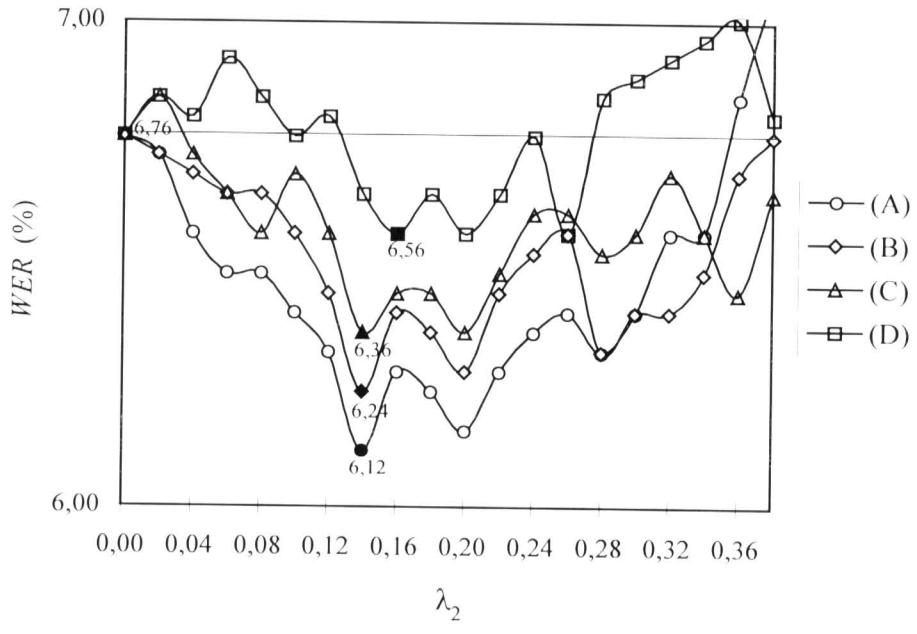

Fig. 3.30 - Resultados correspondentes aos da figura 3.28, mas permitindo vários graus de assincronismo entre os *streams*.

A figura 3.31 apresenta uma perspectiva da redução da taxa do erro obtida pelo sistema *multi-stream*, comparando-a com os resultados do reconhecedor *single-stream* baseado nos MFCC e com a dimensão do *codebook* progressivamente maior.

A redução do erro é menor à medida que a dimensão do sistema *single-stream* aumenta, sendo insignificante quando o *codebook* contém 8K gaussenas, o que evidencia a impossibilidade de melhorar o desempenho continuando apenas a duplicar o *codebook*. Existiriam inúmeras estratégias alternativas, por exemplo uma solução eventualmente capaz de reduzir significativamente a taxa do erro utilizando apenas os MFCC, inspirada no classificador de género (§3.3.4.2), consistiria em treinar e utilizar modelos acústicos diferentes para os falantes em função da estimativa do respectivo valor médio de F_0. Em todo o caso, o aumento, a partir de determinado limiar, do número de parâmetros relativamente ao volume de dados de treino reduziria a capacidade de generalização do sistema.

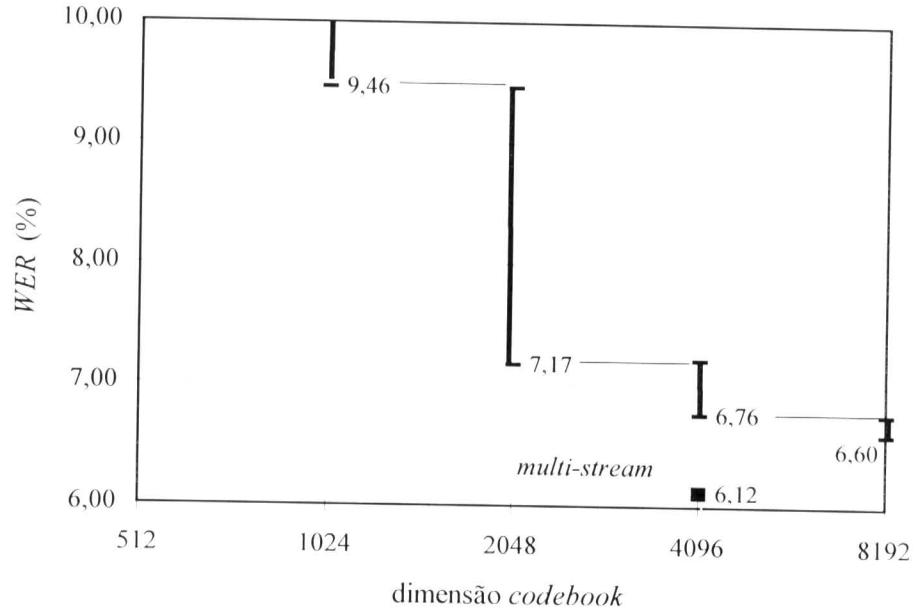

Fig. 3.31 - Evolução do erro nos sistemas *single-stream* baseados nas características MFCC e no sistema *multi-stream*.

Pelo contrário, a estratégia subjacente ao reconhecedor *multi-stream* não exige o aumento do número de parâmetros relativamente ao volume de vectores acústicos de treino. A redução significativa da taxa do erro, tal como mostra a figura 3.31, é antes alcançada através da maior capacidade em extrair e utilizar a informação discriminativa presente no sinal.

Tal como nos reconhecedores *single-stream*, neste sistema foram também simultaneamente optimizados[1] os parâmetros *WP* e α, ambos utilizados no algoritmo de descodificação (§2.9.3.4), tendo-se obtido os valores, respectivamente, 0,50 e 0,60. Considerando também os passos λ_1=0 e λ_2= 0,14, foi alcançada uma significativa redução da taxa de palavras erradas, de aproximadamente 7,2%$_{rel}$, para o valor 5,68%. A tabela 3.6 mostra as diferentes medidas (§2.9.3.4) do desempenho do reconhecedor *multi-stream*, antes e depois de optimizar os parâmetros *WP* e α.

	WER	*WER_Sub*	*WER_Ins*	*WER_Sup*	*SA*
WP=0.00; α=1.00	6,12%	4,23%	0,92%	0,97%	70,67%
WP=0.50; α=0.60	5,68%	3,99%	0,68%	1,01%	73,33%

Tab. 3.6 - Desempenho do sistema *multi-stream* antes e após a optimização de *WP* e α.

[1] Este processo de optimização no reconhecedor *multi-stream* foi substancialmente menos exaustivo do que o levado a cabo relativamente ao sistema *single-stream* baseado nos MFCC.

As tabelas 3.7 e 3.8 facilitam a confrontação dos resultados obtidos com os diversos sistemas desenvolvidos neste trabalho.

Tomando como referência o sistema *single-stream* baseado nas características MFCC, verifica-se que utilizando as características LFCC num sistema *single-stream* a taxa de erro aumenta 45,9% $_{rel}$, mas integrando essas características num sistema *multi-stream* o erro é reduzido em 9,5% $_{rel}$.

Em relação aos sistemas com os parâmetros WP e α já optimizados, verifica-se que em circunstâncias similares a redução da taxa do erro é ainda maior, aproximadamente 12,3% $_{rel}$.

	WER (%)	ε_{WER} (%$_{rel}$)
single-stream MFCC	**6,76**	0,0
single-stream LFCC	9,86	+45,9
multi-stream (MFCC+LFCC)	**6,12**	-9,5

Tab. 3.7 - Desempenho dos reconhecedores *single-* e *multi-stream* (WP e α não optimizados).

	WER (%)	ε_{WER} (%$_{rel}$)
single-stream MFCC	**6,48**	0,0
single-stream LFCC	9,42	+45,3
multi-stream	**5,68**	-12,3

Tab. 3.8 - Desempenho dos reconhecedores *single-* e *multi-stream*, com WP e α optimizados.

Capítulo 4

Conclusão

Neste capítulo são destacadas as principais conclusões deste trabalho e são apresentadas algumas hipóteses de desenvolvimento futuro.

4.1 Conclusões

A principal conclusão resultante deste trabalho é que num reconhecedor automático da fala a utilização simultânea de múltiplas representações do sinal, que encerram informação discriminante diferente embora possivelmente apresentem uma substancial dependência estatística, pode reduzir significativamente o valor esperado do erro de reconhecimento.

De acordo com a abordagem proposta neste trabalho:

- o processo integrado e global de decisão, conducente à hipótese de reconhecimento mais provável, baseia-se, ao nível acústico, num modelo que estima a verosimilhança da observação conjunta dessas representações combinando linearmente as respectivas verosimilhanças, calculadas com distribuições probabilísticas independentes;

- durante o processo de reconhecimento, o modelo de recombinação das verosimilhanças atribui localmente um peso a cada representação, através de coeficientes de recombinação dinâmicos, em função de informação associada à hipótese local de reconhecimento e adaptando-se também a determinadas características extraídas do sinal;

- a forma funcional que suporta os coeficientes de recombinação é definida a partir de conhecimento apriorístico e informação extraída de um conjunto de dados acústicos, com base no qual são também estimados os parâmetros livres através de um procedimento de treino baseado num critério de maximização da verosimilhança.

Os resultados experimentais são obtidos numa tarefa de reconhecimento de fala contínua independente do falante e com um vocabulário de aproximadamente 1000 palavras. O desenvolvimento dos sistemas é suportado pela base de dados *Resource Managment* e a sua avaliação é feita usando o conjunto de teste *February'89*.

O reconhecedor *single-stream* desenvolvido para servir de referência a este trabalho baseia-se em unidades elementares definidas ao nível fonético e modeladas sem contexto através de modelos escondidos de Markov do tipo semi-contínuo, treinados segundo uma aproximação ao critério da máxima verosimilhança. O reconhecimento assenta no algoritmo de descodificação Viterbi e utiliza uma gramática *word-pair* que reduz a perplexidade da tarefa ao valor aproximado de 60. A taxa de palavras erradas alcançada com a representação acústica baseada nos coeficientes mel-cepstrais é 6,48% e substituindo estas características pelos coeficientes lineares-cepstrais é 9,42%.

O reconhecedor *multi-stream* construído baseia-se nestas duas representações do sinal. O modelo linguístico é idêntico ao do sistema *single-stream*. Ao nível acústico são utilizados modelos *compostos* baseados na mesma tecnologia e nos mesmos parâmetros dos respectivos modelos elementares do reconhecedor *single-stream*. Utilizando coeficientes de recombinação constantes e comuns a todas as classes de sons durante o processo de descodificação, confirmou-se a expectativa da taxa do erro ser superior, de modo consistente, ao valor 6,48%.

Os coeficientes dinâmicos de recombinação das verosimilhanças propostos neste trabalho baseiam-se na dependência verificada entre o desempenho relativo destas duas representações e:

- a classe fonética hipoteticamente associada ao segmento local do sinal, de acordo com conhecimento apriorístico;
- a estimativa do valor médio da frequência de excitação fundamental, de acordo com a evidência empírica sobre o conjunto de dados de treino.

Treinando os parâmetros da forma funcional arbitrada para modelar estas dependências segundo uma aproximação ao critério da maximização da verosimilhança da observação conjunta estendida ao conjunto de treino, a avaliação do desempenho deste reconhecedor, efectuada em condições idênticas às dos sistemas anteriores, conduziu ao valor mínimo do erro igual a 5,68%.

Sendo este resultado obtido com representações acústicas significativamente correlacionadas, é legítimo concluir que a utilização de múltiplas representações do sinal é uma abordagem potencialmente muito eficaz na obtenção de reconhecedores de elevado desempenho.

Embora não tenha sido efectuado um estudo exaustivo sobre o efeito do nível de sincronismo imposto aos *streams* na estrutura interna dos modelos compostos, com os resultados experimentais obtidos neste trabalho verificou-se uma tendência consistente

no sentido do melhor desempenho do sistema quando se relaxa esta restrição. Essa tendência foi também verificada estendendo essa relaxação à conexão entre os modelos acústicos no interior de cada palavra.

O facto deste sistema ter alcançado resultados claramente superiores aos obtidos por um reconhecedor convencional numa tarefa de reconhecimento em que não se verificam problemas específicos para os quais a utilização de múltiplas representações tem sido sobretudo explorada, por ser particularmente mais vantajosa, pode constituir também um incentivo para uma aplicação mais abrangente destas soluções.

