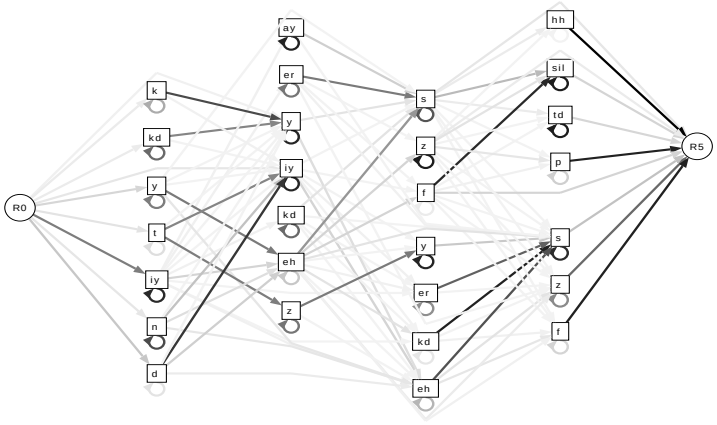
UNIVERSIDADE TÉCNICA DE LISBOA
INSTITUTO SUPERIOR TÉCNICO



Reconhecimento de Fala de Oradores Estrangeiros

Carlos Jorge da Conceição Teixeira

(Mestre)

Dissertação para a obtenção do Grau de Doutor em
Engenharia Electrotécnica e de Computadores

Lisboa, Setembro de 1998

Tese realizada sob a supervisão de

Isabel Maria Martins Trancoso

Professora Associada do
Departamento de Engenharia Electrotécnica e de Computadores
Instituto Superior Técnico
Universidade Técnica de Lisboa

Co-orientação de

António Joaquim dos Santos Romão Serralheiro

Professor Auxiliar do
Departamento de Engenharia Electrotécnica e de Computadores
Instituto Superior Técnico
Universidade Técnica de Lisboa

TÍTULO:	Reconhecimento de Fala de Oradores Estrangeiros
NOME:	Carlos Jorge da Conceição Teixeira
DOUTORAMENTO EM:	Engenharia Electrotécnica e de Computadores
ORIENTADOR:	Isabel Maria Martins Trancoso
CO-ORIENTADOR:	António Joaquim dos Santos Romão Serralheiro
PROVAS CONCLUÍDAS EM:	

RESUMO:

Este trabalho descreve a introdução de algumas metodologias de reconhecimento robusto de fala baseadas em modelos de Markov não observáveis, tendo em vista a sua utilização por parte de oradores estrangeiros.

Um problema que afecta principalmente os reconhecedores com vocabulários pequenos, advém da tendência dos utentes casuais em construir frases coloquiais centradas em cada palavra desse vocabulário. A solução tradicional para este tipo de problema consiste na utilização de um modelo de escoamento que permite eliminar as palavras estranhas ao vocabulário do reconhecedor. Nesta dissertação estudou-se a utilização de diversos modelos de escoamento em simultâneo e a sua articulação com a presença de oradores com sotaque estrangeiro.

Quando a língua para a qual um reconhecedor automático de fala foi desenvolvido não é a língua materna do orador, verificam-se quebras acentuadas no seu desempenho. Este problema só foi equacionado recentemente e as soluções disponíveis não são diferenciadas das existentes para o problema, mais geral, da independência dos reconhecedores em relação ao orador. Neste trabalho propõem-se algumas técnicas para atenuar as referidas quebras de desempenho. Uma destas técnicas consiste na adopção de um modelo estatístico multipronúncia para as transcrições fonotípicas alternativas de cada palavra.

Este estudo baseou-se num corpus de fala que inclui um vocabulário de palavras inglesas recolhido a partir de cerca de 120 oradores em seis países da União Europeia.

PALAVRAS-CHAVE: reconhecimento automático de fala, sotaque estrangeiro, detecção de palavras-chave, modelos de escoamento, modelos multipronúncia, transcrição fonotípica.

TITLE: Speech Recognition for Foreign Speakers

ABSTRACT:

This dissertation describes the integration of some methodologies of robust speech recognition based on hidden Markov models, considering its use by foreign speakers.

A problem that mainly affects small vocabulary recognisers, arises from a tendency of the speaker to use words not included in these vocabularies, namely while attempting to construct a natural sentence around each keyword. In this work, the use of multiple sink models in the recogniser was tested with a multi-accent speech corpus.

When the language for which the recogniser was built is not the native language of the speaker, a severe loss of performance can be detected. This problem has only be equated recently. Thus, the solutions used are usually similar to the ones proposed for the more general problem of the speaker independent recognition. In this dissertation some techniques are proposed in order to reduce the referred loss of performance. One of these techniques is based on a multi-pronunciation statistical model for the alternative phonotypic transcriptions of each word.

The speech corpus used for the experiments includes a vocabulary of English words collected from 120 speakers in six different European countries.

KEY-WORDS: automatic speech recognition, foreign accent, key-word detection, sink models, multi-pronunciation models, phonotypical transcription.

Dedico este trabalho
ao meu filho, Diogo.

Agradecimentos

A presente dissertação foi fruto, para além da inerente componente individual, da ajuda, do entusiasmo, da compreensão e da boa vontade de algumas pessoas que estiveram sempre de alguma forma perto do autor. Outras deram contribuições semelhantes durante períodos de tempo mais limitados, de acordo com as exigências e das circunstâncias que foram possíveis de obter. De facto, o presente trabalho, principalmente devido ao seu carácter interdisciplinar e ao extenso período de tempo durante o qual foi desenvolvido, beneficiou de uma ampla participação de pessoas e meios.

Em primeiro lugar quero agradecer à minha orientadora científica, Professora Isabel Trancoso, pela forma como me influenciou com os seus conhecimentos, o seu trabalho e o seu prestígio na área do processamento da fala. Pelas correcções e sugestões no decorrer das diversas fases deste trabalho. Pelo entusiasmo e amizade.

Agradeço ao meu co-orientador científico, Professor António Serralheiro, pelos conselhos e pela leitura cuidada da tese. Pela amizade e o apoio insubstituível na minha actividade de docente.

Agradeço ao Professor Paul Daalsgard por me ter concedido dois estágios na Universidade de Aalborg. Pelos seus conselhos e motivação em particular para os aspectos da multilingualidade no processamento de fala.

Ao Professor Luís Borges da Almeida, que permitiu a minha inserção no meu actual grupo de trabalho no INESC, para além de outras intervenções decisivas para o desenvolvimento do meu trabalho de doutoramento.

Ao Professor Borge Lindberg pelos inúmeros conselhos e bibliografia cedida que de outro modo nunca teria chegado ao meu conhecimento.

Ao Professor Luís Caldas de Oliveira, que para além de ser o principal responsável pelos meios informáticos utilizados, contribuiu com inúmeras sugestões ao longo do desenvolvimento deste trabalho.

Ao Engenheiro Claus Jacobsen e alguns dos seus colegas da TeleDanmark, pelos

inúmeros esclarecimentos sobre os programas SIRtrain e SIRsignal, sobre o corpus de fala SUNSTAR.

À Professora Maria do Céu Viana pela leitura cuidada de parte da presente dissertação e por muitas salutares discussões.

Ao Professor Tibault Langlois pelo seu encorajamento e amizade.

Ao Engenheiro Bjarne Andersen e a alguns dos seus colegas da Universidade de Aalborg, pelo trabalho que tiveram com pequenas adaptações feitas a meu pedido nos programas do SIRtrain e do SIRsignal, pela experiência transmitida com programas adequados para o processamento dos corpora de fala, assim como pela boa disposição.

Ao Doutor Onofre Moreira por ter disponibilizado a câmara anecóica para a recolha de parte do corpus de fala utilizado e por todos os esclarecimentos que sempre disponibilizou.

Gostaria também de aqui agradecer a um conjunto de pessoas com quem mantive um contacto, necessariamente pontual, mas que repetidamente demonstraram grande disponibilidade e amizade: ao Doutor Reinhold Haeb-Umbach e a alguns dos seus colegas, pelos esclarecimentos sobre o seu trabalho nos laboratórios da Philips de Aachen; à Doutora Kay Berkling dos laboratórios Lincoln (MIT), pelas sugestões e pelos esclarecimentos prestados em relação ao seu trabalho no "Oregon Graduate Institute of Science and Technology"; aos Doutores Mazin Rahim e Rafid Sukkar dos laboratórios Bell pelos esclarecimentos e sugestões prestadas.

À Professora Isabel Lourtie pela excelente estruturação, coordenação e ambiente de trabalho na cadeira de Teoria dos Sinais e Sistemas, que me permitiram conciliar de forma serena a actividade de docente com a escrita da tese.

Aos restantes colegas e amigos dos grupos de processamento de fala, de redes neuronais e de outros grupos de trabalho do INESC com quem partilhei muitos dos problemas que surgiram no decorrer deste trabalho: Carlos Martins, Carlos Ribeiro, Ciro Martins, Diamantino Caseiro, Fernando Silva, Frederico Rodrigues, Hugo Meinedo, Idalina Videira, Ilda Gonçalves, Isabel Mascarenhas, João Neto, Pedro Carvalho e Pedro Lopes.

Às restantes pessoas do INESC e do CLUL que colaboraram na recolha de parte do corpus de sinais de fala utilizado.

A todos os que contribuíram no desenvolvimento de *software* utilizado neste trabalho. Destes programas gostaria de realçar o sistema de preparação de documentos LaTeX, o sistema operativo *Linux*, o sistema de cálculo *Octave* bem como outros programas da *Free Software Foundation* e o sistema interactivo de visualização de grafos *daVinci*.

Agradece-se ao Departamento de Engenharia Electrotécnica e de Computadores do

IST e em particular, à Secção de Sistemas e Controlo por me terem concedido as dispensas de serviço de docente que permitiram a realização deste trabalho.

Agradece-se ainda ao Governo Dinamarquês e à Fundação Calouste Gulbenkien as bolsas concedidas para os estágios na Universidade de Aalborg. À NATO pelos encargos com o curso de Verão de 1993.

Por último, e sem que esta ordem desmereça a sua importância, mencionam-se pessoas em torno das quais se constrói para além de uma carreira, uma vida. Agradeço aos meus Pais e à minha Irmã por muita coisa. À Eugénia e aos meus Sogros, principalmente pelo cuidado com o nosso filho. Também ele de alguma forma compreendeu ou se habituou à ideia de que o pai tinha um trabalho para fazer desde que o conheceu. Aos restantes Amigos e Família.

Índice

Lista de Tabelas

Lista de Figuras

Capítulo 1

Introdução

O presente trabalho descreve modelos e métodos que foram utilizados em reconhecedores automáticos de fala com o objectivo de lhes permitir maior imunidade em relação a alguns factores de variabilidade do respectivo sinal.

Um reconhecedor automático de fala é um sistema capaz de, pelo menos, identificar várias palavras ou frases, quando proferidas oralmente por determinado indivíduo na ausência de qualquer outro sinal acústico. Idealmente, seria também capaz de transcrever qualquer discurso oral, pelo menos nas circunstâncias de audição consideradas aceitáveis por um ouvinte humano. Neste contexto, considera-se como dados para o reconhecimento apenas o sinal acústico resultante do processo da fala. Outros dados que podem ser considerados são os referentes à imagem facial do orador, em particular os movimentos dos lábios (Stork e Hennecke, 1996; Cole et al., 1998).

Os reconhecedores de fala considerados neste estudo são baseados em modelos de Markov não observáveis. O facto deste trabalho se ter baseado neste tipo de reconhecedores não impede no entanto, que um número significativo das suas conclusões se aplique a outros tipos de reconhecedores. Estas aparentam depender directamente das características intrínsecas do sinal de fala estudado e não tanto das particularidades dos instrumentos de análise utilizados, nomeadamente, as metodologias de reconhecimento.

Neste trabalho, considera-se o estudo de dois factores de variabilidade do sinal da fala que, em algumas situações interessantes de aplicação prática, causam quebras apreciáveis no desempenho dos referidos reconhecedores:

- utilização de palavras não pertencentes ao vocabulário do reconhecedor. A identificação forçada dessas palavras com outras pertencentes ao vocabulário compromete o desempenho global da aplicação;

- utilização do sistema por oradores para os quais a língua utilizada no reconhecedor não é a sua língua materna. Surgem, neste caso, sinais evidentes do designado sotaque estrangeiro ou não nativo, com diferenças fonológicas apreciáveis em relação ao material de fala nativo, habitualmente utilizado no desenvolvimento do reconhecedor.

1.1 Reconhecimento automático de fala

1.1.1 Motivação: interface homem-máquina

As aplicações computacionais fazem hoje parte do quotidiano da actividade empresarial, das unidades de produção, dos gabinetes de planeamento, de investigação e de projecto, e de alguns serviços de atendimento público. Para além do seu próprio desempenho, um dos factores essenciais do sucesso e aceitação destas aplicações perante os respectivos utilizadores, reside na rapidez e facilidade de aprendizagem da utilização destes sistemas.

Com o número e complexidade crescente dos meios automáticos ao dispor do Homem, estes factores têm cada vez mais importância. Muitas das aplicações disponíveis deveriam exigir o mínimo esforço da parte do utilizador casual. Mesmo nas aplicações que pressupõem uma utilização regular por parte do mesmo indivíduo, não se justifica, ainda assim, a consulta de extensos manuais para aceder a um número sempre crescente de parâmetros e opções.

A título de exemplo de aplicações de utilização tipicamente casual, considere-se como cenário uma ida ao cinema. Tal poderá implicar o uso de um serviço automático de reserva do lugar, outro de venda de bilhetes de metro, uma máquina de refrescos e outra de chocolates. Apesar de relativamente simples, cada um destes sistemas requer a leitura de instruções específicas que devem ser seguidas com rigor. Isto para uma utilização esporádica que poderá vir a repetir-se apenas meses depois.

Entre as aplicações de uso regular por parte do mesmo utilizador mencionam-se programas de computador, tal como o processador de texto, o vulgar gravador de vídeo, ou o painel de instrumentos dos automóveis mais sofisticados. O condutor destes veículos necessita de consultar regularmente o manual do veículo e dos respectivos acessórios ou, com maior probabilidade, desistirá de usar os mais complexos e supérfluos.

As interfaces homem-máquina tradicionais incluem a apresentação num ecrã de vídeo

de sofisticadas listas de escolha múltipla, seleccionáveis quer através de *rato*, de lápis óptico, ou mesmo por um ecrã sensível ao toque. Contudo, as aplicações com grande diversidade de dados de entrada, tais como o processamento de texto e a consulta de bases de dados, não dispensam ainda o teclado tradicional.

Do texto anterior decorre que as interfaces homem-máquina tradicionais não permitem uma utilização fácil, satisfatória e eficaz dos meios automáticos actualmente disponíveis. Por outro lado, condicionam a introdução de novas aplicações mais complexas. A solução desejável consiste na utilização do suporte natural da comunicação humana, capaz de representar a generalidade do conhecimento humano: a fala.

Os avanços tecnológicos das últimas décadas, nomeadamente nas áreas da electrónica e dos computadores, têm alimentado expectativas crescentes em relação à possibilidade de dispor das facilidades decorrentes do reconhecimento automático da fala.

O desenvolvimento de um robot capaz de executar tarefas domésticas a partir de ordens faladas, ou de um sistema capaz de gerar um texto escrito a partir de ditado, são apenas alguns dos objectivos práticos mais cobiçados. Apesar da motivação e consequente investimento em investigação e desenvolvimento que estas imagens suscitam, elas são ainda na actualidade, apenas objectos da ficção científica[1].

1.1.2 Métodos de reconhecimento automático

Existem na actualidade vários protótipos de reconhecedores, tendo sido a maioria desenvolvidos com base em duas metodologias distintas: a técnica de alinhamento temporal dinâmico (DTW — "dynamic time warping") e a baseada em modelos de Markov não observáveis (HMMs — "hidden Markov models"). Ambas as metodologias pressupõem a existência de material de fala, designado *de treino*, que permita construir modelos ou padrões de referência das unidades de fala a reconhecer.

O método DTW baseia-se na resolução de um problema elementar do sinal de fala referente à sua evolução no domínio do tempo (Itakura, 1975). Não só se fala a diferentes ritmos como existem diferentes temporizações entre frases, palavras e mesmo entre fone-

[1] O reconhecimento automático de fala inspirou as maravilhas da ficção científica tais como o robot R2D2 da série cinematográfica da "Guerra das Estrelas" ou o computador Hal do livro de Arthur C. Clarke eternizado no filme de Stanley Kubrick "2001 — Odisseia no Espaço". Curiosamente, a Mit Press publicou em Janeiro de 1997 (a auto-proclamada data de fabrico do Hal) o livro intitulado "Hal's Legacy: 2001's Computer As Dream and Reality" do editor David G. Stork. O livro recolhe contribuições de investigadores reconhecidos nas áreas tecnológicas mais críticas para os supostos construtores do Hal, entre as quais se destaca o reconhecimento da fala.

mas. Estas temporizações contêm geralmente informação útil para dois interlocutores em conversação, nomeadamente em relação ao respectivo estado emocional (Vroomen et al., 1993). Contudo, os reconhecedores actuais não utilizam esta informação, para a qual não existem ainda modelos suficientemente estudados.

Uma parcela significativa dos reconhecedores em exploração comercial e porventura a maioria dos utilizados pelos investigadores nesta área, utilizam modelos de Markov não observáveis (Baker, 1975; Jelinek et al., 1975; Jelinek, 1976; Rabiner, 1989; Young e Bloothooft, 1997). Devido à sua capacidade de integrar informação estatística detalhada, estes modelos revelaram-se muito úteis no caso do sinal de fala, permitindo atenuar alguns factores que degradam, geralmente de forma acentuada, o desempenho dos reconhecedores (secção 1.2).

Mais recentemente, as técnicas de processamento de redes neuronais artificiais têm sido aplicadas com sucesso no reconhecimento e em outras áreas do processamento de fala (Lippmann e Gold, 1987; Príncipe e Tracey, 1993; Yu e Oh, 1997). Saliente-se, em particular, a utilização dos *modelos híbridos*, os quais combinam o uso destas técnicas com o uso de modelos HMM (Bourlard e Wellekens, 1988; Morgan e Bourlard, 1990; Clary e Hansen, 1992; Cook e Robinson, 1995; Neto et al., 1995; Neto et al., 1996; Neto, 1998).

1.1.3 Tipologia dos modelos do sinal de fala

Uma das características mais importantes dos reconhecedores automáticos de fala diz respeito às unidades do sinal de fala consideradas para representação em modelos elementares: frases completas, palavras e elementos subpalavra. A escolha de um destes tipos de modelo depende essencialmente de algumas características genéricas do sinal de fala que se pretende reconhecer e que são habitualmente tipificadas da seguinte forma:

fala espontânea — sinais de fala recolhidos em situações reais, tais como são produzidos, por exemplo, numa conversa de café ou num debate político. Incluem, para além de ruídos estranhos ao processo de produção da fala, hesitações e fenómenos não linguísticos tal como tosse, risos, cliques, etc. Representa a última fronteira do reconhecimento de fala;

fala contínua — no sentido restrito refere-se a um sinal de fala resultante de um monólogo cuidado tal como o resultante da leitura de um texto. Num sentido mais amplo inclui a fala espontânea. Os modelos utilizados no reconhecimento automático da fala contínua são, quase exclusivamente, baseados em elementos subpalavra (modelos de subpalavras);

fala ligada — distingue-se essencialmente da fala contínua pela utilização exclusiva de um pequeno vocabulário. Os modelos utilizados no reconhecimento são, geralmente, baseados em palavras isoladas;

palavras isoladas — palavras pronunciadas isoladamente, com pausas ou silêncios antes e após cada palavra. Este tipo muito restritivo de fala foi o primeiro a ser utilizado no reconhecimento, correspondendo à utilização de modelos baseados em palavras isoladas (modelos de palavras). Por vezes alarga-se esta designação ao uso de palavras compostas ou mesmo a um pequeno número de frases curtas, rigorosamente especificadas. O uso de vocabulários com dimensão cada vez mais elevada determina a substituição dos modelos de palavras por modelos subpalavra. Neste caso, torna-se dispensável a existência das palavras do vocabulário a reconhecer no corpus de treino dos modelos.