4.2 Desenvolvimento do trabalho

Sendo a tecnologia *multi-stream* relativamente recente e bastante promissora, existe um grande número de questões que requerem investigação. De entre as que emergiram naturalmente ao longo deste trabalho, são aqui destacadas algumas possíveis candidatas a investigação.

Atendendo à natureza deste trabalho, em que o objectivo principal não coincide necessariamente com a optimização de todas as partes de um determinado sistema, sendo até eventualmente vantajoso considerar algumas configurações mais desfavoráveis e assim testar com maior severidade as soluções propostas, nos protótipos desenvolvidos combinam-se representações do sinal acústico que apresentam uma substancial correlação. No entanto, no desenvolvimento de um sistema *multi-stream* definitivo para uma determinada aplicação, a selecção das representações do sinal mais adequadas para o efeito é obviamente uma questão muito importante. Em alternativa à comparação de diferentes combinações de características através da avaliação experimental do seu desempenho no reconhecedor *multi-stream*, um processo que acarreta um elevado custo para treinar os modelos acústicos, seria de grande utilidade poder desenvolver uma metodologia, assente por exemplo na estimação de medidas de Informação sobre um determinado conjunto de dados acústicos representativo da aplicação, que permitisse prever o desempenho dessas combinações.

Admitindo um modelo acústico assente na recombinação linear das verosimilhanças, tal como no sistema desenvolvido neste trabalho, existem alguns aspectos que merecem ser investigados.

Um deles prende-se com o método utilizado na definição da forma funcional que suporta os coeficientes dinâmicos de recombinação das verosimilhanças propostos neste trabalho. O arbítrio dessa função, com base apenas em alguns pressupostos sobre os factores que afectam o desempenho relativo dos *streams*, permitiu reduzir significativamente o erro sem exigir grande esforço. No entanto, em princípio o entendimento mais profundo dos mecanismos envolvidos na relação que se pretende modelar, e o posterior estabelecimento de uma forma funcional adequada, conduziria a

melhores resultados. Por exemplo, essa análise poderia eventualmente conduzir à conclusão de ser vantajoso estender a dependência dos coeficientes de recombinação à identidade dos estados dos modelos correspondentes a algumas classes fonéticas.

Outro aspecto também importante reside no procedimento utilizado na determinação dos parâmetros da função que modela os coeficientes de recombinação. Treinando os parâmetros com base num critério discriminativo, pelo menos em teoria seria possível obter uma maior separação dos modelos correspondentes às classes fonéticas acusticamente mais próximas, aliviando o problema da incorrecção do modelo paramétrico. Foi ensaiada uma abordagem nesse sentido, através de uma função objectivo definida sobre múltiplas hipóteses de reconhecimento, geradas por um descodificador construído com base no algoritmo *N-best* e no formalismo *multi-stream*, de maneira a considerar também a contribuição negativa dos exemplos erradamente reconhecidos.

A estimativa do valor médio da frequência fundamental, a partir dos valores conhecidos nos segmentos correspondentes aos sons vocálicos, é uma questão que merece também ser aprofundada. Um hipótese a investigar seria a possibilidade de melhorar o desempenho do sistema estimando o valor médio de F_0 sobre um intervalo de tempo mais pequeno, relativamente à janela com a duração fixa de 5 segundos utilizada no sistema desenvolvido neste trabalho. Desde logo, essa solução apresentaria uma vantagem imediata ao evitar o significativo atraso no processo de reconhecimento.

Apesar de não serem conhecidos pelo autor resultados publicados que atestem uma vantagem prática consistente dos modelos de recombinação não lineares, utilizados em alguns sistemas, essa poderá ser também uma linha de investigação interessante. À semelhança do que tem sido feito em alguns trabalhos publicados, os MLP poderão ser utilizados como suporte, dada a capacidade de estabelecerem qualquer transformação não linear da entrada e dispensarem grande parte das assunções sobre diferentes componentes do modelo, que em outras abordagens tecnológicas é necessário formular previamente.

Foi iniciado um estudo, que se pretende prosseguir pois não foram ainda alcançados resultados conclusivos, em que se estabelece a dependência da relaxação do sincronismo entre as múltiplas representações, quer na estrutura interna dos modelos *compostos* quer na sua interligação, relativamente à identidade das unidades de fala.

Uma consequência negativa da utilização de múltiplas representações do sinal acústico é o aumento substancial da complexidade computacional do sistema. Apesar de nas experiências realizadas neste trabalho se utilizarem apenas dois *streams* de características, verificou-se já um aumento em cerca de três vezes no tempo de operação relativamente ao reconhecedor *single-stream*. Um número maior de *streams* pode inviabilizar a utilização desta tecnologia em aplicações em tempo real através dos computadores de uso geral disponíveis actualmente. A utilização de *hardware* dedicado para a aceleração de determinadas operações matemáticas, identificadas como particularmente exigentes, pode constituir uma solução para este problema. Com

base em alguma experiência adquirida no desenvolvimento de uma unidade de processamento dedicada à operação de quantificação vectorial [94] num protótipo de reconhecimento baseado nos HMM discretos, existe a intenção de experimentar abordagens do mesmo tipo em partes do processo de reconhecimento com grande peso computacional, nomeadamente no cálculo das funções de Gauss. A tecnologia microelectrónica directamente envolvida nestes módulos baseia-se numa família de circuitos designada *Field Programmable Gate Array* (FPGA), que consiste numa plataforma física reconfigurável apresentando uma estrutura bi-dimensional regular, composta por blocos lógicos e as respectivas vias de interligação [93]. Os FPGA apresentam algumas qualidades muito interessantes na perspectiva de utilização que aqui se propõe, para além de permitirem acelerar efectivamente os cálculos, sobretudo valendo-se do seu elevado potencial para o processamento paralelo - note-se que as funções de Gauss associados às diferentes representações do sinal podem ser estimadas em simultâneo -. Permitem um rápido ciclo de desenvolvimento, com base em poderosas ferramentas de CAD dotadas de bibliotecas bastante completas com blocos funcionais pré-definidos. É possível definir novas funções utilizando uma linguagem de descrição de *hardware* de alto nível a partir das respectivas especificações funcionais. Feita a simulação do sistema, é tarefa simples programar a plataforma física, ou até mesmo reconfigurá-la dinamicamente em tempo real. O custo e também o factor risco associado ao projecto destes sistemas é claramente inferior ao que se verifica com os circuitos integrados fabricados expressamente para uma determinada aplicação.

Para terminar esta breve lista de hipóteses de trabalho relacionado com a tecnologia do reconhecimento da fala, refira-se que uma linha de investigação que se insinua como bastante atractiva, embora árdua, consiste na combinação de múltiplas representações do sinal com os chamados sistemas *multi-model*. Seria muito vantajoso, pelo menos em princípio, aliar a extracção de características com informação diferente do sinal com técnicas de modelação particularmente adequadas a cada uma dessas correntes de características.

BIBLIOTECA F.E. PORTO

Bibliografia

[1] W. W. Gibbs, "Taking computers to task", Scientific American, pp. 64-71, Julho 1997.

[2] F. T. Johansen, Global Discriminative Modeling for Automatic Speech Recognition, PhD Thesys, 1996.

[3] _, Automatic Speech Recognition, Advanced Topics, Kluwer Academic Publishers, 1996.

[4] E. Rich, Artificial Intelligence, McGraw-Hill Book Company, 1986.

[5] J.L. Flanagan, Speech Analysis, Synthesis and Perception, Springer-Verlag, 2ª edição, 1972.

[6] J. R. Deller, J. G. Proakis e J. H. Hansen, Discrete-time Processing of Speech Signals, Macmillan Publishing Company, 1993.

[7] H. Bourlard e N. Morgan, Connectionist Speech Recognition: A Hybrid Approach, Kluwer Academic Publishers, 1994.

[8] E. Keller, Fundamentals of Speech Synthesis and Speech Recognityion: Basic Concepts, State of the Art and Future Challenges, John Wiley & Sons, 1994.

[9] C. Rowden, Speech Processing, McGraw-Hill Book Company, 1994.

[10] A. Martinet, Elementos de Linguística Geral, Livraria Sá da Costa, 1992.

[11] R. J. Schalkoff, Pattern Recognition: Statistical, Structural and Neural Approaches, John Wiley & Sons, 1992.

[12] N. Cremelie e J-P. Martens, "In search of pronunciation rules", Workshop Modeling Pronunciation Variation for Automatic Speech Recognition, Rolduc, 1998.

[13] D. P. Morgan e C. L. Scofield, Neural Networks and Speech Processing, Kluwer Academic Publishers, 1991.

[14] D. Caragea, Curso de Português Básico para Estrangeiros, Publicações Universitárias Portuguesas L.da, 1995.

[15] S. Hayikin, Neural Networks, _, 1997.

[16] H. Ney, "Stochastic Grammars and Pattern Recognition", Speech Recognition and Understanding. Recent Advances, Springer-Verlag Berlin Heidelberg, 1992.

[17] T. Kuhn, H. Niemann e E. G. Schukat-Talamazzini, "Ergodic hidden Markov models and polygrams for language modeling", _.

[18] M. Alfonseca, J. Sancho e M. M. Orga, Teoria de Lenguajes, Gramaticas y Automatas, Ediciones Universidad y Cultura, 1990.

[19] H. Bourlard, "Towards increasing speech recognition error rates", Eurospeech '95, pp. 883-894, Setembro 1995.

[20] Y. K. Park, C. K. Un e O. W. Kwon, "Modeling acoustic transitions in speech by modified hidden Markov models with state duration and state duration-dependent observation probabilities", IEEE Transactions on Speech and Audio Processing, vol. 4, nº 5, Setembro 1996.

[21] J-F. Mari, D. Fohr e J-C. Junqua, "A second-order HMM for high performance word and phoneme-based continuous speech recognition", _, 1996.

[22] G. Qing, Z. Fang, W. Jian e W. Wenhu, "A new method used in HMM for modeling frame correlation", _, 1999.

[23] C-H. Lee, F. K. Soong e K. K. Paliwal, Automatic speech and speaker recognition, Kluwer Academic Publishers, 1996.

[24] C-H. Lee, L. R. Rabiner e R. Pieraccini, "Speaker independent continuous speech recognition using continuous density hidden Markov models", _.

[25] S. B. Davis e P. Mermelstein, "Comparasion of parametric representations for monosyllabic word recognition in continuously spoken sentences", IEEE Transactions on Acoustics, Speech and Signal Processing, vol. 28, nº 4, Agosto 1980.

[26] R. Kuc, Introduction to Digital Signal Processing, McGraw-Hill Book Company, 1988.

[27] F. Fallside, Computer Speech Processing, Prentice Hall International, _.

[28] O. Ghitza, "Auditory models and human performance in tasks related to speech coding and speech recognition", IEEE Transactions on Speech and Audio Processing, vol. 2, nº 1, pp. 115-132, 1994.

[29] C. L. Smith, C. P. Browman, R. S. McGowan e B. Kay, "Extracting dynamic parameters from speech movement data", Journal Acoustic American Society, pp. 1580-1588, Março, 1993.

[30] R. Drullman, J. M. Festen e R. Plomp, "Effect of temporal envelop smearing on speech reception", Journal Acoustic American Society, pp. 1053-1064, Fevereiro, 1994.

[31] S. S. Stevens e J. Volkmann, "The relation of pitch of frequency: a revised scale", American Journal of Psychology, vol. 53, pp. 329-353, 1940.

[32] S. Umesh, L. Cohen e D. Nelson, "Fitting the Mel Scale", IEEE International Conference on Acoustic, Speech and Signal Processing, Março, 1999.

[33] H. Wassner e G. Chollet, "New cepstral representation using wavelet analysis and spectral transformation for robust speech recognition", _.

[34] E. Zwicker, "Subdivision of the audible frequency range into critical bands", Journal of American Acoustic Society, vol. 33, 1961.

[35] A. Biem e S. Katagiri, "Filter bank design based on discriminative feature extraction", _, 1994.

[36] K. Demuynck, J. Duchateau, D. V. Compernolle e P. Wambacq, "Improved feature decorrelation for HMM-based speech recognition", _.

[37] T. Eisele, R. Haeb-Umbach e D. Langmann, "A comparative study of linear feature transformation techniques for automatic speech recognition", _.

[38] B. Milner, "Inclusion of temporal information into features for speech recognition", _.

[39] B. A. Hanson e T. H. Applebaum, "Robust speaker-independent word recognition using static, dynamic and acceleration features", IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 857-860, 1990.

[40] N. Kumar e A. G. Andreou, "Heteroscedastic discriminant analysis and reduced rank HMMs for improved speech recognition", Speech Comunication, vol. 26, pp. 283-297, 1998.

[41] B. P. Milner e S. V. Vaseghi, "Speech modelling using cepstral-time feature matrices and Hidden Markov Models", _, 1994.

[42] N. Morgan, S-L. Wu e H. Bourlard, "Digit recognition with stochastic perceptual models", Eurospeech'95, Setembro 1995.

[43] S. Furui, "On the role of spectral transition for speech perception", Journal of Acoustic American Society, vol. 80, nº 4, pp. 1016-1025, 1986.

[44] H. Hermansky, S. Greenberg e M. Pavel, "A brief (100-200 ms) history of time in feature extraction for speech analysis", 15th Annual Speech Research Symposium (Johns Hopkins University, Baltimore), Maio 1995.

[45] S. Dupont e H. Bourlard, "Using multiple time scales in a multi-stream speech recognition system", Eurospeech'97 , pp. 3-6, Rhodes, 1997.

[46] S. B. Davis e P. Mermelstein, "Comparasion of parametric representations for monosyllabic word recognition in continuously spoken sentences", IEEE Transactions on Acoustic Speech and Signal Processing, vol. 28, nº 4, Agosto 1980.

[47] J. B. Allen, "How do human process and recognize speech?", IEEE Transactions on Speech and Audio Processing, vol. 2, nº 4, pp. 567-577, 1994.

[48] O. Ghitza, "Auditory models and human performance in tasks related to speech coding and speech recognition", IEEE Transactions on Speech and Audio Processing, vol. 2, nº 1, pp. 115-132, 1994.

[49] M. Ostendorf, V. Digalakis e O. Kimball, "From HMMs to segmental models: a unified view of stochastic modelling for speech recognition", IEEE Transactions on Speech and Audio Processing, vol. 4 (5), pp. 360-378, 1996.

[50] H. Hermansky e N. Morgan, "RASTA processing of speech", IEEE Transactions on Speech and Audio Processing, vol. 2, nº 4, pp. 578-589, Outubro 1994.

[51] J. P. Openshaw e J. S. Mason, "On the limitations of cepstral features in noise", _, 1994.

[52] K-H. You e H-C. Wang, "Robust features for noisy speech recognition based on temporal trajectory filtering of short-time autocorrelation", Speech Commnication, nº 28, pp. 13-24, 1999.

[53] J. A. Flores e S. J. Young, "Continuous speech recognition in noise using spectral subtraction and HMM adaptation", _, 1994.

[54] J. Han, M. Han, G-B. Park, J. Park e W. Gao, "Relative Mel-frequency cepstral coefficienets compensation for robust telephone speech recognition", _.

[55] S. Nicholson, B. Milner e S. Cox, "Evaluating feature set performance using the F-ratio and J-measures", _.

[56] D. X. Sun, "Feauture dimension reduction using reduced-rank maximum likelihood estimation for hidden Markov models", _.

[57] F. T. Johansen e M. H. Johnsen, "Non-linear input transformation for discriminative HMms", International Conference on Acoustics, Speech, and Signal Processing, vol. I, pp. 225-228, Abil, 1994

[58] P. J. Price, W. Fisher, J. Bernstein e D. Pallet, "A database for continuous speech recognition in a 1000-word domain", IEEE International Conference on Acoustic, Speech and Signal Processing, Abril 1988.

[59] K-F. Lee, Automatic Speech Recognition: The Development of the SPHINX System, PhD Thesys Carnegie Mellon University, Kluwer Academic Publishers, 1989.

[60] A. H. Waibel, Prosody and speech recognition, PhD Thesys, Computer Science Department, Carnegie Mellon University, Outubro 1986.

[61] C. H. Lee e L. R. Rabiner, "A frame-synchronpus network search algorithm for connected word recognition", IEEE Transactions on Acoustics Speech and Signal Processing, vol. 37 (11), pp. 1649-1658, Novembro 1989.

[62] A. Falaschi, "Phonetic structure inference of phonemic HMM", Speech Recognition and Understanding: Recent Advances, Springer-Verlag Berlin Heidelberg, 1992.

[63] D. Jouvet, L. Mauuary e J. Monné, "Automatic adjustments of the Markov models topology for speech recognition applications over the telephone", Speech Recognition and Understanding: Recent Advances, Springer-Verlag Berlin Heidelberg, 1992.

[64] M. Blanchet e P. Lockwood, "An algorithm for the dynamic inference of hidden Markov models (DIHMM)", IEEE International Conference on Acoustic, Speech and Signal Processing, Minneapolis, vol. 2, pp. 251-254, 1993.

[65] S. E. Levinson, "Continuous variable duration hidden Markov models for automatic speech recognition", Computer Speech and Language, vol. 1(1), pp. 29-45, 1986.

[66] C. Mitchell, M. Harper e L. Jamieson, "On the complexity of explicit duration HMMs", IEEE Transactions on Speech and Audio Processing, vol. 3(3), pp. 213-217, 1995.

[67] T. H. Crystal e A. S. House, "Segmental durations on connected-speech signals: Current results", Journal of the Acoustical Society of America, vol. 83(4), pp. 1553-1573, Abril 1988.

[68] S. Young e G. Bloothooft, Corpus-Based Methods in Language and Speech Processing, Kluwer Academis Publishers.

[69] K-F. Lee, "Context-dependent phonetic hidden Markov models for speaker-independent continuous speech recognition", IEEE Transactions on Acoustics, Speech and Signal Processing, vol. 38, pp. 599-609, Abril 1990.

[70] R. Schwartz, O. Kimball, F. Kubala, M-W. Feng, Y-L. Chow, C. Barry e J. Makhoul, "Robust smoothing methods for discrete hidden Markov models", IEEE International Conference on Acoustic, Speech and Signal Processing, pp. 548-551, Maio 1989.

[71] Y. K. Park, C. K. Un e O. W. Kwon, "Modeling acoustic transitions is speech by modified hidden Markov models with state duration and state duration-dependent observation probabilities", IEEE Transactions on Speech and Audio Processing, vol. 4, nº 5, Setembro 1996.

[72] S. Wan, L. F. Bosh e L. C. Pols, "Integration of context-dependent durational knowledge into HMM-based speech recognition", _.

[73] W-G. Kim, J-Y. Choi e D. H. Youn, "HMM with global path constraint in Viterbi decoding for isolated word recognition", _.

[74] R. N. Sitaram e T. V. Sreenivas, "Phoneme recognition in continuous speech using large inhomogeneous hidden Markov models", _.

[75] J. Li e A. Marua, "A 2D extended HMM for speech recognition", _.

[76] J-F. Mari, D. Fohr e J-C. Junqua, "A second-order HMM for high performance word and phoneme-based continuous speech recognition", _.

[77] L. R. Bahl, P. V. Souza, P. S. Gopalakrishnan, D. Nahamoo e M. A. Picheny, "Robust methods for using context-dependent features and models in a continuous speech recognizer", IEEE International Conference on Acoustic, Speech and Signal Processing, vol. 1, pp. 533-536, 1994.

[78] M-Y. Hwang, X. Huang e F. Alleva, "Predicting unseen triphones with senones", IEEE International Conference on Acoustic, Speech and Signal Processing, vol. 2, pp. 311-314, 1993.

[79] R. Córdoba e J. M. Pardo, "Different strategies for distribution clustering using discrete, semicontinuous and continuos HMMs in CSR", _.

[80] L. R. Bahl, P. V. Souza, P. S. Gopalakrishnan, D. Nahamoo e M. A. Picheny, "Decision trees for phonological rules in continuous speech", IEEE International Conference on Acoustic, Speech and Signal Processing, 1991.

[81] R. K. Moore, M. J. Russell, P. Nowell, S. N. Downey e S. R. Browning, "A comparasion of phoneme decision tree (PDT) and context adaptative phone (CAP) based approaches to vocabulary-independent speech recognition", _.

[82] Y. Zhang, M. Alder e R. Togneri, "Using gaussian mixture modeling in speech recognition", _.

[83] B. Imperl, J. Kohler e Z. Kacic, "On the use of semi-continuous HMM for the isolated digits recognition over the telephone", _.

[84] R. Córdoba e J. M. Pardo, "Different strategies for distribution clustering using discrete, semicontinuous and continuous HMMs in CSR", _.

[85] K-F. Lee e F. Alleva, "Continuous speech recognition", _.

[86] X. D. Huang e M. A. Jack, "Semi-continuous hidden Markov models for speech recognition", Computer Speech and Language, vol. 3(3), pp-239-252, 1989.

[87] R.Schwartz e F. Kubala, "Hidden Markov models and speaker adaptation", Speech Recognition and Understanding: Recent Advances, Springer-Verlag Berlin Heidelberg, 1992.

[88] X. D. Huang, H. W. Hon, M. Y. Hwang e K-F. Lee, "A comparative study of discrete, semicontinuous and continuous HMMs", Computer Speech and Language, nº 7, pp. 359-368, 1993.

[89] X. D. Huang, K-F. Lee, H. W. Hon e M. Y. Hwang, "Improved acoustic modeling with the SPHINX speech recognition system", IEEE International Conference on Acoustic, Speech and Signal Processing, pp. 345-348, 1991.

[90] D. Pallet, J. G. Fiscus, W. M. Fisher e J. S. Garofolo, "1993 benchmark tests for the ARPA spoken language program", ARPA Workshop on Human Language Technology, Março 1994.

[91] C. Espain e V. Pera, "The generalized quickpropagation algorithm", International Conference Signal and Image Processing, Nassau, 1999.

[93] A. Araújo, V. Pera, C. Espain e J. S. Matos, "A vector quantizer architecture for an automatic speech recognizer", XIII Conference on Design of Circuits and Integrated Systems, Madrid, 1998.

[94] A. Araújo e V. Pera, "Vector quantizer acceleration for an automatic speech recognition application", International Conference in Speech and Language Processing, Sidney, 1998.

[95] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano e K. Lang, "Phoneme recognition using time-delay networks", IEEE International Conference on Acoustic, Speech and Signal Processing, New York, 1988.

[96] H. Bourlard e C. J. Wellekens, "Links between Markov models and multilayer perceptrons", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 12, nº 12, Dezembro 1990.

[97] M. Ostendorf, V. V. Digalakis e O. W. Kimball, "From HMM's to segmental models: a unified view os stochastic modeling for speech recognition", IEEE Transactions on Speech and Audio Processing, vol. 4, nº 5, Setembro 1996.

[98] V. V. Digalakis, M. Ostendorf, J. R. Rohlicek, "Fast algorithms for phone classification and recognition using segment-based models", IEEE Transactions on Signal Processing, vol. 40, nº 12, Dezembro 1992.

[99] F. Jelinek, Self-organizd language modelling for speech recognition, Readings in Speech Recognition, A. Waibel and K-F. Lee (eds.), pp. 450-503, 1990.

[100] R. Jacobs, "Increased rates of convergence through learning rate adaptation", Neural Networks, vol. 1, nº 4.

[101] L. B. Almeida, "A learning rule for asynchronous perceptrons with feedback in a combinatorial environment", IEEE First International Conference on Neural Networks, M. Caudill and C.Butler (eds.), vol. 2, pp. 609-618, _.

[102] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano e K. Lang, "Phoneme recognition using time-delay neural networks", IEEE Transactions on Acoustic, Speech and Signal Processing, vol. 37, nº 3, Março 1989.

[103] E. S. Morais, F. Violaro e M. A. Netto, "Modelo oculto de Markov parametrizado através de uma ANN e sua aplicação no reconhecimento da fala contínua", III Congresso Brasileiro de Redes Neuronais, Julho 1997.