Para a obtenção destes últimos, em menor número, não é exigível a disponibilidade de gravações para todas as palavras do vocabulário.

Os modelos de frases inteiras foram utilizados em tarefas de reconhecimento muito simples, para as quais apresentaram um desempenho aceitável. Tal pode ser justificado pela boa representação dos efeitos de coarticulação entre palavras e no interior destas. Contudo, a recolha do material de treino é dispendiosa, não sendo em geral reutilizável para aplicações diferentes. A especificação da aplicação é muito restritiva e qualquer alteração implica um aumento considerável da dimensão do corpus de fala de treino. Além disso, este tipo de reconhecedor obriga o utilizador a usar frases tal como foram definidas nas especificações. Como se verá na subsecção 1.2.2, esta imposição é frequentemente descurada pelo orador.

Os modelos de palavras isoladas são actualmente ainda os mais utilizados em aplicações práticas. Os reconhecedores que utilizam estes modelos requerem quase sempre a existência de pausas entre cada palavra proferida. Em contrapartida obtêm-se bons resultados de reconhecimento nas tarefas simples, com um número reduzido de palavras de comando.

Através de métodos adequados é possível utilizar modelos de palavras no reconhecimento de pequenas sequências de palavras e algumas frases da designada fala ligada. Um caso simples e típico do processamento de fala é o do reconhecimento de numerais cardinais, tal como são pronunciados os números de telefone. Os números extensos são habitualmente segmentados em números de dois ou três dígitos. Deste modo, apenas são necessárias umas escassas dezenas de modelos diferentes de palavras isoladas.

Hoje em dia, na maior parte dos centros de investigação, estudam-se modelos do tipo subpalavra. A combinação destes modelos, por forma a modelarem palavras e frases, permite o reconhecimento de vocabulários de palavras isoladas de elevada dimensão e da fala contínua.

Os reconhecedores podem ainda ser classificados de acordo com outras características, algumas delas porventura menos relacionadas com a evolução tecnológica desta área e mais dependentes do tipo de aplicação. Assim, se um reconhecedor automático de fala se destina ao uso exclusivo de um único orador, poderá ser *dependente do orador*. Se, pelo contrário, se destinar ao uso de um grupo mais ou menos vasto de oradores, em que não é possível identificar cada um de modo a atribuir-lhe um reconhecedor específico, então este reconhecedor deverá ser *independente do orador*. Numa situação intermédia consideram-se os reconhecedores *multi-orador*, destinados a um grupo específico de oradores. Em geral, obtêm-se melhores resultados de reconhecimento quando se treina um reconhecedor para um único orador. Contudo, o esforço requerido ao orador para o treino do "seu reconhecedor" é por vezes excessivo, sobretudo se não estiver devidamente motivado. Além disso, um *reconhecedor dependente do orador* apresenta um desempenho medíocre quando confrontado com qualquer orador diferente daquele a que foi destinado. A solução utilizada nos *reconhecedores independentes do orador* consiste no treino dos modelos com um corpus de fala com um número elevado de oradores, considerados representativos de uma determinada população. Desta forma obtêm-se resultados de reconhecimento aceitáveis com oradores não utilizados no treino do reconhecedor. Ainda assim, apresentam obviamente um desempenho inferior ao dos reconhecedores concebidos exclusivamente para um grupo ou orador específico.

Com a introdução progressiva dos reconhecedores no mercado e a evolução do seu desempenho, exigem-se aplicações cada vez mais elaboradas do tipo independente do orador. Como consequência, aumentam a duração e o número médio de locuções de cada sessão ou acesso individual. Tal facto justifica cada vez mais uma fase inicial de adaptação a cada utilizador a partir de modelos independentes do orador. Esta solução de compromisso, designada por *adaptação ao orador*, tem sido preconizada no sentido de substituir os reconhecedores anteriores (Neto, 1998).

1.1.4 Aplicações da tecnologia existente

O desenvolvimento de determinada aplicação, pressupõe um conhecimento tão rigoroso quanto possível dos utilizadores típicos e do ambiente no qual funcionará o sistema, assim como a existência de uma descrição do respectivo vocabulário. Adicionalmente, com o

conhecimento das regras do diálogo homem-máquina da aplicação, é comum especificar-se uma sintaxe e os subvocabulários correspondentes. Num dado momento de utilização, o reconhecedor emprega apenas um destes subvocabulários: aquele que inclui as palavras permitidas nesse instante do diálogo. A recolha de um corpus adequado ao desenvolvimento de uma aplicação de reconhecimento de fala é uma operação dispendiosa, que justifica todo o empenhamento na especificação do material de fala a ser recolhido.

Tal como acontece com outras tecnologias inovadoras, o reconhecimento de fala debate-se com problemas de adequação às necessidades reais da sociedade neste domínio, nomeadamente, as limitações metodológicas referidas na subsecção anterior, que condicionam de forma muito restritiva o âmbito das aplicações. Assim, a apresentação de produtos comerciais integrando esta tecnologia tem sido, no mínimo, cautelosa. Tal pode ser justificado pelo receio de gorar as expectativas do mercado e criar suspeitas ou mesmo rejeições dos produtos a apresentar no futuro.

Uma questão essencial no desenvolvimento de determinada aplicação reside no facto de o sistema subjacente se encontrar no local do orador (posto local) ou acessível remotamente por via telefónica (posto remoto). O acesso directo ao orador apresenta actualmente algumas vantagens que permitem o desenvolvimento de novas aplicações práticas:

- redução de ruído na comunicação;
- possibilidade de utilização de técnicas de redução de ruído e de eco acústicos, baseadas no uso de dois ou mais microfones (subsecção 1.2.3);
- acesso a outras formas de aquisição de dados. Recolha de dados através de teclado numérico, alfa-numérico ou dedicado, (com figuras alusivas a cada função do sistema) ecrã sensível ao toque, *rato* ou "joystick", sensor de presença, câmara com processamento de imagem, etc. A concepção global deste tipo de interfaces multimodais obedece a especificações de ordem ergonómica e funcional que ultrapassam o âmbito do reconhecimento automático da fala (Morin et al., 1992; Morin e Junqua, 1993; Junqua e Morin, 1996; Benoît e Campbell, 1997; Ciocea et al., 1998);
- acesso a outras formas de apresentação de dados. Utilização de pequenas impressoras ou de um ecrã de vídeo (com o vídeofone e a *internet*, a imagem passa a estar disponível também por acesso remoto);
- possibilidade de adquirir ou de fornecer bens e serviços, tais como: receber ou fornecer documentos ou valores (postos do tipo *multibanco*); emitir bilhetes para transportes ou espectáculos; dar acesso a máquinas de jogos; fornecer os produtos tradicionais de quiosque (por exemplo, jornais e tabaco); etc.

A utilização da rede telefónica pública para aceder a um serviço de atendimento automático, constitui uma alternativa mais restritiva em termos de aplicações mas também menos dispendiosa do que a criação de vários postos locais. Os serviços actualmente disponíveis são accionados na sua grande maioria através da codificação do teclado do telefone tradicional pelo sistema DTMF ("Dual Tone Multifrequency")[2]. Este teclado pode continuar a ser útil mesmo quando se dispõe de uma sofisticada interface de fala, nomeadamente, na introdução de números, uma vez que o utente se encontra particularmente treinado nesta tarefa.

Os postos remotos permitem serviços de despertar, marcação de reservas para transportes e espectáculos, algumas operações bancárias e uma grande variedade de serviços de informação. A expansão actual das redes de comunicações móveis permite antever uma utilização mais eficaz de alguns destes serviços (por exemplo, informação de tráfego). O conceito de serviço de informação pode alargar-se à consulta de sofisticadas bases de dados, embora na actualidade, a maioria das aplicações com reconhecimento de fala se destinem a tarefas consideravelmente mais simples: consulta de mensagens de meteorologia, de movimento de capitais, de espectáculos, de resultados desportivos, etc.

O reconhecimento de fala independente do orador é aplicado, principalmente, nos sistemas previstos para uma utilização de curta duração, por parte de uma população de oradores não identificados. Por sua vez, os reconhecedores automáticos da fala dependentes do orador têm sido incorporados em aplicações utilizadas de forma sistemática e duradoura por um orador ou um conjunto restrito de oradores. Em geral, garantem a interacção com programas de computador especializados, tais como: editores gráficos; sistema operativo (comandos mais utilizados, tais como a gestão de ficheiros); registo de texto ditado, acesso a bases de dados e controlo de robots e maquinaria especializada. A utilização do reconhecimento de fala possibilita uma interacção mais natural, permitindo maior mobilidade ao utilizador, nomeadamente libertando as mãos do teclado para outras tarefas. No caso dos utilizadores com deficiências motoras, invisuais ou iletrados estas aplicações são indispensáveis[3].

O computador pessoal, que actualmente serve de suporte à maioria das aplicações computacionais, pressupõe a utilização por uma única pessoa. Em geral, este utilizador investe algum esforço na personalização do seu ambiente de trabalho de modo a obter maior rendimento nas tarefas específicas que lhe estão atribuídas. Neste caso, é de esperar

[2]Em alguns países, entre os quais Portugal, este tipo de serviços não se encontra muito disseminado, o que se poderá ficar a dever à introdução, ainda recente, da rede digital.

[3]A percepção destas dificuldades foi tida em consideração no desenvolvimento do novo sistema operativo Windows 98, o qual dispõe da possibilidade de ser utilizado através de um reconhecedor de fala (Huang et al., 1995; Huang, 1998) ("URL: http://research.microsoft.com/stg/").

uma atitude colaborante no treino ou adaptação do reconhecedor no sentido de o tornar dependente do orador. O esforço exigido nesta fase é compensado pelo uso continuado do mesmo, com taxas de reconhecimento elevadas na execução de numerosos e repetidos comandos.

As aplicações lúdicas, tais como os jogos electrónicos, não reunem por agora, condições para se configurarem como produtos comercializáveis. Este tipo de aplicações pressupõem pouco esforço da parte do utilizador, quer financeiro quer no próprio produto, pelo que contam em geral com reconhecedores simples do tipo independente do orador (ex: comando remoto de vídeo-gravador (Kuwano et al., 1992)).

Um ambiente acústico que tem sido recentemente alvo de muitos estudos é o da cabina de viaturas. A importância do reconhecimento de fala neste ambiente é sublinhada pela existência de projectos de grande dimensão, suportados com fundos públicos e privados tais como o projecto VODIS[4] (Compernolle, 1997; Trancoso e Viana, 1997) e o projecto MoTiV[5] (Haeb-Umbach, 1997). Estes projectos têm por objectivo o melhoramento das condições de comunicação com um sistema de reconhecimento de fala local ou remoto (através de telefone móvel) e as condições de conforto acústico que permitam o uso convencional desse telefone, a conversação entre os vários ocupantes da cabina ou a audição de música. Para tal têm sido desenvolvidas técnicas de processamento de sinal, nomeadamente de cancelamento passivo ou activo do ruído, com vista a este ambiente acústico (secção 1.2.3).

As futuras aplicações de sistemas com reconhecimento de fala nos automóveis, visam essencialmente a substituição de manípulos de controlo de equipamentos acessórios, não directamente relacionados com a condução do veículo. Elementos como o volante, o travão, o acelerador, etc..., resultam de um processo evolutivo intensamente testado. Muitos destes elementos exigem uma actuação quase contínua e como tal pouco adequada

[4]O projecto VODIS (sigla de "Advanced speech techonologies for Voice Operated Driver Information Systems") é parcialmente financiado pelo sector da "Language Engineering" no âmbito do programa da Comissão Europeia (DG XIII) "Telematics Applications of of Common Interest". Com início nos finais de 1995 tem uma duração prevista de 39 meses ("URL: http://werner.ira.uka.de/VODIS/"). Neste projecto participam as seguintes entidades: Robert Bosch GmbH, Multicom Research Inc., RWTH–Aachen — Institut für Nachrichtengeräte und Datenverarbeitung, Volkswagen AG, Renault Research Innovation, PSA Peugeot Citroën, Lernout & Hauspie Speech Prod., Universidade de Karlsruhe, Center for Research on User-System Interaction/IPO, INESC.

[5]O projecto "Mobilität und Transport im intermodalen Verkehr" (MoTiV) é financiado pelo Ministério Alemão da Educação e Investigação e nele participam as seguintes empresas: Adam Opel AG, BMW AG, Daimler–Benz AG, debis Systemhaus GEI, Deutsche Telekom MobilNet GmbH, IBM Deutschland GmbH, ITF Intertraffic GmbH, MAN Nutzfahrzeuge AG, Philips Car Systems GmbH, Robert Bosch GmbH, Siemens AG, Volkswagen AG ("URL: http://www.tuev-rheinland.de/tsu/IfUE/pt_bvt/motiv/haupts.htm").

ao controlo for fala. A experiência com estes elementos é, para maior parte das pessoas, anterior ao processo formal da aprendizagem da condução, identificando-os com a própria imagem do automóvel. Assim, por motivos de ergonomia, de segurança ou apenas de padronização, estes elementos dificilmente podem ser substituídos. Nos equipamentos considerados acessórios incluem-se os vidros com comando eléctrico, a climatização, o telefone móvel, os espelhos retrovisores, as fechaduras das portas, o rádio, o leitor de cassetes ou de discos compactos e o televisor. Recentemente começaram também a ser introduzidos sistemas de navegação que conjugam informação cartográfica disponível num disco compacto com a localização geográfica do automóvel fornecida por satélite (GPS — "Global Positioning Satellite/System"). Neste caso, a utilização do reconhecimento de fala destina-se essencialmente à inicialização do sistema que necessita de conhecer o destino da viatura. No entanto, pode ser necessário proceder a alterações durante o percurso, nomeadamente devido a condições de tráfego. Em alternativa aos dados fornecidos num ecrã de vídeo prevê-se a utilização de um sintetizador de voz (Compernolle, 1997).

1.2 Factores que degradam o desempenho dos reconhecedores

O desenvolvimento de um sistema automático de reconhecimento de fala é extremamente dificultado pela variabilidade do sinal de fala. Esta variabilidade encontra-se associada à grande capacidade de veicular informação relevante para o ouvinte humano, assim como, à contaminação do sinal por diversos factores associados às condições ambientais em que o sinal é produzido e transmitido. Parte da referida informação, tal como a entoação, o tom de voz, o estilo do discurso e o sotaque, são em geral descurados no reconhecimento automático de fala. Actualmente, apenas o conteúdo textual do sinal de fala é objecto da respectiva descodificação por via automática, isto é, a que resulta da transcrição tal como num ditado normal.

Nesta secção descrevem-se alguns dos factores de variabilidade do sinal de fala que mais contribuem para a degradação do desempenho no reconhecimento automático. Estes factores podem ser considerados a diversos níveis, nomeadamente: o das características do orador (subsecção 1.2.1); o do conteúdo textual do sinal de fala (subsecção 1.2.2); o das condições ambientais em que é produzida a fala (subsecção 1.2.3).

Estes factores interagem entre si de forma por vezes complexa no acto da produção do sinal de fala. Contudo, na análise preliminar aqui apresentada, serão apenas descritos isoladamente. Perante esta panorâmica do universo incomensurável dos sinais de fala,

compreende-se a necessidade de restringir, tanto quanto possível, a influência de alguns destes factores, por forma a obterem-se modelos de complexidade e dimensão aceitáveis. Assim, a generalidade das aplicações existentes referem-se a um universo de oradores, de ambientes acústicos ou mesmo de equipamentos muito limitados, o que, ainda assim, nem sempre permite uma correspondente simplificação do problema.

O estudo destes problemas tem sido enquadrado numa área designada por *processamento robusto de fala* (Teixeira et al., 1993b; Furui, 1997). Uma definição de um sistema robusto, seria a de um sistema que desempenha as suas funções de forma aceitável, em circunstâncias não previstas pelo seu autor. Esta definição é porventura demasiado radical para o âmbito do reconhecimento de fala, no qual as dificuldades se mantêm, mesmo em relação a factores de degradação razoavelmente conhecidos. De facto, mesmo em relação as estes, não foi possível encontrar estratégias que permitam a atenuação, em simultâneo, dos respectivos efeitos. Na prática designam-se por *métodos robustos*, aqueles que reduzem a diferença entre o desempenho de um sistema testado em condições semelhantes às que determinaram o seu desenvolvimento (treino) e o desempenho do mesmo sistema nos designados *testes de campo*, ou seja, nas circunstâncias reais em que o sistema é útil.

1.2.1 Variabilidade das características do orador

No reconhecimento automático, os aspectos da variabilidade do sinal de fala exclusivamente devidos às características do orador são considerados separadamente em duas classes essenciais: a variabilidade intra-orador e a variabilidade interorador. Nos reconhecedores dependentes do orador, interessa essencialmente atenuar os efeitos da primeira, enquanto que nos reconhecedores independentes do orador interessa atenuar a segunda.

A variabilidade intra-orador refere-se a variações temporais das características de um dado orador. Estas são devidas a alterações de dois tipos:

físicas — uma simples constipação altera significativamente os padrões da fala. O surgimento de outras patologias temporárias pode conduzir a casos mais extremos (até à afonia);

emocionais — as alterações do estado emocional do orador ocorrem com mais frequência e mais rapidamente do que as do tipo físico. O estudo da fala sob condições de stress tem sido objecto de diversos estudos com vista, nomeadamente, a aplicações militares (Trancoso e Moore, 1995). Neste contexto, o aspecto mais estudado refere-se às alterações na produção de fala devidas à presença de ruído — efeito de Lombard (Clary e Hansen, 1992).

A variabilidade interorador pode ser relacionada com as inúmeras formas de classificar ou diferenciar os seres humanos, em termos físicos, psicológicos, comportamentais, sociais, económicos, religiosos, políticos, geográficos, etc.. Todas estas categorizações impõem características específicas ao processo de produção de fala que serão identificáveis no respectivo sinal com diversos graus de sucesso (excepção óbvia para os incapacitados de se expressarem oralmente). Entre os factores de variabilidade interorador mais relevantes para o reconhecimento automático destacam-se a idade, o sexo, o peso, o hábito de fumar, o nível cultural, o sotaque, o dialecto, etc. No caso da língua utilizada não ser a língua materna do orador, ou de esta não se encontrar bem definida, (bilíngues) acrescem ainda outros factores tais como a experiência linguística, a capacidade de imitação, etc. As diferenças na fala produzida por diferentes oradores estão portanto relacionadas, não só com a configuração do seu aparelho fonador, como com toda uma série de hábitos linguísticos.

No processo de comunicação por fala o código utilizado é a língua. Ao se considerar um orador capaz de falar duas ou mais línguas é essencial determinar qual delas utiliza em determinado momento. A identificação automática da língua tem sido recentemente objecto de diversos estudos de investigação (House e Neuburg, 1977; Hazen e Zue, 1993; Zissman, 1993; Muthusamy et al., 1994a; Berkling et al., 1994; Zissman, 1995; Caseiro, 1998). Neste trabalho estudam-se sinais de fala de uma única língua mas que são contaminados por características de diversas línguas distintas e conhecidas, sendo os resultados analisados essencialmente com base nesta distinção. Os trabalhos que procuram relacionar aspectos entre duas ou mais línguas distintas são em geral englobados nas áreas de investigação designadas por *multilíngua* ou *interlíngua* ("cross-language").