[104] C. M. Bishop, Neural Networks for Pattern Recognition, Oxford University Press, 1995.

[105] Y. Bengio, Neural Networks for Speech and Sequence Recognition, International Thompson Computer Press, 1996.

[106] D. P. Morgan e C. L. Scofield, Neural Networks and Speech Processing, Kluwer Academic Publishers, 1991.

[107] D. Fohr, J-P. Haton e Y. Anglade, "Recognition of confusable words using HMM and MLP", Workshop Modeling Pronunciation Variation for Automatic Speech Recognition, Rolduc, 1998.

[108] J. E. Díaz-Verdejo, J. C. Segura-Luna, P. García-Teodoro e A. J. Rubio-Ayuso, "SLHMM: a continuous speech recognition system based on alphanet-HMM", _, 1994.

[109] A. Waibel, "Neural Network Approaches for Speech Recognition", _.

[110] F. Jelinek, R. L. Mercer e R. Salim, "Principles of Lexical Language Modeling for Speech Recognition", S. Furui and J.Sondhi, Eds., M. Dekker Publishers, New York, 1991.

[111] J. H. Wright, G. J. Jones e H. Lloyd-Thomas, "A robust language model incorporating a substring parser and extended N-grams", _, 1994.

[112] H. Ney, "Stochastic Grammars and Pattern Recognition", Speech Recognition and Understanding: Recent Advances, P. Laface and R. De Mori Editions, 1992.

[113] S. Okawa, T. Kobayashi e K. Shirai, "Automatic training of phoneme dictionary based on mutual information criterion", _, 1994.

[114] J. M. Kessens, M. Wester e H. Strik, "Improving the performance of a Dutch CSR by modeling pronunciation variation", Workshop Modeling Pronunciation Variation for Automatic Speech Recognition, Rolduc, 1998.

[115] S. T. Polzin e A. Waibel, "Pronunciation variations in emotional speech", Workshop Modeling Pronunciation Variation for Automatic Speech Recognition, Rolduc, 1998.

[116] T. Holter e T. Svendsen, "Maximum likelihood modelling of pronunciation variation", Workshop Modeling Pronunciation Variation for Automatic Speech Recognition, Rolduc, 1998.

[117] T. Fukada, T. Yoshimura e Y. Sagisake, "Automatic generation of multiple pronunciations based on neural networks and laguage statistics", Workshop Modeling Pronunciation Variation for Automatic Speech Recognition, Rolduc, 1998.

[118] B-H. Juang e L. R. Rabiner, Issues in using hidden Markov models for speech recognition, _.

[119] W. A. Woods, Language Processing for Speech Understanding", _.

[120] S. Sagayama, Hidden Markov Network for Precise and Robust Acoustic Modeling, Automatic Speech and Speaker Recognition, Kluwer Academic Publishers, _.

[121] P. Laface e L. Fissore, "Model topology selection for isolated word recognition", _, 1994.

[122] E. Dermatas e G. Kokkinakis, "Algorithm for clustering continuous density HMM by recognition error", IEEE Transactions on Speech and Audio Processing, vol. 4, nº 3, Maio 1996.

[123] A. M. Peinado, J. C. Segura, A. J. Rubio, P. García e J. L. Pérez, "Discriminative codebook design using multiple vector quantization in HMM-based speech recognizers", IEEE Transactions on Speech and Audio Processing, vol. 4, nº 2, Março 1996.

[124] V. V. Diglakis, P. Monaco e H. Murveit, "Genones: generalzed mixture tying in continuous hidden Markov model-basd speech recognizers", IEEE Transactions on Speech and Audio Processing, vol. 4, nº 4, Julho 1996.

[125] J. D. Martino, J. F. Mari, B. Mathieu, K. Perot e K. Smaili, "Which model for future speech recognition systems: hidden Markov models or finite-state automata?", _.

[126] T. Kobayashi, J. Furuyama e K. Masumitsu, "Partly hidden Markov model and its application to speech recognition", _, 1999.

[127] C. J. Wellekens, "Explicit correlation in hidden Markov model for speech recognition", IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 383-386, 1987.

[128] S. Takahashi, "Phoneme HMM's constrained by frame correlations", IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 219-222, 1993.

[129] J. A. Bilmes, "Data-driven extensions to HMM statistical dependencies", International Conference on Speech and Language Processing, Dezembro 1998.

[130] Z. Ghahramani e M. Jordan, "Factorial hidden Markov models", Machine Learning, 29, 1997.

[131] J. He e H. Leich, "Combining Stocastic trajectory model and discriminative feature in speech recognizer", 1994.

[132] L. Deng, "A generalized hidden Markov modl with state-conditioned trend function of time for the speech signal", Signal Processing, pp. 65-78, vol. 27, nº 1, Abril 1992.

[133] X. Wang, L. F. Bosh e L. C. Pols, "Integration of context-dependent durational knowledge into HMM-based speech recognition", _.

[134] C. Chesta e P. Laface, "Connected digit recognition using short and long duration models", _, 1999.

[135] L. R. Rabiner, S. E. Levinson e M. M. Sondhi, "On the application of vector quantization and hidden Markov models to speaker independent, isolated words recognition", Bell Systems Technical Journal, vol. 62, pp. 1075-1105, Abril 1983.

[136] R. Schwartz e S. Austin, "A comparasion of several approximate algorithms for finding multiple (N-best) sentence hypotheses", IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 701-704, 1992.

[137] Y. Chow, M. O. Dunham, O. A. Kimball, M. A. Krasner, G. F. Kubala, J. Makhoul, S. Roukos e R. M. Schwartz, "BYBLOS: The BBN continuous speech recognition system", IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 89-92, Abril 1987.

[138] L. R. Rabiner e B. H. Juang, "Hidden Markov models for speech recognition: strenghts and limitations", Speech Recognition and Understanding: Recent Advances, P. Laface and R. De Mori, 1992.

[139] P. Kenny, P. Labute, Z. Li, D. O'Shaughnessy, "New graph search techniques for speech recognition", _, 1994.

[140] L. R. Bahl, F. Jelinek e R. L. Mercer, "A maximum likelihood approach to continuous speech recognition", IEEE Transactions on Pattern Analysis and Machine Inteligence, vol. PAM I-5, nº 2, pp. 179-190, Março 1983.

[141] H. Sakoe e S. Chiba, "Dynamic programming optimization for spoken word recognition", IEEE Transactions on Acoustics, Speech and Signal Processing, vol. 26, pp. 52-59, 1978.

[142] F. A. Alleva, "Search organization for large vocabulary continuous speech recognition", Speech Recognition and Understanding: Recent Advances, Edited by Pietro Laface and R. De Mori, 1992.

[143] H. Ney, D. Mergel, A. Noll e A. Paeseler, "Data driven search organization for continuous speech recognition", IEEE Transactions on Signal Processing, vol. 40, nº 2, Fevereiro 1992.

[144] A. Sankar e C-H. Lee, "A maximum-likelihood approach to stochastic matching for robust speech recognition", IEEE Transactions on Speech and Audio Processing, vol. 4, nº 3, Maio 1996.

[145] D. S. Pallet, "Benchmark tests for DARPA Resource Managment database performance evaluations", _, 1989.

[146] Q. Huo, C. Chan e C-H. Lee, "On-line adaptation of the SCHMM parameters based on the segmental quasi-Bayes learning fo speech recognition", IEEE Transactions on Speech and Audio Processing, vol. 4, nº 2, Março 1996.

[147] S. M. Kay, Fundamentals of statistical signal processing - Estimation Theory, Prentice Hall International, Inc., 1993.

[148] A. Nádas, "A decision theoretic formulation of a training problem in speech recognition and a comparasion of training by unconditional versus conditional maximum likelihood", IEEE Transactions on Acoustics, Speech and Signal Processing, vol. 31, Agosto 1983.

[149] L. R. Bahl, P. F. Brown, P. V. Souza e R. L. Mercer, "Maximum mutual information estimation of hidden Markov model parameters for speech recognition", IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 49-52, 1986.

[150] J-L. Gauvain e C-H. Lee, "Maximum a posteriori estimation for multivariate gaussian mixture observations of Markov chains", IEEE Transactions on Speech and Audio Processing, vol. 2, pp. 291-298, 1994.

[151] Y. Ephraim, A. Dembo e L. R. Rabiner, "A minimum discrimination information approach for hidden Markov modeling", IEEE Transactions on Information Theory, vol. 35, nº 5, Setembro 1989.

[152] D. Povey e P. C. Woodland, "Frame discrimination training of HMMs for large vocabulary speech recognition", _, 1999.

[153] Q. Huo, C. Chan e C-H. Lee, "Bayesian learning of the SCHMM parameters for speech recognition", _, 1994.

[154] J-K. Chen e F-K. Soong, "Discriminative training of high performance speech recognizer using N best candidates", _, 1994.

[155] A. N. Rodríguez, J. B. Mariño e E. Monte, "An adaptive gradient-search based algorithm for discriminative training of HMM's", _.

[156] R. Schluter, W. Macherey, S. Kanthak, H. Ney e L. Welling, "Comparasion of optimization methods for discriminative training criteria", _.

[157] Y. Normandin, R. Cardin e R. De Mori, "High-performance connected digit recognition using maximum mutual information estimation", IEEE Transactions on Acoustics, Speech and Signal Processing, vol. 2, pp. 299-311, Abril 1994.

[158] A. Ljolje, Y. Ephraim e L. R. Rabiner, "Estimation of hidden Markov model parameters by minimizing empirical error rate", IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 709-712, 1990.

[159] G. Yuqing, E-E. Jan, M. Padmanabhan e M. Picheny, "HMM training based on quality measurement", _, IEEE, 1999.

[160] N. Merhav e Y. Ephraim, "Maximum likelihood hidden Markov modeling using a dominant sequence of states", IEEE Transactions on Signal Processing, vol. 39, nº 9, Setembro 1991.

[161] J. Ming, P. Hanna, D. Stewart, S. Vaseghi e F. J. Smith, "Capturing discriminative information using multiple modeling techniques", _.

[162] H. Bourlard e S. Dupont, "A new ASR approach based on idependent processing and recombination of partial frequency bands", International Conference on Speech and Language Processing, pp. 426-429, 1996.

[163] H. Bourlard, S. Dupont, H. Hermansky e N. Morgan, "Towards subband-based speech recognition", Eusipco'96, pp. 1579-1582, Trieste, 1996.

[164] S. Okawa, E. Bocchieri e A. Potamianos, "Multi-band speech recognition in noisy environments", IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 641-644, Seattle, Maio 1998.

[165] S.Tribewala e H. Hermansky, "Sub-band based recognition of noisy speech", IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 1255-1258, Munich, 1997.

[166] N. Suaudeau e R. André-Obrecht, "An efficient combination of acoustic and supra-segmental informations in a speech recognition system", _, 1994.

[167] P. Shengxi, L. Jia, J. Jintao, W. Zuoying e L. Dajin, "A novel robust speech recognition algorithm based on multi-models and integrated decision method", _.

[168] S-L. Wu, B. E. Kingsbury, N. Morgan e S. Greenberg, "Performance improvements through combining phone- and syllable-scale information in automatic speech recognition", International Conference on Speech and Language Processing, 1998.

[169] I. Rogina e A. Waibel, "Learning state-dependent stream weights for multi-codebook HMM speech recognition systems", _, 1994.

[170] S. M. Chu e Y. Zhao, "Robust speech recognition using discriminative stream weighting and parameter interpolation", _.

[171] C. Cerisara, J-P. Haton e D. Fohr, "Towards a global optimization scheme for multi-band speech recognition", Eurospeech'99, pp. 587-590, Budapest, 1999.

[172] K. Kirchhoff e J. A. Bilmes, "Dynamic classifier combination in hybrid speech recognition systems using utterance-level confidence values", _, 1999.

[173] P. Sivakumaran, A. M. Ariyaeeinia e J. A. Hewitt, "Sub-band based speaker verification using dynamic recognition weights", _.

[174] A.Morris, A. Hagen e H. Bourlard, "The full combination sub-bands approach to noise robust HMM/ANN based ASR", Eurospeech'99, pp. 599-602, Budapest, 1999.

[175] J-K. Chen e F.K. Soong, "Discriminative training of high performance speech recognizer using N best candidates", _, 1994.

[176] S. Okawa, T. Nakagima e K. Shirai, "A recombination strategy for multi-band speech recognition based on mutual information criterion", Eurospeech'99, pp. 603-606, Budapest, 1999.

[177] A. Nogueiras-Rodríguez, J. B. Mariño, E. Monte, "An adaptive gradient-search based algorithm for discriminative training of HMM's", _.

[178] R. Schluter, W. Macherey, S. Kanthac, H. Ney e L. Welling, "Comparasion of optimization methods for discriminative training criteria", _.

[179] M. Rahim, Y. Bengio e Y. Le-Cun, "Discriminative feature and model design for automatic speech recognition", _.

[180] J. Luettin e S. Dupont, "Continuous audio-visual speech recognition", Springer Verlag Heidelberg, 1998.

[181] S. Dupont, "Using the multi-stream approach for continuous audio-visual speech recognition: experiments on the M2VTS database", FPMS-TCTS technical report, Mons, July 1997.

[182] M. J. Tomlinson, M. J. Russel e N. M. Brooke, "Integrating audio and visual information to provide highly robust speech recognition", IEEE International Conference on Acoustics, Speech and Signal Processing, vol. 2, pp. 821-824, 1996.

[183] A. Janin, D. Ellis e N. Morgan, "Multi-stream speech recognition: ready for prime time?", Eurospeech'99, pp. 591-594, Budapest, 1999.

[184] N. Mirghafori e N. Morgan, "Sooner or later: exploring asynchrony in multi-band speech recognition", Eurospeech'99, pp. 591-594, Budapest, 1999.

[185] J. Ming, P. Hanna, D. Stewart, M. Owens e F. J. Smith, "Improving speech recognition performance by using multi-model approaches", _, 1999.

[186] C. J. Wellekens, J. Kangasharju, C. Milesi, "The use of meta-HMM in multistream HMM training for automatic speech recognition", International Conference on Speech and Language Processing, 1998.

[187] A. P. Varga e R. K. Moore, "Hidden Markov model decomposition of speech and noise", IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 845-848, 1990.

[188] M. J. Gales e S. J. Young, "Robust continuous speech recognitio using parallel model combination", IEEE Transactions on Speech and Audio Processing, vol. 4, nº 5, Setembro 1996.

[189] M. J. Tomlinson, M. J. Russel, R. K. Moore, A. P. Buckland e M. A. Fawley, "Modelling asynchrony in speech using elementary single-signal decomposition", IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 1247-1250, Munich, 1997.

[190] H. Bourlard, S. Dupont e C. Ris, "Multi-stream speech recognition", IDIAP Research Report, Mons, 1996.

[191] H. Bourlard e S. Dupont, "Subband-based speech recognition", IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 1251-1254, Munich, 1997.

[192] S. Dupont, H. Bourlard e C. Ris, "Robust speech recognition based on multi-stream features", Workshop on Robust Speech Recognition for Unknown communication channels, Pont-à-Monsson, 1997.

[193] P. C. Loizou e A. S. Spanias, "High-performance alphabet recognition", IEEE Transactions on Speech and Audio Processing, vol. 4, nº 6, Novembro 1996.

[194] L. V. Immerseel e J. P. Martens, "Pitch and voiced/unvoiced determination with an auditory model," Journal Acoustic Soc. Am., 91, pp. 3511-3526, 1992.

[195] Y. Sagisaka, N. Campbell e N. Higuchi, Computing Prosody: Computing Models for Processing Spontaneous Speech, Springer-Verlag, 1997.

[196] A. J. Araújo, Sistema de Apoio ao Projecto de Unidades de Processamento em Vírgula Flutuante: Aplicação ao Cálculo de Funções Transcendentes, Tese de Doutoramento, FEUP, 2000.

[197] C. Brezinski, Accélération de la convergence en analyse numérique, Lecture Notes in Mathematics, p.584, 1977, Springer-Verlag.

Anexo: Léxico

A	ax
A42128	q ey f ao r tcl t uw w ah n tcl t uw ey tcl
AAW	q ey q ey dcl d ah bcl b y uw
ABERDEEN	q ae bcl b axr dcl d iy n
ABOARD	ax bcl b ao r dcl
ABOVE	ax bcl b ah v
ADD	q ae dcl
ADDED	q ae dx ax dcl
ADDING	q ae dx ih ng
AFFECT	ax f eh kcl
AFTER	q ae f tcl t axr
AGAIN	ax gcl g eh n
AJAX	q ey dcl jh ae kcl k s
AJAX+S	q ey dcl jh ae kcl k s ax z
ALASKA	ax l ae s kcl k ax
ALERT	ax l er tcl
ALERTS	ax l er tcl t s
ALEXANDRIA	q ae l ax kcl z ae n d r iy ax
ALL	q ao l
AN	ax n
ANCHORAGE	q ae ng kcl k axr dcl jh
AND	q ae n
ANY	eh n iy
ANYBODY	q eh n iy bcl b ah dx iy
APALACHICOLA	q ae pcl p ax l ae tcl ch ax kcl k ow l ax
APALACHICOLA+S	q ae pcl p ax l ae tcl ch ax kcl k ow l ax z
APRIL	q ey pcl p r el
ARABIAN	ax r ey bcl b iy ax n
ARCTIC	q aa r tcl t ax kcl k
ARE	aa r
AREA	q eh r iy ax
AREAS	q eh r iy ax z
AREN+T	aa axr n
ARKANSAS	q aa r kcl k ax n s ao
ARKANSAS+S	q aa r kcl k ax n s ao z
AROUND	ax r aw n
ARRIVAL	ax r ay v el
ARRIVE	ax r ay v
ARRIVED	ax r ay v dcl
ARRIVING	ax r ay v ih ng
ARROW	q eh r ow
AS	ax z
ASTORIA	ax s tcl t ao r iy ax
ASUW	q ey eh s y uw dcl d ah bcl b y uw
ASW	q ey eh s dcl d ah bcl b y uw
AT	ae tcl
ATLANTIC	ax dcl l ae n ax kcl k
AUGUST	q ao gcl g ax s tcl
AVAILABLE	ax v ey l ax bcl b el
AVERAGE	q ae v axr dcl jh
BAD	bcl b ae dcl
BADGER	bcl b ae dcl jh axr
BADGER+S	bcl b ae dcl jh axr z
BAINBRIDGE	bcl b ey n bcl b r ih dcl jh

BAINBRIDGE+S	bcl b ey n bcl b r ih dcl jh ax z
BANGKOK	bcl b ae ng k aa kcl k
BARGE	bcl b aa r dcl jh
BASS	bcl b ae s
BAY	bcl b ey
BE	bcl b iy
BEAM	bcl b iy m
BEAMS	bcl b iy m z
BEEN	bcl b ih n
BEFORE	bcl b ax f axr
BELOW	bcl b ax l ow
BERING	bcl b eh r ih ng
BETTER	bcl b eh dx axr
BETWEEN	bcl b ax tcl t w iy n
BIDDLE	bcl b ih dx el
BIDDLE+S	bcl b ih dx el z
BISMARK	bcl b ih z m aa r kcl k
BOMBAY	bcl b aa m bcl b ey
BOTH	bcl b ow th
BOX	bcl b aa kcl k s
BRIGHT	bcl b r ay tcl
BRITISH	bcl b r ih dx ax sh
BROOKE	bcl b r uh kcl k
BROOKE+S	bcl b r uh kcl k s
BRUNSWICK	bcl b r ah n z w ax kcl k
BRUNSWICK+S	bcl b r ah n z w ax kcl k s
BUD-TEST	bcl b ah dcl tcl t eh s tcl
BUMP	bcl b ah m pcl p
BY	bcl b ay
C-CODE	s iy kcl k ow dcl
C-CODES	s iy kcl k ow dcl z
C-RATING	s iy r ey dx ih ng
C-RATINGS	s iy r ey dx ih ng z
C1	s iy w ah n
C2	s iy tcl t uw
C3	s iy th r iy
C4	s iy f ao r
C5	s iy f ay v
CALEDONIA	kcl k ae l ax dcl d ow n iy ax
CALIFORNIA	kcl k ae l ax f ao r n y ax
CAMDEN	kcl k ae m d ax n
CAMDEN+S	kcl k ae m d ax n z
CAMPBELL	kcl k ae m b el
CAMPBELL+S	kcl k ae m b el z
CAN	kcl k ax n
CANADA	kcl k ae n ax dcl d ax
CAPABILITIES	kcl k ey pcl p ax bcl b ih l ax dx iy z
CAPABILITY	kcl k ey pcl p ax bcl b ih l ax dx iy
CAPABLE	kcl k ey pcl p ax bcl b el
CAPACITIES	kcl k ax pcl p ae s ax dx iy z
CAPACITY	kcl k ax pcl p ae s ax dx iy
CARRIER	kcl k eh r iy axr
CARRIER+S	kcl k eh r iy axr z
CARRIERS	kcl k eh r iy axr z
CARRY	kcl k eh r iy
CASREP	kcl k ae s r eh pcl p
CASREPED	kcl k ae s r eh pcl tcl t
CASREPS	kcl k ae s r eh pcl p s
CASUALTY	kcl k ae zh uw ax l tcl t iy

CAT-2	kcl k ae tcl t uw
CAT-3	kcl k ae tcl t th r iy
CAT-4	kcl k ae tcl t f ao r
CATEGORIES	kcl k ae dx ax gcl g ao r iy z
CATEGORY	kcl k ae dx ax gcl g ao r iy
CENTER	s eh n tcl t axr
CENTERED	s eh n tcl t axr dcl
CENTERING	s eh n axr ih ng
CEP	s iy iy pcl p iy
CHANGE	tcl ch ey n jh
CHANGED	tcl ch ey n jh
CHANGING	tcl ch ae n el
CHART	tcl ch aa r tcl
CHARTS	tcl ch aa r tcl t s
CHATTAHOOCHEE	tcl ch ae dx ax hh uw tcl ch iy
CHATTAHOOCHEE+S	tcl ch ae dx ax hh uw tcl ch iy z
CHESHIRE	tcl ch eh sh axr
CHINA	tcl ch ay n ax
CHOP	tcl ch aa pcl p
CHOPPED	tcl ch aa pcl tcl t
CHOPPING	tcl ch aa pcl p ih ng
CIRCLE	s er kcl k el
CITRUS	s ih dx axr s
CITRUS+S	s ih dx axr s ax z
CITY	s ih dx iy
CLEAR	kcl k l iy r
CLEARED	kcl k l iy r dcl
CLEARING	kcl k l iy r ih ng
CLEVELAND	kcl k l iy v l ax n
CLEVELAND+S	kcl k l iy v l ax n z
CLOSE	kcl k l ow z
CLOSER	kcl k l ow s axr
CLOSEST	kcl k l ow s ax s tcl
CODAG	kcl k ow dcl d ae gcl
CODE	kcl k ow dcl
CODES	kcl k ow dcl z
COLOR	kcl k ah l axr
COMPARED	kcl k ax m pcl p eh r dcl
CONFIDENCE	kcl k aa n f ax dcl d ax n s
CONFIDENCE+S	kcl k aa n f ax dcl d ax n tcl s ax z
CONIFER	kcl k aa n ax f axr
CONIFER+S	kcl k aa n ax f axr z
CONQUEST	kcl k aa ng kcl k w eh s tcl
CONQUEST+S	kcl k aa ng kcl k w eh s tcl
CONSTANT	kcl k aa n s tcl t ax n
CONSTANT+S	kcl k aa n s tcl t ax n s
CONSTELLATION	kcl k aa n s tcl t ax l ey sh ax n
CONSTELLATION+S	kcl k aa n s tcl t ax l ey sh en z
CONVENTIONAL	kcl k ax n v eh n sh ax n el
COOK	kcl k uh kcl k
COPELAND	kcl k ow pcl p l ax n
COPELAND+S	kcl k ow pcl p l ax n z
CORAL	kcl k ao r el
COULD	kcl k uh dcl
COULDN+T	kcl k uh dx ax n
COUNT	kcl k aw n
COUNTED	kcl k aw n ax dcl
COUNTING	kcl k aw n eng
CROVL	kcl k r ow v el