Sotaque de oradores estrangeiros

Um ouvinte atento é em geral capaz de detectar um orador estrangeiro. Em muitos casos, é possível reconhecer a origem desse orador, ou mais precisamente, qual a sua língua materna. Grande parte dos oradores estrangeiros apresentam um conhecimento deficiente ou pouco treino com a sua segunda língua. O processo de aprendizagem da língua estrangeira é em geral influenciado por inúmeros factores entre os quais se destacam a motivação e determinadas capacidades intrínsecas do aluno, tais como as de imitação. Mesmo aqueles que dominam a língua escrita, podem apresentar problemas de pronúncia que impeçam um ouvinte nativo de entender o que dizem (McAllister, 1995; McAllister, 1998).

O problema do sotaque de oradores estrangeiros classificados de acordo com a sua

língua materna enquadra-se adequadamente nos factores de variabilidade interorador. Tal como acontece com outros destes factores, determina uma forte degradação no desempenho dos reconhecedores automáticos de fala, na ausência de qualquer precaução que permita atenuar este efeito negativo (Teixeira e Trancoso, 1992; Teixeira et al., 1997; Byrne et al., 1998).

O problema dos oradores estrangeiros no reconhecimento de fala só foi equacionado recentemente e as soluções disponíveis não são diferenciadas das existentes para o problema da independência do orador. Este problema assume particular interesse nas aplicações referentes a serviços automáticos de informação nos postos de turismo, de reserva ou compra de bilhetes em aeroportos, estações centrais ferroviárias e de camionagem, em países onde confluem grande variedade de nacionalidades por via do turismo ou da imigração (por exemplo, nos E.U.A. e em alguns países da Europa). Em serviços semelhantes, oferecidos através da rede telefónica pública internacional, este tipo de problemas poderão ocorrer ainda com maior incidência e gravidade.

A maioria destas aplicações foram desenvolvidas para uma única língua e o seu uso por oradores estrangeiros causa uma quebra significativa no seu desempenho, comprometendo a utilidade dos respectivos sistemas. Este problema tem maior incidência quando a língua escolhida para o vocabulário do reconhecedor é uma língua franca tornando-o deste modo acessível a um grande número de oradores de diversas nacionalidades. O inglês é a língua franca mais utilizada no mundo. É a primeira escolha para muitos oradores que pretendem utilizar um serviço de atendimento público no estrangeiro, esperando desta forma obter uma resposta mais rápida e eficaz. Se o receptor for capaz de identificar a origem do sotaque manifestado, poderá optar por solicitar outro receptor capaz de comunicar na língua materna do cliente. Num sistema automático, isto equivale a seleccionar um reconhecedor de fala para essa língua ou, no caso de não existir um disponível, outro reconhecedor também de inglês mas adaptado ao sotaque apresentado pelo cliente. É neste contexto que surge um problema novo, o da identificação do sotaque estrangeiro (Teixeira et al., 1996).

1.2.2 Texto da mensagem oral

A maioria das aplicações do reconhecimento de fala permitem apenas uma interacção simplificada, baseando-se num encadeamento pré-especificado de associações pergunta-resposta. As respostas permitidas são geralmente apresentadas em pequenas listas de palavras isoladas. Considere-se, por exemplo, uma dessas listas com as palavras de comando: começar, parar, ajuda, ligar, etc. Na realidade o utente tem tendência a incluir

palavras adicionais, tais como, por exemplo: "queria ligar por favor". Um reconhecedor que não esteja preparado para esta situação, identificará cada uma destas palavras com palavras existentes no seu vocabulário. Esta substituição de palavras pode ter consequências graves no desempenho global da aplicação. Este fenómeno pode ocorrer também, de forma semelhante, no reconhecimento de fala ligada ou contínua, quando o orador utiliza uma construção frásica não prevista no reconhecedor.

A solução tradicional para este tipo de problema tem sido a utilização de um modelo designado *de escoamento* ou *de lixo* (do inglês "sink" ou "garbage") que é considerado em simultâneo com os modelos das palavras previstas no vocabulário da aplicação. Os diferentes modelos são comparados com cada palavra de entrada através de uma medida de distância ou de verosimilhança. Uma palavra não pertencente ao vocabulário da aplicação (designada por *palavra estranha*) deverá apresentar um valor mínimo para essa medida quando for comparada com o modelo de escoamento. Desta forma, o reconhecedor desprezará a referida palavra que é considerada irrelevante para a aplicação.

A forma de interacção aqui descrita é muito limitativa, tornando-se enfadonha ou mesmo ineficaz para as aplicações mais complexas. No futuro, os reconhecedores para fala contínua e espontânea deverão utilizar modelos de interacção construídos com base em resultados da análise semântica e pragmática do sinal de fala, bem como de outras fontes de conhecimento (Cole et al., 1995).

1.2.3 Condições ambientais

A captação do sinal de fala tal como este é produzido, de acordo com a intenção e capacidade do orador humano, representa apenas uma situação ideal que só se consegue atingir aproximadamente em condições laboratoriais especiais (numa câmara anecóica). Em geral, incorporam-se nesse sinal ideal outros sinais devidos a reverberações e a ruídos produzidos por outras fontes acústicas ou pelo próprio canal de transmissão.

Devido às suas características particulares, o eco é um fenómeno distinto do ruído. Consegue-se obter o cancelamento do eco de forma eficaz, nomeadamente quando originado por uma única fonte (Serralheiro et al., 1991; Dahl et al., 1997; Kellermann, 1997). Nas aplicações do reconhecimento procura-se em geral fornecer informações ao utilizador também através da fala utilizando mensagens pré-gravadas ou de um sintetizador de fala produzida a partir de texto (Dudley, 1939; Oliveira, 1996; Teixeira e Vaz, 1997). Devido ao eco, o sinal de fala à saída do sistema de síntese é realimentado na entrada do reconhecedor com diversos graus de atenuação e de atraso no tempo. Este ecos têm obviamente

consequências desastrosas no desempenho do reconhecedor.

Nos sinais de fala recolhidos através da rede telefónica pública pode surgir ruído eléctrico, habitualmente designado por *ruído do canal* ou por *ruído de linha*, assim como interferências e distorções do sinal. Além disso, existem ecos (não acústicos) que são geralmente atenuados por equipamentos da própria rede. O ruído acústico captado no telefone e em particular no telefone móvel, tem origem em diversas fontes pelo que é difícil considerar estratégias globais para o seu cancelamento. No caso das comunicações móveis existem ainda problemas específicos quer devido à necessidade de codificação digital do sinal de fala quer devido ao problema da captação hertziana do sinal, sempre sujeita a interferências e ao surgimento de réplicas do mesmo sinal ("multi-path").

O ruído acústico é habitualmente caracterizado em função de determinados ambientes típicos seguidamente descritos. Nos espaços públicos no exterior coexistem vários ruídos sobrepostos tais como o de veículos motorizados na via pública e o burburinho citadino. Nas salas de grandes dimensões, onde aflui um grande número de pessoas, surgem ainda importantes efeitos de reverberação. Numa estação de comboios, por exemplo, existem múltiplas fontes de ruído, distantes umas das outras e que sofrem reverberações importantes. O sinal resultante de todas estas componentes pode, em determinadas circunstâncias, ser considerado quase-estacionário. Por outro lado, o ambiente de escritório e o doméstico correspondem a salas de dimensões inferiores onde proliferam ruídos, habitualmente do tipo impulsivo, gerados por fontes muito próximas: máquinas registadoras, impressoras, teclados, telefones, campainhas, electrodomésticos, ou mesmo fala de outros oradores.

O uso de motores mais potentes e mais rotativos acentua os problemas causados com o ruído no interior da cabina dos automóveis. Este problema tem sido estudado de forma integrada com o ruído causado por outros componentes da viatura, pelo tráfego no exterior e pela necessidade de utilização do telefone móvel e da audição do rádio ou do disco compacto. Conforme seguidamente se descreve, a atenuação destes ruídos requer a respectiva recolha através de microfones adicionais quer no interior da cabina, quer no exterior da viatura e no compartimento do motor. Podem ainda ser utilizados sinais não acústicos, como por exemplo, os disponíveis no circuito de ignição, os quais fornecem uma indicação precisa sobre o regime de rotação do motor.

Entre as técnicas de redução de ruído conhecidas destacam-se:

subtracção espectral — particularmente eficaz na atenuação de ruído quase-estacionário (Boll, 1979; Compernolle, 1989; Silva, 1989; Teixeira e Trancoso, 1990; Trancoso et al., 1990; Serralheiro et al., 1991; Teixeira e Trancoso, 1991a; Teixeira e Trancoso, 1991b; Meyer e Simmer, 1997);

cancelamento adaptativo de ruído — utiliza dois ou mais microfones, um para captar o sinal de fala corrompido por ruído e os restantes para captarem o ruído (Widrow et al., 1975; Harrison et al., 1984; Martins et al., 1990; Teixeira et al., 1993b; Martins, 1998b).

Os ruídos de baixa frequência podem ser atenuados acusticamente com interferências destrutivas, produzidas por fontes acústicas secundárias. Também neste caso, é importante utilizar um ou mais microfones secundários. Desta forma, conseguem-se atenuações típicas superiores a 10dB numa zona com um raio de cerca de um décimo do comprimento da onda acústica, em redor de um único destes microfones. Esta técnica, designada por *cancelamento activo de ruído*, é adequada a espaços de pequena dimensão, tais como o da cabina dos automóveis, dos aviões a jacto ou de helicópteros (Elliot e Nelson, 1993; Lopes et al., 1998).

De referir ainda que existem técnicas baseadas em agregados de microfones ("beamforming microphone arrays") que permitem identificar a posição de determinadas fontes acústicas, por exemplo, um orador específico num auditório (Lin et al., 1994; Kellermann, 1997).

Este tipo de soluções são contudo pouco efectivas quando o ruído é do tipo impulsivo. Numa perspectiva mais vasta e integrada com o próprio reconhecedor, surgem alternativas como a da generalização dos HMMs convencionais para uma decomposição óptima de processos simultâneos (Varga e Moore, 1990). Com o uso de modelos perceptuais, que resultam do modelamento dos fenómenos acústicos fisiológicos e psicológicos que ocorrem no ouvinte humano, têm-se conseguido melhorias no desempenho dos sistemas de reconhecimento (Hermanski, 1990b; Hermanski, 1990a; Perdigão, 1997).

1.3 A engenharia da linguagem oral

A investigação do reconhecimento automático da fala congrega e partilha conhecimentos de muitas áreas directa ou indirectamente relacionadas com o sinal da fala. No texto seguinte, procura-se fornecer uma panorâmica do carácter interdisciplinar desta área de investigação.

Quer como instrumento de comunicação, quer como forma de representação do conhecimento, a linguagem manifesta-se sob duas formas: escrita ou oral. Idealmente, estas duas formas seriam equivalentes no sentido de vincularem a mesma informação, contudo tal só excepcionalmente se verifica. Em termos de engenharia existem dois problemas es-

senciais a resolver que estão directamente relacionados com a necessidade de se converter uma destas manifestações da linguagem na outra: o reconhecimento e a síntese de fala. Noutras circunstâncias, pretende-se utilizar uma representação eficiente para o sinal de fala, minimizando as diferenças perceptuais entre o sinal original e o sinal que é possível sintetizar a partir dessa representação. Este é o problema da codificação da fala.

Os problemas referidos correspondem a áreas de investigação bem delimitadas e com algumas metodologias próprias. Contudo, tendo por referência comum o sinal de fala, cada área partilha de muitas das metodologias utilizadas nas outras áreas, como por exemplo: a representação espectral, os modelos de predição linear, a quantificação vectorial, etc.

A síntese de fala representa, pelo menos, uma necessidade para muitas das aplicações descritas utilizando o reconhecimento de fala automático. Quando se tem de falar com uma máquina espera-se uma resposta também oral. Esta necessidade torna-se natural e indispensável para as aplicações mais sofisticadas que utilizam a rede telefónica pública. Conforme se referiu na subsecção 1.2.3, as actuais aplicações do reconhecimento de fala são relativamente simples, sendo a síntese automática de fala a partir de texto substituída com eficácia, em muitos casos, por mensagens pré-gravadas.

A codificação da fala é um problema típico das telecomunicações que tem por objectivo a transmissão ou o armazenamento do sinal de fala de forma económica e segura (Trancoso, 1987; Marques et al., 1990; Ribeiro, 1991; Abrantes, 1992). Trata-se, em geral, de um problema de redução da largura de banda utilizada para a transmissão do sinal. Uma definição preliminar de um codificador ideal poderia ser concretizada com a existência de um reconhecedor e um sintetizador de fala ideais, associados em série. Esta definição só poderia estar certa se a informação contida na linguagem escrita fosse equivalente à do sinal de fala de onde foi extraída. A codificação de fala é utilizada no canal entre dois interlocutores humanos que precisam ter conhecimento, a todo o momento, de dados que lhe permitam identificar o outro interlocutor e o seu estado emocional. Este tipo de dados não existe numa transcrição vulgar do sinal de fala. Ainda assim, a integração de metodologias do reconhecimento e da síntese de fala, perspectiva uma nova geração de codificadores de fala.

O problema do reconhecimento da fala, tema central deste trabalho, encontra soluções essencialmente nas áreas do reconhecimento de padrões e do processamento de sinais, sendo objecto de inúmeros trabalhos e publicações de reconhecido valor nestas áreas. O carácter interdisciplinar deste tema abrange, contudo, muitas outras disciplinas.

O reconhecimento de padrões contribui com os métodos para o agrupamento de dados por forma a determinar modelos representativos desses mesmos dados. Esse

agrupamento é por sua vez função de determinadas medidas de distância para comparação entre esses modelos e os próprios dados. Recentemente, ganham relevância as técnicas associadas às redes neuronais (Lippmann e Gold, 1987; Almeida, 1993; Cook e Robinson, 1995; Neto, 1998).

O processamento de sinais deverá providenciar a informação relevante do sinal de fala, de forma robusta e eficiente (nomeadamente em tempo real). Os parâmetros de natureza espectral são muito utilizados para caracterizar as propriedades de variação no tempo do sinal de fala. A obtenção deste tipo de parâmetros de forma eficaz é objecto do estudo do processamento de sinais. Ainda neste contexto, encontram-se métodos de melhoramento do sinal, tais como os de redução de ruído.

A física encontra-se representada nesta área através da acústica e da sua relação com a fisiologia da produção e da percepção do sinal. O conhecimento destes aspectos tem sido explorado sob diversas perspectivas. Talvez a mais continuada tenha sido a de, na medida do possível, imitar os mecanismos fisiológicos da percepção, nomeadamente os que ocorrem no ouvido até à conversão electroquímica para os nervos auditivos.

A teoria da informação e da comunicação contribui com métodos para a obtenção de estimativas para os parâmetros dos modelos estatísticos e com métodos de detecção de determinados padrões no sinal de fala.

A linguística contribui com todo o conhecimento sobre o sinal de fala, a começar na fonologia, na relação entre as palavras com a sintaxe, até ao seu significado e sentido, com a semântica e a pragmática. Estes últimos aspectos são também comuns à linguagem escrita e são objecto de estudo da disciplina da *compreensão da linguagem natural*, habitualmente classificada na área da *inteligência artificial*.

A informática e a ciência da computação possibilitaram a implementação dos actuais e sofisticados métodos de busca utilizados no reconhecimento de fala. Contudo, estas áreas beneficiam essencialmente dos sucessivos avanços da microelectrónica na tecnologia dos semicondutores, que tem permitido uma vulgarização crescente da utilização de memórias e de processadores de grande capacidade e rapidez.

A psicologia desempenha actualmente um papel importante na adequação das aplicações e respectivas interfaces com o utilizador humano. Existem ainda aspectos relacionados com a aprendizagem da língua, a percepção da fala, o estado emocional, etc., que devem ser considerados no desenvolvimento dos modelos e das aplicações de reconhecimento.

1.4 Objectivos e estrutura deste trabalho

1.4.1 Objectivos

No início deste trabalho procurou-se desenvolver aplicações que demonstrassem o interesse prático do reconhecimento automático de fala. Para este fim e considerando a tecnologia existente, utilizou-se um reconhecedor baseado em modelos HMM de palavras isoladas (SIRtrain, 1991; Jacobsen, 1992). Uma vez que estas aplicações seriam avaliadas por supervisores de diversos países e se destinavam a demonstrações em conferências e outros encontros internacionais, a língua escolhida foi o inglês (Irion et al., 1992; Teixeira et al., 1993b).

O primeiro problema detectado nestas aplicações foi o de, juntamente com as palavras do vocabulário da aplicação, o utilizador ter tendência a pronunciar outras palavras, por vezes no contexto de um diálogo com outra pessoa. Este facto tinha consequências graves no desempenho global dessas aplicações. Assim, o primeiro objectivo do presente trabalho foi o de modificar o reconhecedor de fala por forma a detectar e rejeitar, de forma o mais eficiente possível, as designadas palavras estranhas.

Posteriormente, verificou-se que o desempenho do referido reconhecedor se degradava de forma significativa quando a língua materna dos utilizadores não era o inglês (Teixeira e Trancoso, 1992). Na sequência desta constatação procurou-se obter reconhecedores com um desempenho aceitável, para oradores nativos e não nativos.

Entre outras restrições deste trabalho e que serão mencionadas em local próprio, sublinham-se algumas de carácter geral que delimitam o âmbito deste estudo. Tal como na maioria dos estudos experimentais sobre a fala, estas restrições estão relacionadas com a representatividade do corpus de fala utilizado. O corpus de sinais de fala utilizado neste trabalho foi obtido a partir de uma população de oradores composta por indivíduos de ambos os sexos, provenientes de diversos países do espaço da União Europeia: Alemanha, Dinamarca, Espanha, Itália, Reino Unido e Portugal. O vocabulário deste corpus resulta da tradução inglesa de diversos vocabulários utilizados em protótipos de aplicações com reconhecimento automático de fala, desenvolvidos por importantes companhias europeias no sector das telecomunicações. O espaço europeu representa um mercado potencial para todos os produtos com forte componente tecnológica tais como aqueles que podem incorporar reconhecedores de fala. A língua inglesa é por sua vez a língua franca mais utilizada neste espaço. Estes factos justificam a importância do referido corpus de fala e do trabalho desenvolvido com o mesmo, tanto mais que, na altura em que este trabalho foi iniciado, era o único corpus disponível com estas características (secção 3.2).

1.4.2 Organização da presente dissertação

Este trabalho encontra-se organizado de acordo com a descrição seguinte.

Na presente **introdução** apresentaram-se os objectivos e o contexto em que se inserem. Referiram-se resumidamente algumas das soluções mais conhecidas dos problemas inerentes ao reconhecimento automático da fala. Por fim, perspectivaram-se os capítulos seguintes e salientaram-se os aspectos mais relevantes do presente trabalho.

O **segundo capítulo** descreve a aplicação no reconhecimento da fala dos modelos de Markov não observáveis. Esta descrição centrou-se nos modelos e métodos utilizados nas experiências realizadas no âmbito desta dissertação. São descritos os procedimentos e os respectivos fundamentos, empregues na extracção de características do sinal de fala, relevantes para o reconhecimento automático. São apresentadas duas formulações dos modelos de Markov não observáveis, adequadas ao sinal da fala: uma, mais convencional, baseada em probabilidades de observação contínuas e uma segunda formulação baseada nas designadas probabilidades de observação semicontínuas. Os modelos baseados em elementos subpalavra são essenciais para o reconhecimento de vocabulários de grande dimensão e da fala contínua. Descrevem-se diversos tipos de elementos subpalavra, os compromissos envolvidos na sua escolha e a forma de os obter e utilizar. Faz-se referência aos modelos linguísticos, introduzindo alguns conceitos que serão utilizados nos capítulos seguintes. Por último, definem-se alguns parâmetros utilizados na avaliação dos resultados do reconhecimento de fala.

O **terceiro capítulo** é dedicado à descrição da recolha e do conteúdo do corpus de fala que é utilizado nos capítulos seguintes. São ainda apresentadas outros corpora relevantes para a área em que se insere este trabalho.

No **quarto capítulo**, descreve-se o estudo efectuado sobre o problema da rejeição das palavras estranhas, ou, de outro modo, da detecção das palavras-chave. Procurou-se verificar a vantagem do uso de vários modelos de escoamento em simultâneo. Determinaram-se experimentalmente as condições em que os modelos de escoamento múltiplos devem ser obtidos e utilizados, por forma a serem mais eficazes (Teixeira et al., 1992).

No **quinto capítulo**, descrevem-se diferentes estratégias com vista à atenuação dos efeitos negativos nas taxas de reconhecimento, causados por oradores estrangeiros. Para os reconhecedores mais avançados, baseados em modelos subpalavra, desenvolveu-se um modelo capaz de representar em termos probabilísticos as variações de pronúncia, a partir de um conjunto finito de modelos subpalavra e de um conjunto de repetições de uma dada palavra. Descreve-se este modelo, a forma de determinar os seus parâmetros e os

resultados com ele obtidos (Teixeira et al., 1997). Faz-se também a descrição e comparação com outros métodos da bibliografia que visam a obtenção automática de transcrições do sinal da fala em elementos subpalavra.

O **sexto capítulo** aborda o mesmo problema do capítulo anterior mas na perspectiva de o decompor em dois subproblemas distintos. De um lado, ficam os reconhecedores específicos, obtidos de forma a garantirem o melhor desempenho possível para cada sotaque. Do outro, fica a tarefa de identificar, em função de um determinado orador, qual destes reconhecedores deve ser seleccionado. Este problema é resolvido de forma simples, integrada no mecanismo de reconhecimento. Apresentam-se resultados da identificação de sotaques e dos correspondentes desempenhos do reconhecimento de fala (Teixeira et al., 1996).

O sétimo e **último capítulo** procura sintetizar as conclusões gerais desta dissertação. Adicionalmente referem-se algumas opções a considerar em desenvolvimentos futuros.

1.5 Contribuições originais

Esta dissertação procurou dar um contributo válido para o problema central da robustez e da fiabilidade da nova tecnologia emergente que é o reconhecimento automático de fala. A abordagem directa do problema dos oradores estrangeiros no reconhecimento de fala só se veio a verificar muito recentemente com a maturidade desta tecnologia e nomeadamente com a chegada ao mercado dos seus primeiros produtos. As contribuições originais apresentadas nesta dissertação são as que se enumeram a seguir.

O **corpus de fala** utilizado neste trabalho apresenta características únicas à data da sua recolha (Irion et al., 1992; Teixeira et al., 1993b; Teixeira et al., 1996). A definição, a gravação dos oradores portugueses e a compilação dos sinais, anotações e dados dos oradores foi efectuada no âmbito desta tese (secção 3.2).

O primeiro problema estudado sobre este corpus, foi o da rejeição de palavras estranhas ao vocabulário do reconhecedor (capítulo 4). Desde logo, houve necessidade de abordar os problemas derivados da existência de oradores estrangeiros. Além disso, testou-se com sucesso a possibilidade de se obterem vantagens com a utilização de **modelos de escoamento múltiplos** e determinaram-se as condições em que tal acontece. À data da publicação destes resultados (Teixeira e Lindberg, 1992; Teixeira e Trancoso, 1992; Teixeira et al., 1992), era convicção de um número significativo de investigadores desta área que apenas um único modelo de escoamento poderia ser útil nas tarefas de rejeição

de palavras.

O problema central deste trabalho é o do reconhecimento automático dos sinais de fala de oradores estrangeiros (capítulo 5). Nas experiências preliminares de aferição dos modelos e dos métodos, obtiveram-se resultados práticos satisfatórios com os reconhecedores de palavras isoladas mais simples. Para tal, efectuou-se uma extensão do conceito de reconhecedor independente do orador, ao caso particular dos oradores estrangeiros. Em alternativa a esta estratégia associaram-se diversos reconhecedores, específicos de cada sotaque, num único reconhecedor. Estas estratégias apresentaram resultados equivalentes e foram aplicadas a reconhecedores baseados em modelos subpalavra. Contudo, procurou-se explorar o facto destes modelos oferecerem o detalhe fonético de forma mais explícita, tendo em vista a incorporação de outros tipos de informação, nomeadamente ao nível da transcrição fonética. Pretendeu-se, desta forma, melhorar significativamente o desempenho dos reconhecedores nas referidas condições e simultaneamente contribuir de alguma forma para um melhor conhecimento do fenómeno dos sotaques. Numa primeira tentativa, procurou-se incorporar algumas regras conhecidas de pronúncia da língua nativa referente a cada sotaque, na expectativa de que estas transitassem de forma significativa para a segunda língua. Na maioria dos casos, essas regras não foram possíveis de verificar por inspecção auditiva directa do sinal, nem através de alguma melhoria significativa dos reconhecedores que as incorporaram.

Concebeu-se um modelo que permite descrever de forma probabilística as transcrições de cada palavra em termos de fones, a partir de um corpus de fala. Esta descrição pretende substituir as transcrições fonémicas obtidas de léxicos de pronúncia convencionais e utilizadas no reconhecimento com modelos subpalavra. Desenvolveu-se e testou-se um método que permite calcular os parâmetros deste modelo, por forma a ser utilizável num reconhecedor automático. Embora com algumas limitações práticas, os resultados obtidos são considerados positivos (Teixeira et al., 1997).

Por último, realizaram-se experiências de identificação automática do sexo e do sotaque (capítulo 6) que procuraram, nomeadamente, identificar o papel de cada um destes factores e as relações entre si no contexto do reconhecimento de fala. Os resultados destes trabalhos foram também publicados (Teixeira e Trancoso, 1993; Teixeira et al., 1996).

1.6 Conclusões

Este capítulo foi destinado à apresentação da motivação, dos objectivos e da estrutura do presente trabalho. Introduziu-se igualmente uma breve panorâmica sobre algumas for-

mas típicas de abordar e de resolver os problemas do reconhecimento automático da fala. Tratando-se aqui de problemas que são sobretudo de engenharia, procurou-se fornecer elementos sobre o seu enquadramento quer no âmbito da investigação científica, quer nas consequências práticas em termos tecnológicos e no quotidiano da sociedade actual.

Capítulo 2

Aplicação dos HMMs no reconhecimento de fala

2.1 Introdução

A fala é a representação acústica de uma palavra ou de uma sequência de palavras, caracterizada por uma variação lenta da respectiva envolvente espectral. Os seres humanos percepcionam esta envolvente espectral e convertem-na para a sequência de palavras subjacente e para o significado associado. O objectivo final do processamento da fala e da língua é o de imitar este processo de tal modo que a máquina possa sustentar uma conversa natural com um humano. Em termos práticos espera-se, desta forma, instruir-se facilmente a máquina para a execução de tarefas complexas. Contudo, o processamento da fala e da língua têm um papel mais vasto no desempenho de tarefas menos complexas, tais como na transcrição, na identificação da língua ou no acesso a base de dados, todas estas ao alcance da tecnologia actual. O passo básico de todos estes sistemas é a execução de um mapeamento inverso da fala para a sequência de símbolos subjacente, os quais são, habitualmente, palavras (Young e Bloothooft, 1997).

A automatização do processamento do sinal requer actualmente a sua conversão para uma forma digital. No caso do sinal de fala, considera-se posteriormente a respectiva segmentação de modo a obterem-se segmentos de curta duração do sinal discreto, aproximadamente estacionários. Por forma a reduzir a quantidade de dados a serem processados, cada um destes segmentos pode ser representado por um conjunto de características, determinadas a partir do espectro de curta duração, designado por *observação*. Assim, antes do reconhecimento, o sinal de fala é convertido numa sequência de observações $O = \{o_1, o_2, \ldots, o_T\}$. Os aspectos referentes a este processo de conversão são abordados

na secção 2.2.

Considere-se W uma sequência de palavras ou de unidades subpalavra a identificar numa determinada sequência de observações O. Por sua vez, $Pr(W)$ representa a probabilidade a priori de observar W independentemente da sua realização acústica O. Esta probabilidade é determinada a partir do designado *modelo linguístico* ("language model"). Nas experiências de reconhecimento de palavras isoladas apresentadas nesta dissertação, consideraram-se equiprováveis todas as palavras do vocabulário testado, pelo que $Pr(W)$ é constante. No reconhecimento, para uma dada sequência de observações O, pretende-se determinar a sequência mais provável $\hat{W}$ que maximiza a probabilidade $Pr(W|O)$. Contudo, esta probabilidade não pode ser calculada directamente, sendo necessário utilizar a lei de Bayes[1]:

$$\hat{W} = \arg\max_{W} Pr(W|O) = \arg\max_{W}[\frac{Pr(W).Pr(O|W)}{Pr(O)}]. \tag{2.1}$$

A probabilidade da realização acústica independentemente de W, $Pr(O)$ é um elemento do problema que não é alterável pelo processo de reconhecimento. Assim, determina-se $\hat{W}$ maximizando o produto de $Pr(W)$ e $Pr(O|W)$. A probabilidade $Pr(O|W)$ é determinada a partir do designado *modelo acústico*, para o qual se utilizaram as representações de processos estocásticos de Markov descritos na secção 2.3.

Na secção 2.4 descreve-se a formulação dos modelos HMM semicontínuos e os compromissos nela envolvidos. Na secção 2.5 descrevem-se os conceitos e métodos necessários para o reconhecimento independente do vocabulário. Na secção 2.6 são referidos alguns modelos linguísticos utilizados para a determinação da probabilidade $Pr(W)$. A secção 2.7 apresenta alguns parâmetros utilizados na avaliação dos reconhecedores automáticos de fala. Por último, apresentam-se as conclusões do presente capítulo.

2.2 Extracção de características do sinal de fala

O processamento de sinal pressupõe a amostragem e quantificação do sinal para posterior tratamento digital. Tal deve-se essencialmente à versatilidade, ao poder de cálculo e ao baixo custo dos processadores digitais, bem como ao surgimento de novos algoritmos. O sinal de fala pode ser considerado estacionário em segmentos de curta duração ($\approx$ 10ms). Para reduzir a quantidade de dados a serem processados, é usual extrair-se um conjunto de características a partir de cada um destes segmentos. Na presente secção descrevem-se

[1] Uma formulação equivalente foi apresentada por (Jelinek et al., 1975).

os aspectos referentes à amostragem, à segmentação, à extracção de características e à forma de comparar características de segmentos diferentes.

2.2.1 Audibilidade e inteligibilidade do sinal de fala versus frequência

O sinal da fala é o sinal sonoro que resulta da manifestação oral da linguagem ou, noutra perspectiva, da comunicação verbal. Considera-se um som, qualquer estímulo susceptível de provocar uma sensação auditiva. Num sentido mais lato resulta de uma vibração mecânica que se propaga através da matéria (em geral, a atmosfera terrestre) até ao ouvido. Para poderem estimular o sistema auditivo humano, estas vibrações têm de obedecer a requisitos de intensidade, frequência e duração.

A intensidade e a frequência são habitualmente avaliadas na caracterização de um ouvido normal através do designado audiograma. A gama de frequências do sinal de fala mais relevante para a comunicação situa-se entre os 200 e os 5,6 kHz. Um audiograma típico revela que o ouvido humano é particularmente sensível a frequências entre 1 e 5 kHz. O afastamento desta faixa de frequências requer intensidades progressivamente maiores, por forma ao som ser detectado com igual intensidade. Os limites de audibilidade são diferentes para cada individuo, dependendo, nomeadamente, da idade. Ocorrem tipicamente aos 16 Hz e aos 18 kHz, para valores inferiores ou superiores entra-se, respectivamente, no domínio dos designados infra-sons e ultra-sons.

A contribuição das diferentes bandas de frequência para a inteligibilidade do sinal de fala é objecto de estudo desde os anos 20 (Steinberg, 1929; Steeneken e Houtgast, 1991). Actualmente realizam-se testes de inteligibilidade semelhantes com reconhecedores automáticos de fala que são comparados com testes perceptuais (Leeuwen et al., 1995) (secção 6.7).

Por uma questão de economia de meios de armazenamento e transmissão de dados, interessa limitar o sinal de fala a uma largura de banda mínima mas que contenha toda a informação relevante para o ouvido humano. No caso de um sistema de alta fidelidade, pretende-se abranger os limites de audibilidade que, conforme se referiu, raramente excedem[2] os 18 kHz. Nos sistemas de comunicação mais exigentes consideram-se ritmos de amostragem (f_s) de 20 e de 16 kHz. No sinal de fala, conforme se referiu, as frequências mais importantes para a comunicação não vão além dos 5,6 kHz. As ligações telefónicas

[2]Nos equipamentos de alta fidelidade digitais tais como o CD ("compact disk") e o DAT ("digital audio tape") o ritmo de amostragem é de 44,1 e 48 kHz, garantindo-se o espectro de frequência até 20 e 22 kHz, respectivamente.

convencionais permitem uma largura de banda de cerca de 3,4 kHz, pelo que, para esta qualidade, se utiliza um ritmo de amostragem de 8 kHz. Estes ritmos de amostragem mais baixos eliminam as frequências altas associadas às consoantes fricativas.

O corpus de fala recolhido no decorrer deste trabalho (secção 3.2) inclui sinais amostrados a 20 kHz, quantificados linearmente a 16 bits e armazenados em fitas Exabyte (8mm). Os mesmos sinais ficam igualmente registados por um gravador DAT convencional, com uma amostragem a 48 kHz e quantificação de 20 bits. Estes sinais foram posteriormente reamostrados digitalmente a 8 kHz e filtrados de modo a simular uma banda telefónica típica.

2.2.2 Uso e dimensionamento da janela de análise

O sinal de fala varia no tempo, em parte de forma aleatória, mas essencialmente sob o controlo do orador (O'Shaughnessy, 1987). Os parâmetros da fala utilizados no respectivo processamento automático, apresentam uma variação lenta no tempo que pode ser relacionada com aspectos fisiológicos dos movimentos do tracto vocal. Estes parâmetros são em geral do tipo estatístico, estimados num segmento de sinal contendo várias amostras consecutivas, designado por *trama* ("frame"). Habitualmente considera-se a trama como o resultado de uma segmentação uniforme ou fixa do sinal, ou seja, em segmentos consecutivos com a mesma duração. Existe um compromisso importante no dimensionamento desta duração. Por um lado, existe vantagem em se utilizar um número máximo de amostras possíveis, de modo a garantir a consistência do cálculo destes parâmetros. Por outro lado, este número não deverá corresponder a um tempo superior ao que permite garantir que o parâmetro não sofreu uma variação apreciável, ou seja, que possa ultrapassar um segmento de sinal de fala considerado quase-estacionário. Além disto, se a análise for feita em segmentos sobrepostos no tempo, consegue-se estimar eficazmente os referidos parâmetros nas zonas de transição entre tramas. Estes segmentos de sinal sobrepostos num número fixo de amostras, conduzem ao conceito de *janela*. Assim, uma nova janela contendo N amostras, pode ser obtida em intervalos de tempo inferiores a N/f_s segundos. Designa-se esta janela de *rectangular* se as amostras nela contidas forem consideradas com igual peso na determinação dos respectivos parâmetros. Os parâmetros assim obtidos apresentam flutuações no tempo que não correspondem a alterações reais do próprio sinal e que são causadas essencialmente pelas amostras incluídas nos extremos da janela. Para atenuar este efeito, consideram-se janelas que ponderam com maior peso as amostras centrais em relação às que se aproximam dos extremos. Como exemplo deste tipo de janela referem-se as de Bartlett, Blackman, Hamming, Hanning e Kaiser. A de

uso mais comum é a de Hamming (O'Shaughnessy, 1987; Lee, 1989; Rabiner e Juang, 1993):

$$h(n) = \begin{cases} 0,54 - 0,46\cos(\frac{2\pi n}{N-1}) & \text{para} \quad 0 \leq n \leq N-1 \\ 0 & \text{nos casos restantes.} \end{cases}$$

Nos sinais de fala em que a pronúncia é lenta, o tracto vocal e a forma de excitação podem permanecer sem alterações significativas por períodos de tempo que podem atingir os 200 ms. Contudo, e a avaliar pela duração média dos fonemas, (subsecção 2.5.1) que é aproximadamente de 80 ms, as características da fala variam de forma mais rápida na maioria das situações. Em vogais longas, a variação lenta das formantes pode ser analisada sem inconvenientes em janelas de cerca de 100 ms, enquanto que o relaxamento de uma oclusiva requer uma janela de 5 a 10 ms (rectangular) se se quiser evitar a mistura dos espectros dos sons adjacentes. Neste trabalho, considerou-se uma janela de Hamming de 20 ms ($N = 160$ amostras) que é deslocada em intervalos de 10 ms.

2.2.3 Detecção do início e do fim de palavras

Conforme se poderá verificar nos capítulos seguintes, o desempenho dos reconhecedores automáticos de fala é muito sensível à quantidade e qualidade da informação que lhes é fornecida, quer na fase de treino, quer na de teste (subsecção 2.3.5). Nomeadamente, o desempenho de reconhecedores de palavras isoladas decresce de forma significativa com o número de tramas omitidas ou em excesso, em relação ao segmento de sinal que corresponde exclusivamente à produção acústica da palavra. Estas tramas podem ocorrer devido a uma segmentação deficiente, quer no início, quer no fim da palavra. Isto mesmo foi verificado quando estas tramas são referentes aos extremos de uma segmentação manual, efectuada com base na audição e na visualização dos contornos de energia do sinal (Rabiner e Juang, 1993). Assim, no pré-processamento do sinal de fala para o reconhecimento de palavras isoladas, utilizam-se os designados *detectores de início e fim de palavras*. Estes detectores devem determinar, da forma mais precisa possível, o segmento de sinal entre os instantes em que começa e termina a articulação da palavra a reconhecer.

No reconhecimento de fala contínua utilizam-se por vezes procedimentos muito semelhantes, detectando-se pausas significativas do orador que podem assumir interpretações específicas para cada aplicação. Por exemplo: tomada de vez ("turn-taking") de outro interlocutor ou sintetizador; aguarda execução de uma tarefa; fim de conversação, etc.

De um ponto de vista semelhante, este tipo de detecção destina-se, também, a economizar espaço de armazenamento de dados e tempo de processamento. Assim, é frequente-

mente utilizado na recolha de corpora de sinais de fala que utilizam estes procedimentos "on-line" por forma a não desperdiçar meios de registo para o sinal de fala (capítulo 3).

Os designados *detectores de fala* partilham muitas semelhanças com os detectores de início e do fim de palavras, sendo utilizados em muitas aplicações de telecomunicações. Por exemplo, nos sistemas de transmissão analógica multicanal, é utilizada uma técnica que permite que um determinado número de canais n_C seja utilizado por um número superior de conversações (cerca de $2{,}5.n_C$ conversações). Esta técnica, conhecida por TASI, ("time-assignment speech interpolation") utiliza as pausas da conversação num dado canal para o atribuir a outra conversação onde esteja a ser detectado sinal de fala (Rabiner e Juang, 1993).