CROVLS	kcl k r ow v el z
CRUISER	kcl k r uw z axr
CRUISER+S	kcl k r uw z axr z
CRUISERS	kcl k r uw z axr z
CRUISERS+S	kcl k r uw z axr z
CRUISING	kcl k r uw z ih ng
CURRENT	kcl k er ax n
CURRENTLY	kcl k er ax n tcl t l iy
DALE	dcl d ey l
DALE+S	dcl d ey l z
DARWIN	dcl d aa r w ax n
DATA	dcl d ey dx ax
DATE	dcl d ey tcl
DATED	dcl d ey dx ax dcl
DATES	dcl d ey tcl t s
DAVIDSON	dcl d ey v ax dcl d s ax n
DAVIDSON+S	dcl d ey v ax dcl d s en z
DAY	dcl d ey
DAYS	dcl d ey z
DDD992	dcl d iy dcl d iy dcl d iy n ay n n ay n tcl t uw
DECEMBER	dcl d ax s eh m b axr
DECREASE	dcl d iy kcl k r iy s
DECREASED	dcl d iy kcl k r iy s tcl
DECREASING	dcl d iy kcl k r iy s ih ng
DEFAULT	dcl d iy f ao l
DEFAULTS	dcl d iy f el tcl t s
DEFINE	dcl d ax f ay n
DEFINED	dcl d ax f ay n
DEFINING	dcl d ax f ay n ih ng
DEFINITION	dcl d eh f ax n ih sh ax n
DEFINITIONS	dcl d eh f ax n ih sh en z
DEGRADATION	dcl d eh gcl g axr dcl d ey sh ax n
DEGRADATIONS	dcl d eh gcl g axr dcl d ey sh en z
DEGRADE	dcl d ax gcl g r ey dcl
DEGRADED	dcl d ax gcl g r ey dx ax dcl
DEGRADING	dcl d ax gcl g r ey dx ih ng
DEGREES	dcl d ax gcl g r iy z
DELETE	dcl d ax l iy tcl
DELETED	dcl d ax l iy dx ax dcl
DELETING	dcl d ax l iy dx ih ng
DENVER	dcl d eh n v axr
DENVER+S	dcl d eh n v axr z
DEPLOYED	dcl d ax pcl p l oy dcl
DEPLOYMENT	dcl d ax pcl p l oy m ax n
DEPLOYMENTS	dcl d ax pcl p l oy m ax n s
DEPTH	d eh pcl p th
DEPTHS	dcl d eh pcl s
DESTINATION	dcl d eh s tcl t ax n ey sh ax n
DESTINATIONS	dcl d eh s tcl t ax n ey sh en z
DID	d ih dcl
DIDN+T	dcl d ih dcl d ax n
DIEGO-GARCIA	dcl d iy ey gcl g ow gcl g axr s iy ax
DIESEL	dcl d iy z el
DIM	dcl d ih m
DISPLACEMENT	d ax s pcl p l ey s m ax n
DISPLACEMENTS	d ax s pcl p l ey s m ax n s
DISPLAY	d ax s pcl p l ey
DISPLAYED	d ax s pcl p l ey dcl
DISPLAYING	d ax s pcl p l ey ih ng

DISTANCE	dcl d ih s tcl t ax n s
DIXON	dcl d ih kcl k s ax n
DIXON+S	dcl d ih kcl k s en z
DMDS	dcl d iy eh m dcl d iy eh s
DO	d uw
DOES	d ax z
DOESN+T	dcl d ah z ax n
DON+T	dcl d ow n
DOWNES	dcl d aw n z
DOWNES+S	dcl d aw n z
DOWNGRADE	dcl d aw n g r ey dcl
DOWNGRADED	dcl d aw n g r ey dx ax dcl
DRAFT	dcl jh r ae f tcl
DRAFTS	dcl jh r ae f tcl t s
DRAW	dcl jh r ao
DUBUQUE	d ax bcl b y uw kcl k
DUBUQUE+S	d ax bcl b y uw kcl k s
DUE	dcl d uw
DURING	dcl d er ih ng
EACH	q iy tcl ch
EARLIER	q er l iy axr
EARLIEST	q er l iy ax s tcl
EARLY	q er l iy
EAST	iy s tcl
EASTERN	q iy s tcl t axr n
EASTPAC	q iy s pcl p ae kcl k
EASTPAC+S	q iy s pcl p ae kcl k s
ECG041	q iy s iy dcl jh iy z iy r ow f ao r w ah n
ECHO	q eh kcl k ow
ECONOMIC	q eh kcl k ax n aa m ax kcl k
EDIT	q eh dx ax tcl
EDITED	q eh dx ax dx ax dcl
EDITING	q eh dx ax dx ih ng
EIGHT	ey tcl
EIGHTEEN	q ey tcl t iy n
EIGHTEENTH	q ey tcl t iy n th
EIGHTH	q ey tcl th
EIGHTY	q ey dx iy
EISENHOWER	q ay z ax n hh aw axr
EISENHOWER+S	q ay z ax n hh aw axr z
ELEVEN	ax l eh v ax n
ELEVENTH	ax l eh v ax n th
ELIZABETH	ax l ih z ax bcl b ax th
EMPLOYED	ax m pcl p l oy dcl
END	q eh n
ENDING	q eh n d ih ng
ENGLAND	q ih ng g l ax n
ENGLAND+S	q ih ng g l ax n z
ENGLISH	q ih ng g l ax sh
ENOUGH	ax n ah f
ENROUTE	ax n r uw tcl
ENTERPRISE	q eh n axr pcl p r ay z
ENTERPRISE+S	q eh n axr pcl p r ay z ax z
EQUIPMENT	iy kcl k w ih pcl m ax n
EQUIPPED	iy kcl k w ih pcl tcl t
ESTEEM	ax s tcl t iy m
ESTEEM+S	ax s tcl t iy m z
ESTIMATED	q eh s tcl t ax m ey dx ax dcl
ETA	q iy tcl t iy q ey

ETR	q iy tcl t iy aa r
EVER	q eh v axr
EVERETT	q eh v r ax tcl
EXPECTED	ax kcl k s pcl p eh kcl t ax dcl
FANNING	f ae n ih ng
FANNING+S	f ae n ih ng z
FAR	f aa r
FARTHER	f aa r dh axr
FARTHEST	f aa r dh ax s tcl
FAST	f ae s tcl
FASTER	f ae s tcl t axr
FASTEST	f ae s tcl t ax s tcl
FEBRUARY	f eh bcl b y uw eh r iy
FEET	f iy tcl
FFF088	q eh f eh f eh f z iy r ow ey dcl d ey tcl
FIFTEEN	f ih f tcl t iy n
FIFTEENTH	f ih f tcl t iy n th
FIFTH	f ih f th
FIFTY	f ih f tcl t iy
FIGURE	f ih gcl g y axr
FIGURES	f ih gcl g y axr z
FIJI	f iy dcl jh iy
FIND	f ay n
FIREBUSH	f ay axr bcl b uh sh
FIREBUSH+S	f ay axr bcl b uh sh ax z
FIRST	f er s tcl
FIVE	f ay v
FIXED	f ih kcl k s tcl
FLASHER	f l ae sh axr
FLASHER+S	f l ae sh axr z
FLEET	f l iy tcl
FLEETS	f l iy tcl t s
FLINT	f l ih n
FLINT+S	f l ih n s
FOOTER	f uh dx axr
FOR	f axr
FORMOSA	f axr m ow s ax
FORTY	f ao r dx iy
FOUR	f ao r
FOURTEEN	f ao r tcl t iy n
FOURTEENTH	f ao r tcl t iy n th
FOURTH	f ao ax th
FOX	f aa kcl k s
FOX+S	f aa kcl k s ax z
FREDERICK	f r eh dcl jh r ax kcl k
FREDERICK+S	f r eh dcl d axr kcl k s
FRENCH	f r eh n tcl ch
FRESNO	f r eh z n ow
FRESNO+S	f r eh z n ow z
FRIDAY	f r ay dcl d ey
FRIDAY+S	f r ay dx iy z
FRIGATE	f r ih gcl g ax tcl
FRIGATE+S	f r ih gcl g ax tcl t s
FRIGATES	f r ih gcl g ax tcl t s
FRIGATES+S	f r ih gcl g ax tcl s
FROM	f r ax m
FUEL	f y uw l
FULL	f uh l
GALLONS	gcl g ae l ax n z

GALVESTON	gcl g ae l v ax s tcl t ax n
GAS	gcl g ae s
GET	g eh tcl
GIVE	g ih v
GLACIER	gcl g l ey sh axr
GLACIER+S	gcl g l ey sh axr z
GNOMONIC	n ax m aa n ax kcl k
GO	gcl g ow
GOING	gcl g ow ih ng
GONE	gcl g aa n
GREAT	gcl g r ey tcl
GREATER	gcl g r ey dx axr
GREATEST	gcl g r ey dx ax s tcl
GREEN	gcl g r iy n
GRID	gcl g r ih dcl
GRIDLEY	gcl g r ih dcl d l iy
GRIDLEY+S	gcl g r ih dcl l iy z
GRILL	gcl g r ih l
GROSS	gcl g r ow s
GROUP	gcl g r uw pcl p
GROUPS	gcl g r uw pcl p s
GUARDFISH	gcl g aa r dcl d f ih sh
GUARDFISH+S	gcl g aa r dcl f ih sh ax z
GUITARRO	gcl g ih tcl t aa r ow
GUITARRO+S	gcl g ih tcl t aa r ow z
GULF	gcl g ah l f
HAD	hh ae dcl
HALF	hh ae f
HARBOR	hh aa r bcl b axr
HARPOON	hh aa r pcl p uw n
HAS	hh ae z
HASN+T	hh ae z ax n
HAVE	hh ae v
HAVEN+T	hh ae v ax n
HAWKBILL	hh ao kcl b ih l
HAWKBILL+S	hh ao kcl b ih l z
HE	hh iy
HE+S	hv iy z
HECTOR	hh eh kcl t axr
HECTOR+S	hh eh kcl t axr z
HEPBURN	hh eh bcl b axr n
HEPBURN+S	hh eh pcl b axr n z
HER	hh axr
HERS	hh er z
HFDF	q ey tcl ch eh f dcl d iy eh f
HIGH	hh ay
HIGHER	hh ay axr
HIGHEST	hh ay ax s tcl
HIM	hh ih m
HIS	hh ih z
HOME	hh ow m
HOMER	hh ow m axr
HONG-KONG	hh ao ng kcl k ao ng
HONOLULU	hh aa n ax l uw l uw
HOOKED	hh uh kcl t
HORNE	hh ao r n
HORNE+S	hh ao r n z
HOUR	q aw axr
HOURS	aw axr z