A utilização de detectores de fala é também comum em métodos de redução de ruído do tipo subtracção espectral. Neste contexto, os detectores são usados para obter estimativas em segmentos do sinal exclusivamente com ruído. Um exemplo de um método deste tipo, que foi implementado para processamento em tempo real, encontra-se descrito em (Teixeira e Trancoso, 1991a; Teixeira e Trancoso, 1991b). Este detector de fala baseia-se num único limiar de energia, calculado a partir de duas estimativas da energia do sinal: uma para o sinal com fala e outra para o sinal sem fala. Ambas as estimativas são actualizadas ao longo do tempo, com factores de esquecimento do tipo exponencial.

No contexto do reconhecimento de fala, a detecção do início e do fim de palavras pode ser realizada de três formas distintas. Se a detecção ocorre numa fase anterior ao reconhecimento, determinando uma segmentação do sinal de forma independente da decisão do reconhecedor, é designada por *detecção explícita*. Se pelo contrário, a detecção é feita de forma implícita nos próprios mecanismos do reconhecedor (subsecção 2.3.4) então é designada por *detecção implícita*. Uma outra forma de efectuar a detecção baseia-se em métodos de detecção explícita capazes de gerarem diversas segmentações possíveis. Estas segmentações são fornecidas ao reconhecedor que escolhe aquela que apresenta o maior valor para a probabilidade $Pr(W|O)$ (equação 2.1). Este tipo de detecção é designada por *híbrida*. Por vezes é possível dispor de uma indicação da qualidade de cada uma destas segmentações, com a qual se determina a ordem da sua apresentação ao reconhecedor. O reconhecedor descodifica as sucessivas segmentações até surgir uma que apresente um valor de $Pr(W|O)$ superior a um determinado limiar.

Neste trabalho, adoptou-se um método de detecção explícita capaz de gerar uma lista de ordenadas de segmentações. Este método foi inicialmente desenvolvido para um detector do tipo híbrido mas provou ser também muito eficaz se se considerasse exclusivamente a primeira segmentação da referida lista. De facto, 84% dos testes realizados com o de-

tector híbrido determinaram apenas uma única segmentação. A taxa de reconhecimento obtida com a utilização desta segmentação foi de 74,9%, enquanto que um detector do tipo explícito convencional, utilizando exclusivamente a energia do sinal, obteve apenas o valor de 68,8% (Lamel et al., 1981).

O detector do início e do fim de palavras utilizado no presente trabalho, pode ser descrito em três etapas distintas:

1. Equalização adaptativa do nível de energia;
2. Localização de segmentos do sinal com uma energia elevada, aqui designados de impulsos;
3. Determinação e ordenamento das segmentações. De facto, neste caso, apenas se pretende determinar qual a segmentação mais adequada.

Os referidos impulsos são agrupados de forma heurística criando diversos pares de início e fim de palavra possíveis (segmentações). A ordenação destas segmentações numa lista é determinada pela aplicação sucessiva dos seguintes pressupostos:

1. O segmento correspondente à palavra isolada que se pretende determinar inclui um ou mais impulsos;
2. A trama que apresentar um valor máximo da energia, será sempre incluída no segmento referido em 1.;
3. Quanto maior for a duração de tempo entre dois impulsos, menos provável será a possibilidade de ambos pertencerem a uma palavra;
4. Os impulsos que ocorram com uma diferença de tempo superior a 150 milissegundos em relação ao impulso com maior energia, não deverão ser considerados como fazendo parte da palavra.

2.2.4 Análise espectral

No processamento de sinais distinguem-se dois tipos fundamentais de técnicas de análise do sinal: a análise temporal e a análise espectral. Os parâmetros mais úteis para o reconhecimento de fala são referentes ao domínio da frequência. O tracto vocal produz sinais que são analisados de forma mais consistente e fácil no domínio espectral do que no domínio temporal. De facto, conforme se descreve nos parágrafos seguintes, o ouvido

humano dá maior relevância a variações de características referentes à distribuição de amplitudes na frequência, do que a aspectos relacionados com a respectiva fase e de temporização.

O ouvido humano

Uma das formas mais conhecidas de analisar o espectro do sinal de fala baseia-se na utilização de um *banco de filtros*: um conjunto de filtros passa-banda que cobrem intervalos de frequência consecutivos do espectro do sinal de fala. O critério mais comum na definição do banco de filtros é a de considerar filtros com a mesma largura de banda e uniformemente distribuídos ao longo do espectro disponível. Contudo, existem estudos psico-acústicos que revelam que o ouvido humano atribuí importância diferente a cada uma destas bandas de frequência (Rabiner e Juang, 1993). É de facto possível identificar aí estruturas anatómicas que actuam como filtros na frequência do sinal sonoro antes de este atingir o cérebro. O ouvido médio actua como um filtro passa-baixo com uma atenuação de cerca de 15 dB/oitava acima de 1 kHz. Contudo, é no ouvido interno que ocorrem as alterações mais interessantes em termos espectrais.

O ouvido interno é classificado em duas partes designadas por *labirinto ósseo* e *labirinto membranoso*, este último contido no primeiro. No labirinto membranoso existe a *cóclea* que é um tubo de cerca de 35 mm preenchido por um fluído, o *licor de Cotunni*. O tubo enrola-se em cerca de 2,5 voltas numa espiral com forma de caracol. Duas membranas dividem o interior da cóclea ao longo da espiral em três câmaras. A membrana *de Reissner* separa a câmara de volume superior, a *scala vestibuli*, da denominada câmara média (*scala media*). Por sua vez membrana *basilar* separa a câmara média da *scala tympani*.

Na base da cóclea encontram-se a *janela oval* e a *janela redonda*. As vibrações provenientes do ouvido médio são transmitidas à *scala vestibuli* através da membrana da janela oval. O líquido na *scala vestibuli* comunica com a *scala tympani* através de uma pequena abertura *helicotrema* no ápice da cóclea. As pressões geradas na *scala tympani* são aliviadas através da janela redonda.

A propagação da energia sonora como ondas hidrodinâmicas no fluído através das câmaras funciona essencialmente como um filtro passa-baixo distribuído. Conforme as ondas são propagadas ao longo da cóclea, ocorre uma acção de filtragem em que as frequências altas são fortemente atenuadas. Designa-se de *lugar* a localização dentro da cóclea medida pela distância a partir da base na direcção do ápice. A frequência de corte baixa gradualmente em função do lugar. Uma das teorias clássicas da audiologia classifica a cóclea como um analisador espectral e considera os níveis de saída em função do lugar

como um bom processo de caracterização do som.

Dentro da câmara média, sobre a membrana basilar e por baixo da designada *membrana tectorial*, existe o *órgão de Corti*. Trata-se de uma estrutura com cerca de 30.000 células ciliadas e organizadas em diversas linhas ao longo da cóclea. Dentro do órgão de Corti existe uma linha com cerca de 3.500 destas células, designadas por *células internas*. As terminações do nervo auditivo deverão ser estimuladas por cerca de 40 a 140 cílios das células internas, que se agitam entre a membrana basilar e a tectorial. Estes movimentos são convertidos pelas células internas numa corrente de iões que é transmitida ao cérebro através de cerca de 28.000 fibras nervosas (O'Shaughnessy, 1987).

A onda de pressão na cóclea faz oscilar a membrana basilar que se desloca num plano paralelo à membrana tectorial. Quando os cílios se movem numa determinada direcção, as células internas estimulam os nervos auditivos primários que geram os disparos neuronais. Se após este movimento os cílios ficarem imóveis numa qualquer posição, os disparos acabam por desaparecer. Quando se movimentam na direcção oposta da anterior também não se produzem quaisquer disparos. Assim, as células internas apresentam um comportamento que pode ser descrito em termos eléctricos como: anulamento da componente contínua com rectificação de meia onda. Por este motivo, os modelos auditivos conhecidos consideram geralmente o sinal na membrana basilar, rectificado de meia onda (Lyon e Mead, 1988; Perdigão, 1997).

Além da linha de células internas existem três linhas das designadas *células externas*. Estas células não transmitem informação acerca do som para o cérebro actuando principalmente como um músculo sobre a membrana basilar. Se não forem inibidas pelos nervos eferentes, produzem retroacção positiva na membrana e em determinadas condições conseguem colocar energia suficiente para provocar oscilações audíveis. Este fenómeno é designado por *tinnitus*. Em condições normais o efeito das células externas é o de diminuir o factor de amortecimento da membrana basilar quando o som seria de outra forma demasiado baixo para ser ouvido e vice-versa, isto é, aumentar o amortecimento quando os sons forem intensos. Esta acção pode ser descrita como a de um amplificador activo com controlo automático de ganho. Este é um sistema de controlo de ganho inteligente que actua antes do sinal ser convertido para potenciais eléctricos, tal como acontece com o sistema visual de controlo de ganho na retina. Os impulsos nervosos são muito pouco adequados à transmissão dos sinais sonoros: são ruidosos, erram e têm uma gama dinâmica limitada para os ritmos dos disparos. Esta realimentação, actuando antes da geração dos impulsos nervosos, diminui o ruído associado com esta forma de quantificação do sinal (Lyon e Mead, 1988).

Estudos psico-acústicos determinaram um modelo do tipo banco de filtros para a membrana basilar com um conjunto de 24 filtros passa-banda com larguras de banda crescentes com a frequência. Cada uma das bandas, designada por *banda crítica*, corresponde aproximadamente a 1.200 fibras nervosas primárias ou a um espaçamento de cerca de 1,5 mm sobre a membrana basilar (Rabiner e Juang, 1993). Pretende-se desta forma equilibrar a importância relativa, em termos perceptuais, de cada uma destas bandas. Uma primeira aproximação a esta abordagem determina a adopção de uma escala de frequências logarítmica. A designada escala de *Bark* constitui um critério para o desenho de um banco de filtros que identificam as referidas bandas críticas. Por sua vez a conhecida escala de *mel* constituí uma outra variante baseada nas bandas críticas (Davis e Mermelstein, 1980). Esta escala é aproximadamente linear até frequências próximas de 1 kHz e logarítmica para frequências superiores. As diferenças entre as escalas de *Bark* e de *mel* são geralmente pequenas e insignificantes do ponto de vista da sua aplicação para o reconhecimento de fala (Rabiner e Juang, 1993).

2.2.5 Análise autorregressiva

A análise autorregressiva tem sido utilizada em diversas áreas do processamento digital de sinal. Os exemplos tradicionais incluem a análise dos sinais sísmicos (Makoul, 1975), neurofisiológicos tais como o electroencefalograma (Zetterberg, 1969; Gersh, 1970; Teixeira, 1989) e os próprios sinais de fala (Markel e Gray, 1976; Serralheiro, 1990). Este tipo de análise é também por vezes designada por *análise de predição linear* ou de LPC ("linear predictive coding").

Este tipo de análise baseia-se na obtenção dos parâmetros de um sistema causal, discreto, linear, invariante no tempo e sem zeros fora da origem. O modelo autorregressivo de ordem p permite predizer uma amostra $\hat{x}(n)$ num dado instante n de um sinal discreto $x(n)$ com base numa combinação linear das p amostras anteriores:

$$\hat{x}(n) = -\sum_{i=1}^{p} a_i x(n-i).$$

A esta estimativa corresponde um erro de predição

$$e(n) = x(n) - \hat{x}(n) = \sum_{i=0}^{p} a_i x(n-i), \qquad a_0 = 1. \tag{2.2}$$

Cálculo dos coeficientes autorregressivos

O cálculo dos coeficientes autorregressivos a_i pode ser conseguido através de duas estratégias diferentes baseadas na minimização do erro quadrático médio $\mathcal{E}[e(n)]^2$. No designado *método da autocorrelação* assume-se que $\mathcal{E}[e(n)]^2$ é minimizado num período de tempo infinito ($-\infty < n < \infty$) enquanto que no *método da covariância* esta minimização ocorre num intervalo finito ($0 < n < N-1$). Na prática este tipo de análise só é aplicável a pequenos troços do sinal de fala que se possam considerar com características aproximadamente estacionárias ou quase-estacionárias. Assim, no método da autocorrelação, é aplicada uma janela (no caso presente uma janela de Hamming) ao sinal, antes do cálculo do erro, enquanto que no método da covariância essa janela é aplicada ao erro $e(n)$. O método da autocorrelação introduz distorção nos procedimentos de estimação espectral uma vez que a utilização da janela corresponde a uma convolução da resposta em frequência da janela com o espectro de curta duração original do sinal. O método da covariância evita esta distorção mas o respectivo cálculo é, em geral, mais complexo. De facto, no método da autocorrelação é necessária a resolução de um sistema de equações em que intervém a matriz de autocorrelação que é uma matriz de Toeplitz (simétrica com todos os elementos diagonais iguais). Este tipo de equações podem ser eficientemente resolvidas através de procedimentos bem conhecidos tal como o procedimento recursivo de Levinson-Durbin (Makoul, 1975). Por este motivo, utilizou-se este procedimento na presente dissertação para o cálculos dos coeficientes autorregressivos. No método da covariância, o sistema de equações correspondente substitui a matriz de autocorrelação pela matriz de covariância que é simétrica mas não de Toeplitz. Neste caso, os métodos de resolução disponíveis necessitam de um número superior de cálculos. Existem outros tipo de métodos que permitem estimar os parâmetros autorregressivos. Em particular destacam-se os métodos capazes de actualizar as estimativas autorregressivas para cada amostra disponível e que podem ser classificados em dois grupos distintos: os preditores transversais e as formas designadas por "lattice".

O valor mínimo de $\mathcal{E}[e(n)]^2$ determinado pelo método da autocorrelação é dado por

$$E_p = R(0) + \sum_{k=1}^{p} a_k R(k), \tag{2.3}$$

em que $R(k) = \sum_{-\infty}^{+\infty} x(n)x(n+k)$ são os coeficientes de autocorrelação.

Descrição do espectro a partir do modelo autorregressivo

Aplicando a transformada Z à expressão 2.2 obtém-se

$$X(z)/E(z) = 1/\sum_{i=0}^{p} a_i z^{-i},$$

em que $X(z)$ e $E(z)$ representam as transformadas Z de $x(n)$ e de $e(n)$, respectivamente. O erro de predição obtido com a expressão 2.2 quando os coeficientes autorregressivos foram determinados pelo método da autocorrelação, apresenta uma variância mínima $\sigma^2 = E_p$ (equação 2.3). Ou seja, $e(n) = \sigma w(n)$, em que $w(n)$ tem média nula e variância unitária. Se $w(n)$ for também uma sequência de amostras não correlacionadas (ruído branco) a sua transformada Z é dada por $W(z) = 1$, obtendo-se

$$X(z) = \sigma/\sum_{i=0}^{p} a_i z^{-i}. \tag{2.4}$$

A densidade espectral de potência de $w(n)$ é, por sua vez, de valor igual ao período de amostragem[3] $T_s = 1/f_s$. Assim, a densidade espectral de potência de $x(n)$ pode ser representada por (Kay e Marple, 1981)

$$\frac{\sigma^2 T_s}{|\sum_{i=0}^{p} a_i e^{j\omega i T_s}|^2}.$$

Ordem do modelo autorregressivo

Por motivos práticos, é geralmente desejável usar o número mínimo de parâmetros necessários para descrever com precisão as características significativas do sinal. Na modelação do espectro da fala, estas características são as ressonâncias do tracto vocal (formantes) e, embora com menos importância, as regiões entre estas ressonâncias (Markel e Gray, 1976). Um critério heurístico para a determinação da ordem p do modelo autorregressivo considera que esta deverá ser superior ao dobro do número de ressonâncias do espectro. Por vezes consideram-se cerca de dois a quatro pólos adicionais que deverão aproximar a possível existência de alguns zeros (O'Shaughnessy, 1987). Na teoria, cada zero só pode ser substituído por um número infinito de pólos; consegue-se, contudo, uma boa aproximação com um número reduzido de pólos, tendo em consideração que os vales (zonas de baixa energia do espectro) não necessitam de ser descritos com a mesma precisão exigida para os picos (zonas do espectro com maior energia). A ordem p deve ainda

[3] O integral de $w(n)$ na banda $[-1/2T_s, 1/2T_s]$ tem de ser igual à potência real do respectivo sinal analógico.

aumentar com a frequência de amostragem, uma vez que, por exemplo no sinal de fala, se devem incluir um maior número de formantes.

Existem diversos critérios que permitem dar uma indicação aproximada da ordem do modelo autorregressivo dos quais se destacam: o critério da variação relativa do erro ("relative variation error") (Makoul, 1975); o critério do erro de predição final ("final prediction error"); o critério de MAICE ("minimum A information criterion estimate") (Akaike, 1974) e o critério da função de transferência autorregressiva. Estes critérios e algumas das suas variações têm sido objecto de alguns estudos dedicados a determinados tipos de sinais incluindo os de fala (Markel e Gray, 1980). No caso destes últimos tem sido utilizado o erro de predição normalizado

$$V_p = E_p/R(0),$$

em que $R(0) = E[x^2(n)]$ corresponde à energia do sinal. V_p decresce monotonamente com a ordem p.

De acordo com os critérios e estudos referidos existem valores de p conhecidos para as frequência de amostragem mais comuns, que representam compromissos geralmente aceitáveis para a maioria das aplicações em processamento de fala. Desta forma estabeleceu-se neste trabalho $p = 8$, o que pode ser considerado um valor que preserva aproximadamente as primeiras quatro formantes (f_s=8 kHz).

Validade do modelo autorregressivo para o sinal de fala

Da descrição anterior do modelo autorregressivo decorrem duas limitações básicas em relação ao sinal da fala:

- A inexistência zeros (fora da origem) no modelo autorregressivo. Trata-se de um modelo simplificado, uma vez que surgem zeros associados ao espectro da fala devido à forma do sinal resultante da glote em combinação com a radiação a partir dos lábios. Existem ainda zeros associados à resposta do tracto vocal nas nasais e nos sons não vozeados (O'Shaughnessy, 1987).

- A excitação com ruído branco, que corresponde a um espectro plano. Este pressuposto é razoável para os sons não vozeadas mas não o é para os sons vozeados em que a excitação deverá ser periódica e consequentemente com um espectro de riscas (Almeida, 1982).

Parametrizações baseadas no modelo autorregressivo

Existem diversas formas equivalentes para os coeficientes autorregressivos a_i, algumas delas com significado físico mais evidente, tais como: os coeficientes de reflexão; a resposta impulsiva $h(n)$ do filtro LPC[4]; a autocorrelação dos parâmetros a_i ou dos valores de $h(n)$; os correspondentes coeficientes espectrais das transformadas de Fourier discretas das respectivas autocorrelações; o cepstrum referente aos valores de a_i ou de $h(n)$. Outras formas de representação foram concebidas no sentido de permitirem uma quantificação mais eficiente, tais como as designadas "log-area ratios"; as funções do inverso de seno e as "line spectrum pairs" (LSP) (O'Shaughnessy, 1987).

Neste trabalho adoptou-se o modelo autorregressivo para o sinal de fala (SIRtrain, 1991; Jacobsen, 1992; Young et al., 1996). Os parâmetros autorregressivos foram calculados utilizando o método da autocorrelação (Makoul, 1975). Contudo, estes parâmetros foram previamente convertidos em coeficientes cepstrais, antes de serem utilizados pelos métodos de reconhecimento de fala. Na secção 2.2.6, apresentam-se alguns detalhes deste tipo de representação.