HOW	hh aw
HUNDRED	hh ah n d axr dx
ICE-NINE	q ay s n ay n
ID	q ay dcl d iy
IDENTIFICATION	ay dcl d eh n ax f ax kcl k ey sh ax n
IDENTIFICATIONS	ay dcl d eh n ax f ax kcl k ey sh en z
IF	ih f
IN	ih n
INCLUDE	ax ng kcl k l uw dcl
INCLUDED	ax ng kcl k l uw dx ax dcl
INCLUDING	ax ng kcl k l uw dx ih ng
INCREASE	q ih ng kcl k r iy s
INCREASED	q ih ng kcl k r iy s tcl
INCREASING	q ih ng kcl k r iy s ih ng
INDEPENDENCE	q ih n d iy pcl p eh n d ax n s
INDEPENDENCE+S	q ih n d iy pcl p eh n d ax n tcl s ax z
INDIAN	q ih n d iy ax n
INDONESIA	q ih n d ow n ey zh ax
INFORMATION	q ih n f ax m ey sh ax n
INSTALLED	ax n s tcl t aa l dcl
INSTEAD	ax n s tcl t eh dcl
INSUFFICIENT	q ih n s ax f ih sh ax n
INVOLVED	ax n v aa l v dcl
INVOLVING	ax n v aa l v ih ng
IRONWOOD	q ay axr n w uh dcl
IRONWOOD+S	q ay axr n w uh dcl z
IS	ih z
ISLANDS	q ay l ax n z
ISN+T	q ih z ax n
IT	ih tcl
IT+S	ih tcl t s
ITS	ih tcl t s
JANUARY	dcl jh ae n y uw axr iy
JAPAN	dcl jh ax pcl p ae n
JARRETT	dcl jh eh r ax tcl
JARRETT+S	dcl jh eh axr tcl t s
JARVIS	dcl jh aa r v ax s
JARVIS+S	dcl jh aa r v ax s ax z
JASON	dcl jh ey s ax n
JASON+S	dcl jh ey s en z
JULY	dcl jh uw l ay
JUNE	dcl jh uw n
JUPITER	dcl jh uw pcl p ax dx axr
JUPITER+S	dcl jh uw pcl p ax dx axr z
KENNEDY	kcl k eh n ax dx iy
KENNEDY+S	kcl k eh n ax dx iy z
KILOMETER	kcl k el aa m ax dx axr
KILOMETERS	kcl k el aa m ax dx axr z
KIRK	kcl k er kcl k
KIRK+S	kcl k er kcl k s
KISKA	kcl k ih s kcl k ax
KISKA+S	kcl k ih s kcl k ax z
KNOT	n aa tcl
KNOTS	n aa tcl t s
KODIAK	kcl k ow dcl iy ae kcl k
KOREA	kcl k ao r iy ax
KOREAN	kcl k ao r iy ax n
LAMPS	l ae m p s
LANTFLT	l ae n t f l iy tcl

LARGE	l aa r dcl jh
LARGER	l aa r dcl jh axr
LARGEST	l aa r dcl jh ax s tcl
LAST	l ae s tcl
LAT	l ae tcl
LAT-LON	l ae tcl q l aa n
LAT-LONS	l ae tcl q l aa n z
LATER	l ey dx axr
LATEST	l ey dx ax s tcl
LATITUDE	l ae dx ax tcl t uw dcl
LATITUDES	l ae dx ax tcl t uw dcl z
LATS	l ae tcl t s
LEAST	l iy s tcl
LEFT	l eh f tcl
LENGTH	l eh ng th
LENGTHS	l eh ng th s
LENINGRAD	l eh n ax n g r ae dcl
LESS	l eh s
LETTER	l eh dx axr
LETTERS	l eh dx axr z
LEVEL	l eh v el
LEVELS	l eh v el z
LINK-11	l ih ng kcl k ax l eh v ax n
LIST	l ih s tcl
LOCATED	l ow kcl k ey dx ax dcl
LOCATION	l ow kcl k ey sh ax n
LOCATIONS	l ow kcl k ey sh en z
LOCKWOOD	l aa kcl k w uh dcl
LOCKWOOD+S	l aa kcl k w uh dcl z
LON	l aa n
LONG	l ao ng
LONGER	l ao ng g axr
LONGEST	l ao ng g ax s tcl
LONGITUDE	l aa n jh ax tcl t uw dcl
LONGITUDES	l aa n jh ax tcl t uw dcl z
LONS	l aa n z
LOW	l ow
LOWER	l ow axr
LOWEST	l ow ax s tcl
M-CODE	q eh m kcl k ow dcl
M-CODES	q eh m kcl k ow dcl z
M-RATING	q eh m r ey dx ih ng
M-RATINGS	q eh m r ey dx ih ng z
M1	q eh m w ah n
M2	q eh m tcl t uw
M3	q eh m th r iy
M4	q eh m f ao r
M5	q eh m f ay v
MADAGASCAR	m ae dcl d ax gcl g ae s kcl k aa r
MADE	m ey dcl
MAKE	m ey kcl k
MAKING	m ey kcl k ih ng
MANCHESTER	m ae n tcl ch ax s tcl t axr
MANHATTAN	m ae n hh ae tcl t ax n
MANHATTAN+S	m ae n hh ae dx ax n z
MANILA	m ax n ih l ax
MANY	m eh n iy
MARCH	m aa r tcl ch
MARS	m aa r z

MARS+S	m aa r z
MAX	m ae kcl k s
MAXIMUM	m ae kcl k s em ax m
MAY	m ey
MCCLUSKY	m ax kcl k l ah s kcl k iy
MCCLUSKY+S	m ax kcl k l ah s kcl k iy z
ME	m iy
MEASURE	m ey zh axr
MERCATOR	m axr kcl k ey dx axr
MERCURY	m er kcl k y axr iy
MERCURY+S	m er kcl k y axr iy z
METEOR	m iy tcl t iy axr
METEOR+S	m iy tcl t iy axr z
METERS	m iy dx axr z
METRIC	m eh tcl ch r ax kcl k
MEXICO	m eh kcl k s ax kcl k ow
MIAMI	m ay ae m iy
MIDGETT	m ih dcl jh ax tcl
MIDGETT+S	m ih dcl jh ax tcl t s
MIDPAC	m ih dcl pcl p ae kcl k
MIDPAC+S	m ih dcl pcl p ae kcl k s
MIDWAY	m ih dcl d w ey
MIDWAY+S	m ih dcl d w ey z
MILE	m ay el
MILES	m ay el z
MIND	m ay n
MINUTES	m ih n ax tcl t s
MISHAWAKA	m ih sh ax w aa kcl k ax
MISHAWAKA+S	m ih sh ax w aa kcl k ax z
MISSION	m ih sh ax n
MISSIONS	m ih sh en z
MISSISSIPPI	m ih s ax s ih pcl p iy
MISSISSIPPI+S	m ih s ax s ih pcl p iy z
MIW	q eh m ay dcl d ah bcl b y uw
MOB	q eh m ow bcl b iy
MONDAY	m ah n d ey
MONDAY+S	m ah n d ey z
MONGOLIA	m aa ng g ow l iy ax
MONTH	m ah n th
MONTH+S	m ah n th s
MONTHS	m ah n th s
MONTICELLO	m aa n ax s eh l ow
MONTICELLO+S	m aa n ax s eh l ow z
MORE	m ao r
MOST	m ow s tcl
MOZAMBIQUE	m ow z ae m bcl b iy kcl k
MUCH	m ah tcl ch
N92762	q eh n n ay n tcl t uw s eh v ax n s ih kcl k s tcl t uw
NAME	n ey m
NAMES	n ey m z
NAPLES	n ey pcl p el z
NASHUA	n ae sh uw ax
NASHUA+S	n ae sh uw ax z
NEAR	n iy r
NEARER	n iy r axr
NEAREST	n iy r ax s tcl
NEVER	n eh v axr
NEW	n uw
NEW-YORK	n uw y ao r kcl k

NEW-ZEALAND	n uw z iy l ax n
NEWCASTLE	n uw kcl k ae s el
NEWER	n uw axr
NEWEST	n uw ax s tcl
NEXT	n eh kcl k s tcl
NINE	n ay n
NINETEEN	n ay n tcl t iy n
NINETEENTH	n ay n tcl t iy n th
NINETY	n ay n iy
NINTH	n ay n th
NO	n ow
NOME	n ow m
NORTH	n ao ax th
NORTHERN	n ao r dh axr n
NOT	n aa tcl
NOVA	n ow v ax
NOVEMBER	n ow v eh m b axr
NOW	n aw
NTDS	q eh n tcl t iy dcl d iy eh s
NUCLEAR	n uw kcl k l iy axr
NUMBER	n ah m b axr
OAKLAND	ow kcl k l ax n
OCEAN	ow sh ax n
OCTOBER	aa kcl t ow bcl b axr
OF	ax v
OFF	q ao f
OLD	ow dx
OLDER	ow dcl d axr
OLDEST	ow l dcl d ax s tcl
OLYMPIA	ow l ih m pcl p iy ax
ON	q aa n
ONE	w ah n
ONLY	ow n l iy
OPEN	ow pcl p ax n
OR	ao r
ORANGE	ao axr n jh
OSGP	q ow eh s dcl jh iy pcl p iy
OVERALL	q ow v axr ao l
OVERLAY	q ow v axr l ey
OVERLAYS	q ow v axr l ey z
PAC	pcl p ae kcl k
PACFLT	pcl p ae kcl k f l iy tcl
PACIFIC	pcl p ax s ih f ax kcl k
PANAMA	pcl p ae n ax m aa
PARAMETER	pcl p axr ae m ax dx axr
PARAMETERS	pcl p axr ae m ax dx axr z
PEARL	pcl p er l
PEORIA	pcl p iy ao r iy ax
PEORIA+S	pcl p iy ao r iy ax z
PERCENT	pcl p axr s eh n
PERSIAN	pcl p er zh ax n
PERSONNEL	pcl p axr s ax n eh l
PHILIPPINE	f ih l ax pcl p iy n
PHILIPPINES	f ih l ax pcl p iy n z
PIGEON	pcl p ih dcl jh ax n
PIGEON+S	pcl p ih dcl jh ax n z
PLH003	pcl p iy eh l ey tcl ch z iy r ow z iy r ow th r iy
PLUCK	pcl p l ah kcl k
PLUCK+S	pcl p l ah kcl k s

PLUNGER	pcl p l ah n jh axr
PLUNGER+S	pcl p l ah n jh axr z
POLLACK	pcl p aa l ax kcl k
POLLACK+S	pcl p aa l ax kcl k s
POLYNESIA	pcl p aa l ax n iy zh ax
PORT	pcl p ao r tcl
PORTS	pcl p ao r tcl t s
POSIT	pcl p aa z ax tcl
POSITION	pcl p ax z ih sh ax n
POSITIONS	pcl p ax z ih sh en z
POSITS	pcl p aa z ax tcl t s
POUGHKEEPSIE	pcl p uw kcl k ih pcl p s iy
POUGHKEEPSIE+S	pcl p uw kcl k ih pcl p s iy z
POWERED	pcl p aw axr dcl
PRAIRIE	pcl p r eh r iy
PRAIRIE+S	pcl p r eh r iy z
PRESENT	pcl p r eh z ax n
PREVIOUS	pcl p r iy v iy ax s
PROBLEM	pcl p r aa bcl l ax m
PROBLEMS	pcl p r aa bcl l ax m z
PROJECTION	pcl p axr dcl jh eh kcl k sh ax n
PROPELLED	pcl p axr pcl p eh l dcl
PROPULSION	pcl p axr pcl p ah l sh ax n
PUFFER	pcl p ah f axr
PUFFER+S	pcl p ah f axr z
PUGET-1	pcl p y uw dcl jh ax tcl q w ah n
QUARTER	kcl k w ao r dx axr
QUARTERS	kcl k w ao r dx axr s
QUEENFISH	kcl k w iy n f ih sh
QUEENFISH+S	kcl k w iy n f ih sh ax z
RADAR	r ey dx aa r
RAMSEY	r ae m z iy
RAMSEY+S	r ae m z iy z
RANGER	r ey n jh axr
RANGER+S	r ey n jh axr z
RATED	r ey dx ax dcl
RATHBURNE	r ae th bcl b axr n
RATHBURNE+S	r ae th bcl b axr n z
RATING	r ey dx ih ng
RATINGS	r ey dx ih ng z
READINESS	r eh dx iy n ax s
REASON	r iy z ax n
REASONER	r iy z ax n axr
REASONER+S	r iy z ax n axr z
REASONS	r iy z en z
RECENT	r iy s ax n
RECLAIMER	r iy kcl k l ey m axr
RECLAIMER+S	r iy kcl k l ey m axr z
RED	r eh dcl
REDEFINE	r iy dcl d iy f ay n
REDEFINED	r iy dcl d iy f ay n
REDEFINING	r iy dcl d iy f ay n ih ng
REDISPLAY	r iy dcl d ax s pcl p l ey
REDO	r iy dcl d uw
REDRAW	r iy dcl jh r ao
REEVES	r iy v z
REEVES+S	r iy v z
REMAINING	r iy m ey n ih ng
REMARK	r iy m aa r kcl k