Pré-ênfase

Uma das primeiras operações habituais no sinal de fala, logo após a sua digitalização e antes da respectiva análise autorregressiva, consiste numa filtragem digital de primeira ordem designada por pré-ênfase:

$$\tilde{x}(n) = x(n) - \alpha x(n-1), \tag{2.5}$$

com $0{,}9 < \alpha < 1{,}0$. O filtro de pré-ênfase é um filtro passa alto que torna o espectro do sinal de fala mais plano e a análise LPC menos susceptível a erros devidos a representações numéricas de precisão finita. No presente trabalho, utilizou-se $\alpha = 0{,}95$. A metade da frequência de amostragem, este filtro introduz um ganho de cerca de 32 dB (Rabiner e Juang, 1993).

Os segmentos não vozeados do sinal de fala apresentam um espectro relativamente plano. A pré-ênfase deveria ser aplicada exclusivamente às zonas vozeadas do sinal da fala que apresentam uma quebra de amplitude de cerca de -6 dB/oitava. No entanto, a degradação introduzida por esta filtragem na análise dos segmentos não vozeados é

[4]Da equação 2.4 obtém-se $x(n) = -\sum_{i=1}^{p} a_i x(n-i) + \sigma\delta(n)$ em que $\delta(n)$ é a função generalizada impulso unitário de Dirac. Bastam apenas os primeiros $p+1$ valores para especificar o filtro inequivocamente.

considerada desprezável. Assim, a pré-ênfase é aplicada a todo o sinal de fala e a referida degradação ponderada na determinação do factor α (O'Shaughnessy, 1987).

2.2.6 Análise cepstral

O espectro de curta duração do sinal de fala inclui uma envolvente espectral de variação lenta correspondente ao filtro que representa o tracto vocal e, no caso da fala vozeada, de uma estrutura espectral mais fina, de variação rápida, correspondente à excitação periódica e às respectivas harmónicas. Este espectro pode ser obtido pelo produto da resposta em frequência deste filtro com o espectro da excitação. Utilizando-se logaritmos, pode-se decompor este produto numa soma dos espectros representativos do tracto vocal e da respectiva excitação.

O logaritmo da transformada de Fourier de um sinal $x(n)$ é conhecido pela designação de *cepstrum* (Oppenheim, 1974):

$$C(z) = \log X(z).$$

Os parâmetros da transformada Z inversa de C(z) são designados por *quefrency* e são considerados parâmetros (pseudo) temporais (Huang et al., 1990). Na prática, não é necessário o cálculo do cepstrum complexo e utiliza-se a designação para o cepstrum real (O'Shaughnessy, 1987):

$$c(n) = \frac{1}{2\pi} \int_0^{2\pi} \log |X(e^{j\omega})| e^{j\omega} d\omega.$$

Em termos da implementação digital, a transformada de Fourier contínua tem de ser substituída pela transformada de Fourier discreta, em N frequências discretas equi-espaçadas entre 0 e 2π:

$$c(n) = \frac{1}{N} \sum_{k=0}^{N-1} \log |X(k)| e^{j2\pi kn/N} \quad \text{para} \quad n = 0, 1, \ldots, N-1.$$

Os coeficientes cepstrais de índice superior estão relacionados com a excitação, devido às correspondentes componentes de frequências mais altas. Os coeficientes de índice inferior, por sua vez, dependem essencialmente do tracto vocal e da correspondente envolvente espectral de baixa frequência. Esta análise, designada por *análise cepstral*, permite separar a convolução da excitação com o filtro representativo do tracto vocal. Designa-se de *análise homomórfica* aquela que utiliza uma transformação que permite efectuar a desconvolução de determinadas propriedades numa soma. Esta análise decorre de uma generalização do princípio da sobreposição para sistemas não lineares (Oppenheim, 1974).

A vantagem mais relevante que decorre da utilização dos coeficientes cepstrais advém do facto de estes apresentarem correlações entre si extremamente baixas. Tal facto permite simplificações apreciáveis para o processamento baseado nestes tipo de parâmetros (subsecção 2.2.7).

Nos sinais de fala, os coeficientes cepstrais são geralmente obtidos a partir de um dos dois tipos de métodos de análise espectral anteriormente referenciados: bancos de filtros ou modelo autorregressivo.

Devido à eficiência e à simplicidade do cálculo dos coeficientes cepstrais a partir dos coeficientes LPC optou-se, neste trabalho, por este método. A expressão recursiva utilizada é dada por (Makoul, 1975):

$$c(n) = a(n) - \sum_{m=1}^{n-1} \frac{m}{n} c(m) a(n-m), \qquad \text{para} \quad n \geq 1, \tag{2.6}$$

com $a(n) = 0$ para $n \geq p$.

Levantamento cepstral

A variabilidade da *quefrency* devido às limitações inerentes à análise autorregressiva (inexistência de zeros, posição da janela de análise, interacção da excitação com o sistema, etc.) é maior nos coeficientes de ordem mais elevada do que nos de ordem inferior. Os coeficientes de ordem inferior são, portanto, menos influenciáveis pelas particularidades do método de análise, dependendo predominantemente do canal de transmissão, das características do orador e restantes factores inerentes ao próprio sinal da fala. Por outro lado, se se considerar por exemplo fala via telefónica, a resposta em frequência das diversas linhas telefónicas é particularmente notada nos primeiros coeficientes cepstrais. O mesmo acontece com as variações da forma da glote e das cordas vocais, bem como o designado declive espectral ("spectral tilt") (Rabiner e Juang, 1993). Estas características são essenciais para tarefas de reconhecimento do orador, mas são fontes indesejáveis de variabilidade no reconhecimento de fala independente do orador.

Pelos motivos descritos é de esperar resultados de reconhecimento melhorados atribuindo um peso adequado a cada um dos coeficientes cepstrais. Esta ponderação dos coeficientes cepstrais é muito utilizada no reconhecimento de fala, sendo conhecida pela designação de levantamento ou pesagem cepstral ("cepstral liftering") (Juang et al., 1987):

$$c_n^{'} = [1 + \frac{L}{2} \sin(\frac{\pi n}{L})] c_n. \tag{2.7}$$

De acordo com a experiência de outros investigadores (Jacobsen, 1992; Rabiner e Juang, 1993; Young et al., 1996), utilizou-se no presente trabalho o levantamento cepstral descrito pela equação 2.7 para o valor de $L = 12$.

Variação dos coeficientes no tempo

As variações do espectro no tempo são importantes na percepção da fala. A inclinação crescente ou descendente das formantes são dados importantes para um perito na leitura dos espectrogramas. De facto, a localização das formantes varia de orador para orador. No entanto, os respectivos declives são relativamente independentes do orador (Lee, 1989). Com base nestes factos, é de esperar que o desempenho de um sistema de reconhecimento de fala possa ser substancialmente melhorado com a utilização das diferenças entre os parâmetros básicos da fala no tempo (Young et al., 1996). Conforme o descrito na secção 2.3, os reconhecedores baseados em modelos de Markov não observáveis consideram em geral cada segmento de sinal independentemente dos anteriores. Pretende-se atenuar esta limitação com a integração do conhecimento das diferenças entre parâmetros sucessivos que acumulam assim vantagens adicionais para este tipo de reconhecedores. A inclusão destas diferenças como parâmetros para o reconhecimento de fala só se tem verificado nos reconhecedores mais recentes, nomeadamente a partir dos coeficientes cepstrais também designados por delta-cepstrum (Lee, 1989; Lee et al., 1990b). De igual forma, se podem calcular as diferenças entre os parâmetros assim obtidos. Estes últimos são por vezes designados por parâmetros de aceleração (por analogia à utilização da segunda derivada na cinemática).

Neste trabalho, considerou-se uma solução adoptada para o cálculo das variações dos parâmetros $c(i)$ que se baseia na seguinte fórmula de regressão (Young et al., 1996):

$$\Delta_c(i) = \frac{\sum_{n=1}^{\Theta} n[c(i+n) - c(i-n)]}{2\sum_{n=1}^{\Theta} n^2},$$

com $\Theta = 2$.

Caracterização de um segmento de fala

O vector de parâmetros escolhido para representar os segmentos do sinal de fala de curta duração, (subsecção 2.2.2) inclui os coeficientes do cepstrum e do delta-cepstrum. O reconhecedor disponível no início deste trabalho não previa o uso dos coeficientes $c(0)$ e Δ_{c_0} (SIRtrain, 1991; Jacobsen, 1992). Alguns investigadores verificaram que estes parâmetros,

referentes à energia do sinal, não contribuem de forma relevante para o melhor desempenho de reconhecedores dependentes do orador, embora o mesmo não se passe com os reconhecedores independentes do orador (Lee, 1989). Posteriormente, utilizou-se outro reconhecedor (HTK) que não apresentava esta restrição, tendo sido efectuados testes de reconhecimento independente do orador para determinar eventuais vantagens no uso destes parâmetros. As diferenças de desempenho detectadas não foram consideradas significativas. Optou-se assim por considerar apenas os restantes oito coeficientes cepstrais mais os respectivos parâmetros do delta-cepstrum, num total de 16 parâmetros.

2.2.7 Medidas de similaridade entre segmentos de fala

Nas metodologias da área do reconhecimento de padrões é essencial dispor de uma forma de comparar os elementos de determinado conjunto de objectos com características semelhantes. As características de cada objecto devem poder ser representadas num ponto de um espaço de vectores $\mathcal{V}$, por forma a permitir uma abordagem analítica adequada. Deste modo é possível definir uma função real $d(x, y)$ no produto cartesiano $\mathcal{V} \times \mathcal{V}$ que forneça uma medida de similaridade ou de dissemelhança entre dois objectos $x, y \in \mathcal{V}$. Uma função de dissemelhança (ou de distorção) $d(x, y)$ é designada por medida de distância se: for nula para $x = y$; for positiva para $x \neq y$; for simétrica ($d(x, y) = d(y, x)$); e verificar $d(x, y) \leq d(x, z) + d(y, z)$, $\forall x, y, z \in \mathcal{V}$.

Nas secções anteriores foram descritos alguns conjuntos de parâmetros, que permitem caracterizar um dado segmento de fala. De seguida são descritas algumas medidas de dissemelhança e de similaridade que permitem a comparação entre diversos segmentos de fala.

Medida de distância espectral

Na subsecção 2.2.4 referiu-se a importância da utilização da análise espectral no processamento da fala. Procura-se agora uma medida de distância que permita a comparação entre diferentes segmentos de fala a partir das respectivas características espectrais.

Um sinal discreto $x(n)$, com transformada Z dada por $X(z)$, apresenta um espectro de potência $|X(z)|^2$. Considere-se a seguinte medida de distorção entre dois sinais $x(n)$ e $y(n)$

$$d_{x,y}(z) = \log |X(z)|^2 - \log |Y(z)|^2, \tag{2.8}$$

e o respectivo valor quadrático médio:

$$d_2^2(x,y) = \int_{-\pi/T}^{\pi/T} d_{x,y}(e^{j\omega T})^2 \frac{d\omega}{2\pi}.$$

Prova-se (Huang e Jack, 1989) que $d_2^2(x,y)$ corresponde a uma medida de distância L_2 entre os respectivos coeficientes do cepstrum:

$$d_2^2(x,y) = \sum_{-\infty}^{\infty} [c_x(n) - c_y(n)]^2.$$

Considere-se um espectro de potência só com pólos e sem zeros fora da origem, tal como o de LPC. Prova-se que o cepstrum correspondente (equação 2.6) é assimptoticamente limitado e que constitui uma sequência decrescente, pelo que se pode aproximar d_2 por uma sequência finita (Rabiner e Juang, 1993). De facto, pode considerar-se a distância d_c expressa na equação 2.9, com apenas L parcelas, como uma boa representação para a distância espectral entre dois espectros só com pólos. Para tal, L não deverá ser inferior à ordem estimada para o modelo autorregressivo (subsecção 2.2.5):

$$d_c^2(x,y) = \sum_{n=1}^{L} [c_x(n) - c_y(n)]^2. \tag{2.9}$$

Medida de similaridade baseada na função gaussiana

No processamento da fala pode ser vantajoso substituir uma medida de similaridade pelo uso de uma função densidade de probabilidade contínua. Tal deve-se essencialmente ao carácter não determinístico de que se revestem os sinais de fala. Os parâmetros da densidade de probabilidade de um padrão podem ser estimados com o uso de um número significativo de realizações do sinal, o que resulta numa representação mais robusta quando comparada com as medidas de distorção tradicionais, nomeadamente, as referidas nos parágrafos anteriores (Huang e Jack, 1989).

A conhecida família de funções designada por gaussiana ou normal possui várias propriedades matemáticas importantes (por exemplo o teorema de DeMoivre-Laplace) sendo uma representação adequada à descrição de inúmeros fenómenos experimentais (Papoulis, 1984). A forma multivariável para um vector o_t com dimensão n, pode ser representada por

$$\mathcal{N}(o;\mu,\Sigma) = \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} \exp\{-\frac{1}{2}(o_t - \mu)'\Sigma^{-1}(o_t - \mu)\}, \tag{2.10}$$

em que o vector μ e a matriz Σ podem ser considerados parâmetros fixos da função. Estas funções são particularmente importantes quando se considera o_t um vector de variáveis

aleatórias. O designado teorema do limite central determina que, se as variáveis em o_t forem o resultado da soma de um número elevado de variáveis aleatórias independentes entre si, então a função densidade de probabilidade de o_t tenderá para uma função semelhante à descrita pela equação 2.10. Neste caso, os parâmetros fixos μ e Σ assumem, respectivamente, o significado de vector das médias e o de matriz de covariância de o_t. Uma função densidade de probabilidade gaussiana pode representar uma probabilidade condicional de se observar um vector o_t dado um determinado conjunto de vectores da mesma natureza, com média μ e com matriz de covariância Σ. Esta probabilidade pode ser interpretada como uma medida de similaridade entre os vectores o_t e μ quando ambos se referem ao mesmo espaço de vectores aleatórios com matriz de covariância Σ. De facto, calculando o logaritmo natural da expressão 2.10 é possível identificar um único termo que depende dos vectores o_t e μ com a forma

$$d_M(o_t, \mu) = (o_t - \mu)' \Sigma^{-1} (o_t - \mu),$$

que representa a conhecida distância de Mahalanobis (Duda e Hart, 1973). Deste modo, verifica-se que o uso das densidades gaussianas tem uma relação próxima com a medida de distância espectral que utiliza a distância euclidiana para os coeficientes cepstrais (equação 2.9). O facto destes coeficientes apresentarem em geral uma baixa correlação entre si permite aproximar a matriz de covariância Σ pela respectiva diagonal e acentuar desta forma as semelhanças com a norma L_2.

Além disso, a distância de Mahalanobis pode ser aplicada com sucesso a diversos conjuntos de parâmetros sem necessidade de outras justificações baseadas na natureza de cada um deles. Conforme já se referiu, conhecem-se outros parâmetros, para além do cepstrum, que permitem representar adequadamente um segmento de sinal de fala. Uma medida de similaridade deste tipo garante uma integração no reconhecedor, de modo simples, de diversos destes parâmetros. Além disso permite comparar, em condições semelhantes, o desempenho de reconhecedores que utilizam diferentes conjuntos de parâmetros.

Aproximação de funções densidade de probabilidade utilizando uma composição de funções gaussianas

De acordo com os elementos descritos nos parágrafos anteriores, os parâmetros estimados sobre segmentos de sinal de fala são habitualmente descritos, em primeira aproximação, por uma distribuição gaussiana. Contudo, para uma descrição mais adequada, nomeadamente para sinais originários de diversos oradores, esta aproximação revela-se demasiado grosseira. Ainda assim, é vantajoso considerar funções gaussianas, quando

utilizadas em conjunto numa composição ponderada, tal como a seguinte:

$$b(o_t) = \sum_{m=1}^{M} c_m \mathcal{N}(o_t; \mu_m, \Sigma_m). \tag{2.11}$$

Os pesos c_m deverão satisfazer a $\sum_{m=1}^{M} c_m = 1$ de modo a que $\int_{-\infty}^{\infty} b(o_t) = 1$. Este somatório ponderado de funções gaussianas é designado por *mistura de gaussianas* e é particularmente adequado para aproximar qualquer função densidade de probabilidade contínua (Huang et al., 1990). Na subsecção 2.3.5 é descrito um procedimento de estimação destas misturas que se integra de forma simples no treino dos modelos de Markov não observáveis.

Em determinadas circunstâncias, interessa isolar determinados subconjuntos de parâmetros em S funções densidade de probabilidade separadas b_s ("streams"). Neste caso, a probabilidade conjunta vem dada por

$$B(o_t) = \prod_{s=1}^{s=S} [b_s(o_t)]^{\gamma_s},$$

em que γ_s representa o peso de cada subconjunto. Tal aplica-se, por exemplo, quando se utilizam simultaneamente os parâmetros referentes ao cepstrum, ao delta-cepstrum e às respectivas acelerações, em que cada um destes tipos de parâmetros é descrito por uma densidade gaussiana ou por uma mistura de gaussianas (Young et al., 1996).

2.3 Modelos de Markov não observáveis

No capítulo de introdução foram referidos muitos dos aspectos que conferem ao sinal de fala características não determinísticas. A modelação estocástica é um método flexível e geral para este tipo de problemas.

Muitos dos métodos de reconhecimento de padrões tradicionais baseiam-se na ideia de determinar um padrão típico da sequência de tramas de fala através do cálculo de médias e usando medidas de distância do espectro local. Outra característica exigida para o reconhecimento de fala é a capacidade de se alinharem temporalmente os referidos padrões de forma a poderem representar diferenças de ritmo de fala entre diversas locuções da mesma palavra. Neste caso é reconhecido o sucesso do método de programação dinâmica geralmente designado por DTW ("dynamic time warping").

Este tipo de técnicas baseadas na ideia do padrão típico, não podem ser consideradas no sentido restrito, como técnicas estatísticas de modelamento de sinal. Contudo, as

técnicas estatísticas propriamente ditas, têm sido usadas sistematicamente para agrupamento ("clustering") na criação dos padrões de referência. A designação mais adequada será a de técnicas não paramétricas em que várias sequências de referência são usados para caracterizar as diversas sequências possíveis. Deste modo, a caracterização estatística do sinal baseada na representação de um padrão típico é apenas implícita e usualmente inadequada.

Os modelos de Markov não observáveis representam uma técnica estocástica adequada para problemas de dados incompletos associados a séries temporais (Huang et al., 1990). Na última década, estes modelos têm sido aplicados sistematicamente e com sucesso no reconhecimento de fala. Grande quantidade de ferramentas têm vindo a ser desenvolvidas, quer teóricas, quer práticas, contribuindo para que estes modelos adquirissem maior versatilidade e de modo a suportarem novos paradigmas. Nomeadamente, a modelação de elementos subpalavra a par com os conhecimentos da fonética acústica permitiram avanços práticos na integração com a linguística e com o processamento da língua natural.

As técnicas de processamento com redes neuronais artificiais constituem actualmente uma alternativa competitiva para o reconhecimento de fala. Destaca-se, nomeadamente, a possibilidade de utilização em tempo real (Yu e Oh, 1997) e o uso de modelos híbridos integrando estas técnicas com as dos modelos de Markov não observáveis (Bourlard e Wellekens, 1988; Morgan e Bourlard, 1990; Clary e Hansen, 1992; Neto et al., 1995; Neto et al., 1996; Neto, 1998).

2.3.1 Processos de Markov

Um processo estocástico $x(t)$ pode ser interpretado como uma regra para atribuir a cada acontecimento ou amostra ζ de uma experiência $\mathcal{X}$ uma função $x(t,\zeta)$ que se assume aqui definida no tempo (t). Assim, um processo estocástico é uma família de funções no tempo dependendo do parâmetro ζ. Equivalentemente, é uma função de t e de ζ em que o domínio de ζ é o de todas as realizações da experiência $\mathcal{X}$ e o domínio de t $(\mathcal{F})$ é, em termos gerais, um conjunto de números reais.