REMARKS	r iy m aa r kcl k s
REPAIR	r iy pcl p eh r
REPAIRED	r iy pcl p eh r dcl
REPAIRING	r iy pcl p eh r ih ng
REPLACED	r iy pcl p l ey s tcl
REPORT	r iy pcl p ao r tcl
REPORTED	r iy pcl p ao r dx ax dcl
REPORTING	r iy pcl p ao r dx ih ng
REPORTS	r iy pcl p ao r tcl t s
RESET	r iy s eh tcl
RESOLUTION	r eh z ax l uw sh ax n
RESOLVED	r iy z ao l v dcl
RESOURCE	r iy s ao r s
RESOURCES	r iy s ao r s ax z
REVIEW	r iy v y uw
ROSS	r ao
SACRAMENTO	s ae kcl k ax r m eh n t ow
SACRAMENTO+S	s ae kcl k axr m eh n tcl t ow z
SAIL	s ey l
SAME	s ey m
SAMPLE	s ae m pcl p el
SAMPLE+S	s ae m pcl p el z
SAN-DIEGO	s ae n d iy ey gcl g ow
SAN-FRAN	s ae n f r ae n
SASSAFRAS	s ae s ax f r ae s
SASSAFRAS+S	s ae s ax f r ae s ax z
SATURDAY	s ae dx axr dcl d ey
SATURDAY+S	s ae dx axr dcl d ey z
SAVE	s ey v
SCHENECTADY	s kcl k ax n eh kcl t ax dx iy
SCHENECTADY+S	s kcl k ax n eh kcl t ax dx iy z
SCREEN	s kcl k r iy n
SCREENS	s kcl k r iy n z
SEA	s iy
SEAWOLF	s iy w uh l f
SEAWOLF+S	s iy w uh l f s
SECOND	s eh kcl k ax n
SECURITY	s ax kcl k y er ax dx iy
SENSOR	s eh n s axr
SENSORS	s eh n s axr z
SEPTEMBER	s ax pcl t eh m b axr
SET	s eh tcl
SETTING	s eh dx ih ng
SETTINGS	s eh dx ih ng z
SEVEN	s eh v ax n
SEVENTEEN	s eh v ax n tcl t iy n
SEVENTEENTH	s eh v ax n tcl t iy n th
SEVENTH	s eh v ax n th
SEVENTY	s eh v ax n iy
SHASTA	sh ae s tcl t ax
SHASTA+S	sh ae s tcl t ax z
SHE	sh iy
SHE+S	sh iy z
SHERMAN	sh er m ax n
SHERMAN+S	sh er m ax n z
SHIP	sh ih pcl p
SHIPS	sh ih pcl p s
SHIPS+S	sh ih pcl p s
SHOW	sh ow

SHOWING	sh ow ih ng
SHOWN	sh ow n
SIBERIAN	s ay bcl b iy r iy ax n
SIDNEY	s ih dcl n iy
SINCE	s ih n s
SINGAPORE	s ih ng ax pcl p ao r
SIX	s ih kcl k s
SIXTEEN	s ih kcl k s tcl t iy n
SIXTEENTH	s ih kcl k s tcl t iy n th
SIXTH	s ih kcl k s th
SIXTY	s ih kcl k s tcl t iy
SIZE	s ay z
SLOW	s l ow
SLOWER	s l ow axr
SLOWEST	s l ow ax s tcl
SLQ-32	q eh s eh l kcl k y uw th er dx iy tcl t uw
SMALL	s m ao l
SMALLER	s m ao l axr
SMALLEST	s m ao l ax s tcl
SOHO	s ow hh ow
SOLOMON	s ao l ax m ax n
SONAR	s ow n aa r
SOON	s uw n
SOONER	s uw n axr
SOUTH	s aw th
SOUTHERN	s ah dh axr n
SOVIET	s ow v iy ax tcl
SPEED	s pcl p iy dcl
SPEEDS	s pcl p iy dcl z
SPS-40	q eh s pcl p iy eh s f ao r dx iy
SPS-48	q eh s pcl p iy eh s f ao r dx iy ey tcl
SQQ-23	q eh s kcl k y uw kcl k y uw tcl t w eh n iy th r iy
START	s tcl t aa r tcl
STARTED	s tcl t aa r dx ax dcl
STARTING	s tcl t aa r dx ih ng
STATES	s tcl t ey tcl t s
STATION	s tcl t ey sh ax n
STATUS	s tcl t ey dx ax s
STEAM	s tcl t iy m
STEREOGRAPHIC	s tcl t eh r iy ow gcl g r ae f ax kcl k
STERETT	s tcl t eh r ax tcl
STERETT+S	s tcl t eh r ax tcl t s
STRAIT	s tcl ch r ey tcl
SUB	s ah bcl b
SUB+S	s ah bcl b z
SUBIC	s uw bcl b ax kcl k
SUBMARINE	s ah bcl m axr iy n
SUBMARINE+S	s ah bcl m axr iy n z
SUBMARINES	s ah bcl m axr iy n z
SUBS	s ah bcl b z
SUFFICIENT	s ax f ih sh ax n
SUMMARIZE	s ah m axr ay z
SUNDAY	s ah n d ey
SUNDAY+S	s ah n d ey z
SUPPLIES	s ax pcl p l ay z
SUPPLY	s ax pcl p l ay
SUPPOSED	s ax pcl p ow z
SURFACE	s er f ax s
SUSTAINED	s ax s tcl t ey n

SWITCH	s w ih tcl ch
SWITCHES	s w ih tcl ch ax z
SWORDFISH	s ao r dcl d f ih sh
SWORDFISH+S	s ao r dcl f ih sh ax z
SYSTEM	s ih s tcl t ax m
T-LAM	tcl t iy l ae m
TACAN	tcl t ae kcl k ae n
TAIWAN	tcl t ay w aa n
TAKE	tcl t ey kcl k
TAKEN	tcl t ey kcl k ax n
TASM	tcl t ae s ax m
TEN	tcl t eh n
TENTH	tcl t eh n th
TEST	tcl t eh s tcl
TEXAS	tcl t eh kcl k s ax s
TEXAS+S	tcl t eh kcl k s ax s ax z
TFCC	tcl t iy eh f s iy s iy
THAILAND	tcl t ay l ae n
THAN	dh ae n
THAT	dh ae tcl
THE	dh ax
THEIR	dh eh r
THEM	dh eh m
THERE	dh eh r
THESE	dh iy z
THEY	dh ey
THIRD	th er dcl
THIRTEEN	th er tcl t iy n
THIRTEENTH	th er tcl t iy n th
THIRTIETH	th er dx iy ax th
THIRTY	th er dx iy
THIS	dh ih s
THOSE	dh ow z
THOUSAND	th aw z ax n
THREAT	th r eh tcl
THREATS	th r eh tcl t s
THREE	th r iy
THURSDAY	th er z dcl d ey
THURSDAY+S	th er z dcl d ey z
TICONDEROGA	tcl t ay kcl k aa n d axr ow gcl g ax
TICONDEROGA+S	tcl t ay kcl k aa n d axr ow gcl g ax z
TIME	tcl t ay m
TIMES	tcl t ay m z
TO	tcl t uw
TODAY	tcl t ax dcl d ey
TODAY+S	tcl t ax dcl d ey z
TOGGLE	tcl t aa gcl g el
TOGGLED	tcl t aa gcl g el dcl
TOGGLING	tcl t aa gcl g el ih ng
TOKYO	tcl t ow kcl k iy ow
TOMORROW	tcl t ax m ao r ow
TOMORROW+S	tcl t ax m ao r ow z
TONKIN	tcl t ao ng kcl k ax n
TONS	tcl t ah n z
TOTAL	tcl t ow dx el
TOWNSVILLE	tcl t aw n z v el
TRACK	tcl ch r ae kcl k
TRACKS	tcl ch r ae kcl k s
TRAINING	tcl ch r ey n ih ng

TRIPOLI	tcl ch r ih pcl p ow l iy
TRIPOLI+S	tcl ch r ih pcl p ow l iy z
TRUE-VIEW	tcl ch r uw v y uw
TUESDAY	tcl t uw z dcl d ey
TUESDAY+S	tcl t uw z dcl d ey z
TURBINE	tcl t er bcl b ax n
TURN	tcl t er n
TURNED	tcl t er n
TURNING	tcl t er n ih ng
TUSCALOOSA	tcl t ah s kcl k ax l uw s ax
TUSCALOOSA+S	tcl t ah s kcl k ax l uw s ax z
TWELFTH	tcl t w eh l f th
TWELVE	tcl t w eh l v
TWENTIETH	tcl t w ah n iy ax th
TWENTY	tcl t w ah n iy
TWO	tcl t uw
TYPE	tcl t ay pcl p
TYPES	tcl t ay pcl p s
UNION	y uw n y ax n
UNIT	y uw n ax tcl
UNITED	y uw n ay dx ax dcl
UNTIL	ax n tcl t ih l
UOM	y uw ow q eh m
UPDATE	ah pcl d ey tcl
UPDATED	ah pcl d ey dx ax dcl
UPDATES	ah pcl d ey tcl t s
UPGRADE	ah pcl g r ey dcl
UPGRADED	ah pcl g r ey dx ax dcl
USA	y uw eh s q ey
USE	y uw z
USING	y uw z ih ng
USN	y uw eh s eh n
VALUE	v ae l y uw
VALUES	v ae l y uw z
VANCOUVER	v ae ng kcl k uw v axr
VANCOUVER+S	v ae ng kcl k uw v axr z
VANDERGRIFT	v ae n d axr gcl g r ih f tcl
VANDERGRIFT+S	v ae n d axr gcl g r ih f tcl t s
VARIOUS	v eh r iy ax s
VESSEL	v eh s el
VESSELS	v eh s el z
VICTORIA	v ax kcl t ao r iy ax
VIRGINIA	v axr dcl jh ih n y ax
VIRGINIA+S	v axr dcl jh ih n y ax z
VISUAL	v ih zh uw el
WABASH	w aa bcl b ae sh
WABASH+S	w aa bcl b ae sh ax z
WADSWORTH	w aa dcl z w axr th
WADSWORTH+S	w aa dcl z w axr th s
WAS	w ah z
WASN+T	w ah z ax n
WASP	w aa s pcl p
WASP+S	w aa s pcl p s
WE	w iy
WEDNESDAY	w eh n z dcl d ey
WEDNESDAY+S	w eh n z dcl d ey z
WEEK	w iy kcl k
WEEK+S	w iy kcl k s
WEEKS	w iy kcl k s

WELLINGTON	w eh l ax ng tcl t ax n
WENT	w eh n
WERE	w axr
WEREN+T	w axr n
WEST	w eh s tcl
WESTERN	w eh s tcl t axr n
WESTPAC	w eh s pcl p ae kcl k
WESTPAC+S	w eh s pcl p ae kcl k s
WHAT	w ah tcl
WHAT+RE	w ah dx axr
WHAT+S	w ah tcl t s
WHEN	w eh n
WHEN+LL	w eh n el
WHEN+S	w eh n z
WHERE	w eh r
WHERE+S	w eh r z
WHICH	w ih tcl ch
WHIPPLE	w ih pcl p el
WHIPPLE+S	w ih pcl p el z
WHO	hh uw
WHO+S	hh uw z
WHOSE	hh uw z
WHY	w ay
WHY+S	w ay z
WICHITA	w ih tcl ch ax tcl t ao
WICHITA+S	w ih tcl ch ax tcl t ao z
WILL	w el
WILLAMETTE	w ih l ax m eh tcl
WILLAMETTE+S	w ih l ax m eh tcl t s
WINAMAC	w ih n ax m ae kcl k
WINAMAC+S	w ih n ax m ae kcl k s
WINDOW	w ih n d ow
WINDOWS	w ih n d ow z
WITH	w ax th
WITHIN	w ax th ih n
WITHOUT	w ax th aw tcl
WON+T	w ow n
WORSE	w er s
WORST	w er s tcl
WOULD	w uh dcl
WOULDN+T	w uh dcl d ax n
YANKEE	y ae ng kcl k iy
YEAR	y iy r
YEARS	y iy ax z
YELLOW	y eh l ow
YESTERDAY	y eh s tcl t axr dcl d ey
YESTERDAY+S	y eh s tcl t axr dx iy z
YET	y eh tcl
YORKTOWN	y ao r kcl t aw n
YORKTOWN+S	y ao r kcl t aw n z
ZULU	z uw l uw

FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO
BIBLIOTECA

0000056105