Se $\mathcal{F} = \mathcal{R}$, (eixo real) então $x(t)$ é um processo contínuo no tempo. Se $\mathcal{F} = \mathcal{I}$ (conjunto de números inteiros) então $x(t)$ é um processo discreto, ou seja, uma sequência de variáveis aleatórias representável por $x(t_n)$. Define-se o espaço de estados como o conjunto de valores distintos assumidos pelo processo estocástico. Se este espaço for contável ou finito, o processo é designado *de estados discretos* ou *cadeia*. De outro modo, será designado por processo de estados contínuos.

Um processo de Markov é um processo estocástico no qual o valor de $x(t_n)$ num dado instante t_n é completamente especificado a partir de $x(t_{n-1})$, com $t_{n-1} \leq t_n$. Esta é a conhecida *propriedade de Markov* definida pela condição

$$Pr[x(t_n) \leq X_n | x(t), t \leq t_{n-1}] = Pr[x(t_n) \leq X_n | x(t_{n-1})], \tag{2.12}$$

Se se considerar apenas os instantes: $t_1 < t_2 < \cdots < t_n$ obtém-se

$$Pr[x(t_n) \leq X_n | x(t_{n-1}), \ldots, x(t_1)] = Pr[x(t_n) \leq X_n | x(t_{n-1})]. \tag{2.13}$$

Esta definição serve também para processos discretos no tempo, se se substituir $x(t_n)$ por x_n (Papoulis, 1984). Designam-se *de primeira ordem* os processos de Markov assim definidos. Em geral podem considerar-se processos semelhantes que dependem dos n valores mais recentes e que, consequentemente, são designados por processos de Markov de n-ésima ordem. No que se segue, consideram-se apenas processos de Markov de estados discretos e discretos no tempo, ou seja, cadeias discretas de Markov.

2.3.2 Cadeias de Markov discretas

Uma cadeia de Markov para tempo discreto é um processo de Markov com um número contável de estados. Por conveniência da notação, nas secções seguintes representam-se os valores do espaço de estados por s_{t_n} ou, abreviadamente s_t, em vez de $x(t_n)$. Na representação s_t assume-se t como uma variável inteira.

Considere-se um conjunto finito N de estados. A cada instante de tempo discreto t é atribuído um único dos N estados. Uma cadeia de Markov é especificada a partir das probabilidades $Pr(s_t = i)$ de cada estado i (usadas para estimar um estado inicial) e das probabilidade de permanecer num estado i no instante seguinte ou de transitar para um dos outros estados j

$$a_{ij}(t) = Pr(s_{t+1} = j | s_t = i) \qquad 1 \leq i, j \leq N.$$

Nos processos de Markov de primeira ordem e de acordo com a propriedade de Markov expressa na equação 2.12, apenas se considera a dependência do último estado visitado. Além disso, nesta dissertação considera-se a probabilidade a_{ij} como independente do instante de tempo em que ocorre a transição entre os estados i e j.

$$a_{ij} = Pr(s_{t+1} = j | s_t = i) = Pr(s_{t+k+1} = j | s_{t+k} = i),$$

com $1 \leq i, j \leq N, \forall k \geq 0$. As cadeias que satisfazem esta propriedade são designadas por *cadeias homogéneas.*

Definição: considere-se uma matriz $\Gamma(P \times Q)$ de elementos γ_{ij} com as seguintes propriedades:

$$\gamma_{ij} \geq 0 \qquad \sum_{j=1}^{Q} \gamma_{ij} = 1, \tag{2.14}$$

com $1 \leq i \leq P$ e $1 \leq j \leq Q$, então Γ é uma *matriz de Markov*.

As probabilidades de transição são representadas na designada matriz das probabilidades de transição $A = \{a_{ij}\}$ a qual, de acordo com a definição anterior, é uma matriz de Markov quadrada (N^2). A cadeia de Markov homogénea pode ser completamente descrita a partir desta matriz e do conjunto de probabilidades de cada estado ser o estado inicial

$$\Pi = \{\pi_i : \pi_i = Pr(s_1 = i)\}, \qquad 1 \leq i \leq N.$$

Se os elementos da matriz A, forem nulos ($a_{ij} = 0$) para $i \geq j$, então a respectiva cadeia $S = \{s_1, s_2, \ldots, s_m\}$ é finita com $m \leq N$. Por outro lado, se $a_{ij} = 0$ apenas para $i > j$, a dimensão de S pode ser muito superior a N (*modelos esquerda-direita*). Se existirem elementos $a_{ij} > 0$ por cima e por baixo da diagonal principal de A, a dimensão de S poderá ser infinita. A título de exemplo, numa primeira abordagem ao fenómeno da língua escrita, considere-se que cada estado corresponde a uma letra do alfabeto. Desta forma pode obter-se um modelo, embora simplista, para o processamento da língua natural. A limitação mais evidente prende-se com as restrições impostas pela própria propriedade de Markov. O mesmo acontece em relação ao sinal de fala.

2.3.3 Estados não observáveis

O processo estocástico anteriormente descrito é designado por modelo observável de Markov, porque se pode considerar a sequência de estados como a saída do sistema em que cada estado corresponde a um acontecimento observável ou observação. Assim, o processo é completamente descrito a partir da saída. Este modelo é demasiado limitado, nomeadamente, para os fins em vista. Considere-se agora que não existe uma observação única gerada por cada estado visitado. Em vez disso, cada estado j possui um gerador aleatório de observações o_t cujas características são descritas por uma função probabilidade:

$$b_j(o_t) = Pr(o_t|s_t = j), \qquad 1 \leq j \leq N.$$

Obtém-se assim um novo processo estocástico gerado sobre o primeiro processo constituído pela sequência de estados. Esta sequência inicial já não é directamente observável mas apenas através desta espécie de véu constituído pelas probabilidades de observação. Este

tipo de modelos são assim designados por *modelos de Markov não observáveis* (Serralheiro, 1990) ou HMMs ("hidden Markov models"). Um modelo HMM pode ser representado por $\lambda = (A, B, \Pi)$, com $B = \{b_j(o_t)\}$.

2.3.4 Os problemas elementares dos HMMs

São geralmente identificados como três os problemas elementares dos HMMs de cuja resolução depende a sua aplicabilidade. As três secções seguintes descrevem esses problemas e apresentam as respectivas soluções.

A formulação seguinte costuma ser apresentada na literatura para observações discretas. Esta tradição tem na maior parte dos casos, intuitos didácticos, uma vez que a maioria dos trabalhos actuais utiliza na prática observações contínuas e consequentemente funções de probabilidade contínuas.

O problema da avaliação

Seguidamente, descreve-se o designado problema da avaliação ou do teste do modelo. Dada uma sequência de observações $O = \{o_1, o_2, \ldots, o_T\}$, pretende-se calcular a probabilidade $Pr(O|\lambda)$ de esta sequência ter sido produzida pelo modelo λ. Quando se dispõe de vários modelos, a solução deste problema permite decidir qual deles tem maior probabilidade de ter produzido determinada sequência para efeitos de classificação ou reconhecimento.

Para calcular a probabilidade $Pr(O|\lambda)$ directamente, enumeram-se todas as sequências possíveis de estados $\mathcal{S}$ com comprimento T (o número de observações). Uma sequência fixa de estados $S \in \mathcal{S}$ tem a seguinte probabilidade

$$Pr(S|\lambda) = \pi_{s_1} \prod_{t=1}^{T-1} a_{s_t, s_{t+1}}. \tag{2.15}$$

A probabilidade da sequência de observações O ter sido gerada a partir dessa sequência de estados é dada por

$$Pr(O|S, \lambda) = \prod_{t=1}^{T} b_{s_t}(o_t).$$

A probabilidade conjunta de O e S ocorrerem simultaneamente é dada por

$$Pr(O, S|\lambda) = Pr(O|S, \lambda) Pr(S|\lambda).$$

Assim, a probabilidade de O dado o modelo λ é dada por

$$Pr(O|\lambda) = \sum_{\mathcal{S}} Pr(O|S,\lambda).Pr(S|\lambda) \tag{2.16}$$

$$= \sum_{\mathcal{S}} \pi_{s_1} b_{s_1}(o_1). \prod_{t=2}^{T} a_{s_{t-1},s_t} b_{s_t}(o_t).$$

Para efectuar este cálculo são necessárias $(2T-1)N^T$ multiplicações e N^T-1 adições no total de $2N^T-1$ operações. Mesmo valores relativamente pequenos de N e de T não permitem que este cálculo seja praticável. Uma forma de analisar este cálculo é o de que a cada instante t existem N estados possíveis de atingir. Como só existem N estados, todas as sequências de estado possíveis têm de passar repetidamente nestes N estados, qualquer que seja o comprimento da sequência. Esta análise permite reduzir muito o número de operações anteriores conforme se explicará de seguida.

Considere-se a probabilidade de obter a parte da sequência de observações gerada até ao instante t, quando nesse instante se chegou ao estado i com o modelo λ

$$\alpha_t(i) = Pr(o_1, o_2, \ldots, o_t, s_t = i|\lambda),$$

em que a probabilidade $\alpha_t(i)$ é designada por *progressiva* ("forward"). Estas probabilidades podem ser calculadas por indução pelas seguintes expressões:

$$\alpha_1(i) = \pi_i b_i(o_1), \qquad 1 \leq i \leq N;$$

$$\alpha_t(j) = b_j(o_t) \sum_{i=1}^{N} \alpha_{t-1}(i) a_{ij}, \qquad 2 \leq t \leq T, \qquad 1 \leq j \leq N. \tag{2.17}$$

Por fim, obtém-se a pretendida probabilidade da sequência de observações

$$Pr(O|\lambda) = \sum_{i=1}^{N} \alpha_T(i).$$

Esta forma de cálculo de $Pr(O|\lambda)$ é designada por *algoritmo progressivo*. O número de operações necessárias é agora de $N(N+1)(T-1)+N$ multiplicações e de $N(N-1)(T-1)$ adições no total de $2N^2(T-1)+N$ operações. Tendo em conta que N é em geral inferior a 10 e que T é muito superior a 2, em geral na casa das centenas, verifica-se que a economia de cálculo é aproximadamente da ordem de $N^{(T-2)}$ vezes menos operações com o algoritmo progressivo quando comparado com o cálculo directo (ex: para os valores típicos de N=5 e T=100 obtém-se aproximadamente 3×10^{68}). De forma semelhante, pode-se considerar

a probabilidade de obter as $T-t$ últimas observações até ao final da sequência no instante T, a partir da observação o_t produzida no instante t pelo estado i:

$$\beta_t(i) = Pr(o_{t+1}, o_{t+2}, \ldots, o_T, s_t = i|\lambda).$$

A probabilidade $\beta_t(i)$, designada por *regressiva*, ("backward") pode ser calculada por indução com as seguintes expressões:

$$\beta_T(i) = 1, \qquad 1 \leq i \leq N$$

e

$$\beta_t(i) = \sum_{j=1}^{N} a_{ij} b_j(o_{t+1}) \beta_{t+1}(j), \qquad 1 \leq t \leq T-1, 1 \leq i \leq N. \tag{2.18}$$

Obtém-se de novo a probabilidade da sequência de observações

$$Pr(O|\lambda) = \sum_{i=1}^{N} \pi_i b_i(o_1) \beta_1(i),$$

designando-se este procedimento por *algoritmo regressivo*. Os algoritmos progressivo e regressivo podem ser utilizados na resolução dos outros dois problemas elementares seguidamente apresentados.

O problema da descodificação

O problema da descodificação é habitualmente enunciado como o segundo problema dos modelos HMM. Dada uma sequência de observações $O = o_1, o_2, \ldots, o_T$ e um modelo λ, pretende-se determinar a sequência de estados $S = s_1, s_2, \ldots, s_T$ subjacente. Trata-se portanto de recuperar a informação escondida pelo modelo. Excluindo-se os modelos degenerados, não existe uma solução correcta para este problema. Existem contudo vários critérios razoáveis que podem ser seleccionados de acordo com o fim em vista. A utilização típica destas soluções prende-se com o estudo da estrutura do modelo, segmentação do sinal, obtenção de estatísticas para cada estado, etc. O critério mais utilizado consiste em escolher a sequência mais provável de estados. Ou seja, maximizar $Pr(S|O, \lambda)$ o que é equivalente a maximizar $Pr(S, O|\lambda)$. Determina-se a probabilidade mais alta de, ao fim das primeiras t observações, ter ocorrido uma sequência de estados que terminou no estado i:

$$\delta_t(i) = \max_{s_1, s_2, \ldots, s_{t-1}} Pr[s_1, s_2, \ldots, s_{t-1}, s_t = i, o_1, o_2, \ldots, o_t|\lambda].$$

Por indução obtém-se

$$\delta_t(j) = \max_{1 \leq i \leq N} [\delta_{t-1}(i) a_{ij}].b_j(o_t), \quad 2 \leq t \leq T, 1 \leq j \leq N. \tag{2.19}$$

Para se obter a melhor sequência de estados, é necessário tomar nota dos argumentos que maximizam a expressão anterior num vector que se representa por

$$\psi_t(j) = \arg\max_{1\leq i\leq N}[\delta_{t-1}(i)a_{ij}], \qquad 2 \leq t \leq T, \qquad 1 \leq j \leq N.$$

Desta forma estabelece-se um algoritmo que é iniciado com

$$\delta_1(i) = \pi_i b_i(o_1), \qquad \psi_1(i) = 0.$$

No final determinam-se os valores optimizados:

$$Pr^*(S, O|\lambda) = \max_{1\leq i\leq N}[\delta_T(i)];$$

$$s_T^* = \arg\max_{1\leq i\leq N}[\delta_T(i)];$$

$$s_t^* = \psi_{t+1}(s_{t+1}^*), \qquad t = T-1, T-2, \ldots, 1.$$

As últimas duas expressões permitem determinar a melhor sequência de estados, a que determina a melhor probabilidade $Pr^*(S, O|\lambda)$. Este algoritmo é conhecido pelo nome de algoritmo de *Viterbi*.

Como se viu, o algoritmo progressivo, ou o regressivo, pode ser utilizado para o cálculo da probabilidade $Pr(O|\lambda)$ que é igual à soma de $Pr(S, O|\lambda)$ (equação 2.16) para todas as sequências possíveis de estados $S \in \mathcal{S}$. No reconhecimento de fala substitui-se habitualmente $Pr(O|\lambda)$ pelo valor de $\max_{S\in\mathcal{S}}[Pr(O, S|\lambda)]$, determinado pelo algoritmo de Viterbi. A experiência mostra que os resultados obtidos podem não ser iguais, especialmente nos procedimentos de estimação dos parâmetros dos HMMs. Nesses casos, o algoritmo progressivo, ou o regressivo, pode ser mais robusto do que o de Viterbi. Contudo, os resultados obtidos são em geral muito semelhantes e o algoritmo de Viterbi é mais eficiente em termos computacionais, em particular se se tomarem os logaritmos das expressões anteriores (conversão dos produtos em adições). Ao mesmo tempo, determina a sequência mais provável dos estados, a qual pode ser necessária se se pretender uma segmentação explícita do sinal. Estas vantagens fazem do algoritmo de Viterbi um dos mais usados nos sistemas de reconhecimento de fala.

O problema da estimação ou do treino do modelo

O mais difícil dos problemas básicos dos HMMs é o do ajuste dos parâmetros destes modelos de forma a satisfazer determinado critério. Dada uma sequência de observações $O = o_1, o_2, \ldots, o_T$, pretende-se estimar os parâmetros de $\lambda = (A, B, \Pi)$ de forma a

maximizar a probabilidade $Pr(O|\lambda)$. Esta forma de optimização é conhecida pela designação de *critério da máxima verosimilhança* (MV)[5]:

$$\lambda_{MV} = \arg\max_{\lambda} Pr(O|\lambda).$$

Não existe um procedimento conhecido para determinar analiticamente os parâmetros do modelo que conduzem ao máximo global λ_{MV}. Pode-se, no entanto, escolher A, B e Π por forma a que $Pr(O|\lambda)$, seja localmente maximizada. Para tal são utilizados procedimentos iterativos, tal como o método do gradiente ou o algoritmo de Baum-Welch, também designado de *reestimação*. Este último foi utilizado no presente trabalho para a estimação dos parâmetros dos HMMs.

A formulação do algoritmo de reestimação seguidamente apresentada refere-se ao caso dos HMMs contínuos, ou seja, em que as probabilidades de observação são funções de probabilidade contínuas. Utiliza-se também a designação CHMMs, ("continuous hidden Markov models"). Neste caso, as expressões de probabilidades de observação condicionais ou conjuntas, $Pr(O, \ldots)$ devem ser substituídas por funções densidade de probabilidade na forma $f(O, \ldots)$. Para que estas funções possam ser estimadas de forma consistente é necessário considerar algumas restrições. A forma mais geral de representação de uma função densidade de probabilidade, para a qual o algoritmo de reestimação tem sido formulado, é a de uma mistura (subsecção 2.2.7) finita de M componentes da forma

$$b_j(o_t) = \sum_{k=1}^{M} c_{jk}\mathcal{N}(o_t, \mu_{jk}, \Sigma_{jk}), \qquad 1 \leq j \leq N, \tag{2.20}$$

em que c_{jk} é o coeficiente ou peso da k-ésima componente no estado j e $\mathcal{N}$ é qualquer função densidade log-côncava ou elíptica (gaussiana por exemplo). Pode demonstrar-se que um estado com uma mistura de densidades de probabilidade de observação é equivalente a um conjunto de M estados, cada um deles com apenas uma destas densidades (Rabiner e Juang, 1993).

Sem perda de generalidade, assume-se que $\mathcal{N}$ é gaussiana com vector de médias μ_{jk} e matriz de covariância Σ_{jk} para a k-ésima componente no estado j. Os pesos das componentes satisfazem a seguinte restrição estocástica

$$\sum_{k=1}^{M} c_{jk} = 1, \qquad c_{jk} \geq 0, 1 \leq j \leq N, 1 \leq k \leq M,$$

[5]Deve-se referir o facto de existirem outros critérios possíveis (Franco e Serralheiro, 1991). Contudo, o critério MV tem sido o mais utilizado no reconhecimento de fala, dispondo dos métodos mais robustos em termos numéricos.

de forma a que a função densidade de probabilidade seja normalizada adequadamente, isto é

$$\int_{-\infty}^{\infty} b_j(o)do = 1, \qquad 1 \leq j \leq N.$$

Considere-se $a_{s_0,s_1} = \pi_{s_1}$ para maior simplicidade. A densidade de probabilidade de se observar O para uma sequência de estados S no modelo λ vem expressa por

$$f(O,S|\bar{\lambda}) = \prod_{t=1}^{T} \bar{a}_{s_{t-1},s_t} b_{s_t}(o_t)$$

$$= \sum_{k_1=1}^{M}\sum_{k_2=1}^{M}\cdots\sum_{k_T=1}^{M}[\prod_{t=1}^{T}\bar{a}_{s_{t-1},s_t}b_{s_tk_t}(o_t)]c_{s_1k_1}c_{s_2k_2}\cdots c_{s_Tk_T}.$$

Considera-se: Ω^T como o produto cartesiano de ordem T de $\Omega = \{1,2,\ldots,M\}$; $K \in \Omega^T$ um conjunto de índices k_t das gaussianas $\mathcal{N}(o_t,\mu_{s_tk_t})$ e dos respectivos pesos $c_{s_tk_t}$ associados a uma sequência de estados S. Assim, é possível considerar a densidade de probabilidade (Huang et al., 1990):

$$f(O,S,K|\bar{\lambda}) = \prod_{t=1}^{T}\bar{a}_{s_{t-1},s_t}c_{s_tk_t}\mathcal{N}(o_t,\mu_{s_tk_t},\Sigma_{s_tk_t})$$

e a partir desta expressão determinar a densidade de probabilidade conjunta de observação de O

$$f(O|\bar{\lambda}) = \sum_{S\in\mathcal{S}}\sum_{K\in\Omega^T} f(O,S,K|\lambda).$$

Considere-se agora a função auxiliar de Baum (Kullback e Leiber, 1951; Baum et al., 1970)

$$Q(\lambda,\bar{\lambda}) = \frac{1}{f(O|\lambda)}.\sum_{S\in\mathcal{S}}\sum_{K\in\Omega^T} f(O,S,K|\lambda).\log f(O,S,K|\bar{\lambda}),$$

com

$$\log f(O,S,K|\bar{\lambda}) = \sum_{t=1}^{T}\log\bar{a}_{s_{t-1},s_t} + \sum_{t=1}^{T}\log\bar{c}_{s_t,k_t} + \sum_{t=1}^{T}\log\mathcal{N}(o_t,\bar{\mu}_{s_t,k_t},\bar{\Sigma}_{s_t,k_t}).$$

Prova-se que a maximização desta função em ordem a $\bar{\lambda}$, com λ fixo, implica o aumento da verosimilhança (Baum, 1972):

$$Q(\lambda,\bar{\lambda}) \geq Q(\lambda,\lambda) \Rightarrow f(O|\bar{\lambda}) \geq f(O|\lambda).$$

O designado algoritmo de Baum-Welch, efectua esta maximização de forma iterativa substituindo λ pelos parâmetros $\bar{\lambda}$ determinados após maximização de $Q(\lambda,\bar{\lambda})$ na iteração

anterior. A função Q é maximizada através de técnicas conhecidas de optimização com restrições (Huang et al., 1990).

Os modelos HMM utilizados no reconhecimento de fala devem representar convenientemente as variações acústicas referentes a diversas locuções (repetições) da mesma palavra ou elemento subpalavra. Estas repetições podem ser do mesmo orador ou de oradores diferentes no caso de se pretender um reconhecedor independente do orador (subsecção 1.1.3). O material de fala para o treino de cada modelo deve incluir um conjunto representativo destas repetições, ou seja, em termos da formulação dos HMMs, devem ser utilizadas simultaneamente múltiplas sequências de observações $O^R = \{O^1, \ldots, O^r\}$. Assumindo que estas sequências de observações são independentes entre si, a estimação dos parâmetros baseia-se na maximização de

$$\log f(O^R|\lambda) = \sum_{n=1}^{r} f(O^n|\lambda).$$

Para tal, a função Q pode ser redefinida (Q^R) como a soma de funções parcelares Q^n referentes a cada sequência de observações O^n. Considerem-se as seguintes funções densidade de probabilidade:

$$\begin{aligned} \gamma_t^n(i,j) &= f(s_t = i, s_{t+1} = j|O^n, \lambda) \\ &= \frac{\alpha_t^n(i) a_{ij} \beta_{t+1}^n(j) \sum_{k=1}^{M} c_{jk} \mathcal{N}(o_{t+1}^n, \mu_{jk}, \Sigma_{jk})}{\sum_i \alpha_{T_n}^n(i)}; \end{aligned} \tag{2.21}$$

$$\begin{aligned} \gamma_t^n(i) &= f(s_t = i|O^n, \lambda) \\ &= \frac{\alpha_t^n(i)\beta_t^n(i)}{\sum_i \alpha_{T_n}^n(i)}; \end{aligned} \tag{2.22}$$

$$\begin{aligned} \zeta_t^n(j,k) &= f(s_t = j, k_t = k|O^n, \lambda) \\ &= \frac{\beta_t^n(j) c_{jk} \mathcal{N}(o_t^n, \mu_{jk}, \Sigma_{jk}) \sum_i \alpha_{t-1}^n(i) a_{ij}}{\sum_i \alpha_{T_n}^n(i)}. \end{aligned} \tag{2.23}$$

Como resultado da maximização de Q obtêm-se as seguintes equações de reestimação:

$$\bar{a}_{ij} = \frac{\sum_{n=1}^{r} \sum_{t=1}^{T_n - 1} \gamma_t^n(i,j)}{\sum_{n=1}^{r} \sum_{t=1}^{T_n - 1} \gamma_t^n(i)}; \tag{2.24}$$

$$\bar{c}_{jk} = \frac{\sum_{n=1}^{r} \sum_{t=1}^{T_n} \zeta_t^n(j,k)}{\sum_{n=1}^{r} \sum_{t=1}^{T_n} \gamma_t^n(j)}; \tag{2.25}$$

$$\bar{\mu}_{jk} = \frac{\sum_{n=1}^{r}\sum_{t=1}^{T_n}\zeta_t^n(j,k)o_t^n}{\sum_{n=1}^{r}\sum_{t=1}^{T_n}\zeta_t^n(j,k)}; \tag{2.26}$$

$$\bar{\Sigma}_{jk} = \frac{\sum_{n=1}^{r}\sum_{t=1}^{T_n}\zeta_t^n(j,k)(o_t^n - \bar{\mu}_{jk})(o_t^n - \bar{\mu}_{jk})^t}{\sum_{n=1}^{r}\sum_{t=1}^{T_n}\zeta_t^n(j,k)}. \tag{2.27}$$

2.3.5 Aspectos da implementação

A implementação prática dos modelos HMM é considerada em duas fases distintas: a do treino e a do teste. Na fase do treino estimam-se os modelos a serem utilizados na fase do teste, na qual se consuma o objectivo do reconhecimento automático da fala. Os sinais de fala utilizados na fase de treino são quase sempre diferentes dos utilizados na fase de teste. Desta forma, pretende-se garantir a representatividade das medidas de avaliação dos resultados obtidos na fase de teste. Por outras palavras, pretende-se que estas sejam independentes dos dados de treino (secção 2.7).

Conforme se referiu no capítulo da introdução, os reconhecedores independentes do orador são muito importantes nas aplicações do reconhecimento de fala. Estes reconhecedores exigem, em particular, um conjunto de sinais de teste de um número significativo de oradores, os quais devem ser distintos dos recrutados para a produção do conjunto de sinais de treino. As experiências de reconhecimento realizadas no decorrer do presente trabalho são exclusivamente deste tipo.

Na maioria dos casos, as fases de treino e de teste são repetidas em sucessivas iterações, durante a quais se procura eliminar alguns dos erros detectados na fase de teste, alterando a fase de treino. De acordo com esta estratégia, designam-se estes testes *de desenvolvimento*, considerando-se por vezes um outro teste final no qual se utiliza um conjunto de sinais de fala diferente dos anteriores. Nas experiências realizadas no âmbito desta dissertação, não se considerou este tipo de testes.

Geração dos parâmetros iniciais a reestimar

Neste trabalho utilizaram-se topologias lineares[6] para os modelos HMM (figura 2.1). Considera-se geralmente a existência de um estado inicial e de outro final que não emitem observações sendo por isso designados por *não emissores*. Assim, os restantes estados

[6]No capítulos 4,5 e 6 foram também empregues outras topologias mais complexas.

são geralmente designados de *emissores*[7]. A topologia de um HMM pode ser imposta inicializando a zero as probabilidades de transição não previstas.

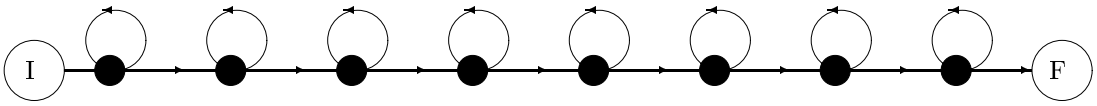


Figura 2.1: Topologia linear um modelo HMM para uma palavra com oito estados emissores e sem excluir qualquer estado intermédio.

A fase de treino dos modelos HMM resolve um problema de maximização utilizando o algoritmo de Baum-Welch. Contudo, a solução geralmente obtida corresponde apenas a um máximo local. De modo a que este máximo seja o mais próximo possível do máximo global, é necessário obter estimativas com alguma qualidade para os parâmetros iniciais dos modelos HMM. Para tal, desenvolveu-se um procedimento de inicialização destes parâmetros, cujos passos elementares são seguidamente descritos:

1. Segmentação uniforme. Cada sinal destinado ao treino de um dado HMM é seccionado em segmentos com igual duração e em número igual ao número de estados necessários para percorrer o modelo do estado inicial ao estado final[8]. Este procedimento é designado por *segmentação uniforme*.

2. Estimação das probabilidade de observação (B). Com base nos segmentos associados a cada estado no passo anterior estimam-se as distribuições gaussianas correspondentes às densidades de probabilidade de observação para cada estado.

3. Inicialização das probabilidades de transição (A). Estas probabilidades foram inicializadas por forma a impor uma topologia linear do tipo representado na figura 2.1. De acordo com o procedimento disponível no primeiro reconhecedor utilizado (SIRtrain, 1991; Jacobsen, 1992), atribuiu-se o valor de 0,6 à probabilidade de transição de um estado emissor para si próprio e de 0,4 para o estado seguinte. No último estado emissor estes valores são substituídos por 0,7 e 0,3 respectivamente.

4. Segmentação com o algoritmo de Viterbi. Com os passos anteriores obtêm-se todos os valores dos parâmetros do modelo λ. Com este modelo inicial determina-se uma segmentação do sinal utilizando o algoritmo de Viterbi (subsecção 2.3.4). Deste

[7] Os estados não emissores são particularmente importantes no reconhecimento de fala contínua onde representam o elemento de ligação entre os modelos das diversas palavras ou unidades subpalavra.

[8] Este procedimento só funciona com os modelos esquerda-direita existindo, contudo, outras alternativas para modelos ergódicos (Young et al., 1996).

modo, é possível atribuir a cada observação um determinado estado, de acordo com a sequência mais provável desses estados.

5. Paragem ou regresso ao passo anterior. Com os resultados obtidos no passo anterior é possível calcular novos parâmetros para o modelo λ. Com o novo modelo, o qual deverá apresentar parâmetros com mais qualidade do que o modelo inicial, repete-se o passo anterior. Estes dois passos são repetidos num processo iterativo que só é interrompido se se verificar uma convergência do valor global da verosimilhança acumulada de todas as locuções ou, caso esta não se verifique, ao fim de realizadas 10 iterações.

O escalamento das probabilidades

A partir das equações 2.17 e 2.18 verifica-se que, para valores elevados de t, as probabilidades progressivas e regressivas tendem para valores de tal forma baixos, que nem os processadores com precisão dupla (64 bits) lhes podem atribuir um valor diferente de zero. Existem dois tipos de soluções para este problema: o uso de logaritmos e o escalamento.

Em termos computacionais, as operações de produto e divisão apresentam geralmente custos superiores às de adição e subtracção. É assim particularmente conveniente a aplicação de logaritmos numa expressão matemática contendo exclusivamente produtos e divisões. Contudo, este procedimento não é aplicável nas restantes expressões que originalmente incluem adições e subtracções. Nestes casos, utiliza-se um *escalamento*, ou seja, multiplica-se a variável em questão por um factor de escala, por forma a manter $\alpha_t(i)$ e $\beta_t(i)$ numa gama de precisão aceitável (Rabiner e Juang, 1993).

Treino de múltiplas componentes gaussianas

Conforme se referiu, o uso de misturas de gaussianas é adoptado com frequência para modelar adequadamente as funções densidade de probabilidade de observação (subsecção 2.2.7). No presente trabalho, utilizou-se uma estratégia para o treino destas misturas, integrada no algoritmo de reestimação e denominada de *divisão de misturas* ("mixture splitting") (Young et al., 1996). Esta estratégia considera como ponto de partida modelos devidamente estimados com uma única gaussiana representando a função densidade de probabilidade de observação de cada estado. Para cada uma destas gaussianas são geradas duas cópias, nas quais as respectivas médias são ligeiramente alteradas por um valor igual a um quinto do respectivo desvio padrão. Assim, numa das cópias este valor é adicionado à média enquanto na outra lhe é subtraído. Os pesos de cada componente são iguais a

metade do peso da gaussiana original (unitário uma vez que até agora existia apenas uma única componente). Posteriormente, os novos modelos assim criados, são sujeitos a um novo ciclo de reestimações. O processo pode continuar, de forma iterativa, a produzir mais componentes para cada mistura de gaussianas.

Nas iterações seguintes é necessário escolher a componente gaussiana que será eliminada para dar origem a duas novas componentes. Em primeiro lugar, seleccionam-se as componentes resultantes de um menor número de divisões, depois, é escolhida aquela que apresente um peso com maior valor. Assume-se que esta componente representa um maior número de observações, sendo aquela que, potencialmente, necessita de uma representação mais detalhada. Esta quantidade acrescida de material de treino deverá, também, de permitir o modelamento mais consistente de uma componente adicional.

2.4 Modelos semicontínuos

Nos modelos descritos nas secções anteriores, as probabilidades de observação eram determinadas a partir de uma distribuição contínua (uma gaussiana ou uma mistura de gaussianas). Os primeiros HMMs utilizados no reconhecimento da fala não utilizavam estas distribuições. Em vez destas produzia-se, na fase de treino, um conjunto com um número limitado (L=256 (Lee, 1989)) de vectores, designado por *dicionário das observações*. Estes vectores constituem protótipos representativos das observações de fala a reconhecer e são determinados utilizando um método de quantificação vectorial. Para cada estado (i) dos modelos HMM e para um (v_j) dos vectores do dicionário das observações, estimava-se a respectiva probabilidade $b_i(v_j)$. Para qualquer vector de observação O_t é possível, através de uma medida de similaridade adequada, determinar a entrada do dicionário das observações correspondente. Desta forma existe também um número finito de probabilidades de observação que podem ser representadas numa matriz ($L \times N$). Por forma a distinguir estes HMMs dos baseados em probabilidades de observação contínuas designam-se os primeiros por *modelos discretos* e os segundos por *modelos contínuos*.

Os modelos discretos apresentavam uma formulação teórica mais simples que a dos contínuos. Os cálculos envolvidos eram mais simples, permitindo implementações mais eficazes em termos de tempo de processamento. Além disso, permitiam modelar qualquer distribuição sem exigirem grandes quantidades de material para a estimação dos parâmetros. O problema mais relevante prende-se com a utilização do dicionário das observações. Este tipo de solução introduz inevitavelmente erros de quantificação. Outro inconveniente, associado a este dicionário, reside no facto deste ser gerado antes do pro-

cesso de treino dos HMMs e não ser posteriormente reestimado em conjunto com esses modelos. Os modelos contínuos não apresentam este tipo de problema, garantindo um modelamento acústico detalhado. Uma outra vantagem, habitualmente referida para os modelos contínuos, reside no facto de ser possível melhorar as taxas de reconhecimento (secção 2.7) substituindo o critério de MV pelo critério de máxima informação mútua ("maximum mutual information") (Huang et al., 1990). Contudo, estes aumentos no desempenho não são verificados com a utilização deste mesmo critério para os modelos discretos.

Referem-se seguidamente alguns inconvenientes dos modelos contínuos:

- Assumem mais pressupostos do que os necessários para os modelos discretos, exigindo particular atenção os referentes à utilização da matriz de covariância diagonal;
- Se por um lado, o uso de uma única componente gaussiana pode introduzir singularidades (Nadas, 1983), o uso de misturas de gaussianas aumenta a complexidade computacional. De facto, aumenta o número de parâmetros a estimar, pelo que se torna também necessária uma quantidade superior de observações;
- Para o caso do reconhecimento independente do orador, é importante a utilização das misturas de gaussianas, por forma a ser possível a modelação eficaz da variabilidade acrescida destes sinais.

Os problemas referidos a propósito dos modelos discretos e contínuos motivaram o desenvolvimento de novos modelos baseados nos HMMs que conjugam alguns dos pressupostos anteriores e que se revelaram promissores para a obtenção de melhores desempenhos no reconhecimento automático da fala: os modelos semicontínuos e os modelos discretos com múltiplos dicionários de quantificação.

Os modelos HMM designados por *semicontínuos* ou SCHMMs, ("semi-continuous hidden Markov models") substituem o dicionário das observações dos modelos discretos por um *dicionário de gaussianas*. Nos modelos discretos, em cada estado existem probabilidades associadas para cada protótipo de vector de observações contido no dicionário das observações. Nos modelos semicontínuos, estas probabilidades são substituídas pelos pesos c_{jk} que associam gaussianas (existentes no dicionário de gaussianas) em misturas equivalentes àquelas que podem ser obtidas nos modelos contínuos (equação 2.20). Consegue-se, deste modo, reduzir o número de parâmetros a estimar relativamente aos modelos contínuos, atenuando simultaneamente os efeitos referentes ao erro de quantificação nos modelos discretos.

Os modelos discretos com múltiplos dicionários de quantificação foram apresentados mais recentemente (Segura et al., 1994) e não obtiveram uma divulgação na comunidade científica semelhante à obtida com os modelos semicontínuos. Baseiam-se na utilização de dicionários de quantificação específicos para cada palavra. Os testes preliminares apresentaram desempenhos superiores aos obtidos com os modelos discretos. Contudo, em relação aos modelos semicontínuos, só os superaram quando os dicionários de gaussianas eram de pequena dimensão, (64 e 128 gaussianas) verificando-se um desempenho ligeiramente inferior para dimensões maiores (256 e 512).

2.4.1 Equações dos modelos semicontínuos

A função densidade de probabilidade do vector de observação o condicionada para que este seja produzido a partir de um estado s_t do modelo semicontínuo pode ser descrita por:

$$f(o_t|s_t) = \sum_{j=1}^{L} f(o_t|v_j, s_t) Pr(v_j|s_t),$$

em que L é a dimensão do dicionário com entradas $v_j : 1 < j < L$. Para maior simplicidade, considera-se $f(o_t|v_j, s_t) = f(o_t|v_j)$ independente do estado de Markov $i = s_t$. Assumindo que $f(o_t|v_j)$ é uma gaussiana e que v_j são os respectivos parâmetros, obtém-se

$$f(o_t|i) = \sum_{j=1}^{L} \mathcal{N}(o_t, \mu_j, \Sigma_j) c_{ij}, \tag{2.28}$$

em que $c_{ij} = Pr(v_j|i)$ é o peso da componente j na mistura de gaussianas associada ao estado i.

As probabilidades de transição a_{ij} podem ser reestimadas de acordo com a expressão 2.24, apresentada para os modelos contínuos. A expressão de reestimação dos pesos de cada componente gaussiana c_{ij} tem uma forma idêntica à da equação 2.25 dos modelos contínuos. Contudo, para que se possa utilizar esta expressão sem ambiguidade, deve-se redefinir o significado de $\zeta_t(j, k)$

$$\zeta_t(j, k) = f(s_t = j, o_t \sim v_k|O, \lambda) \tag{2.29}$$

e consequentemente de

$$\zeta_t(k) = \sum_i \zeta_t(i, k) = f(o_t \sim v_k|O, \lambda), \tag{2.30}$$

em que $o_t \sim v_k$ significa que o_t é quantificado em v_k.

As equações de reestimação 2.26 e 2.27, respectivamente para as médias e as covariâncias dos modelos contínuos, podem facilmente ser convertidas para os modelos semicontínuos. Para tal, basta substituir os dois índices referentes à componente k da mistura de gaussianas do estado j do modelo contínuo, pelo índice do vector v_j do dicionário de gaussianas.

Por forma a reduzir o número de parâmetros sem perda de informação, é importante utilizar um único dicionário de gaussianas para todos os modelos que se pretende estimar. Assume-se assim que seja possível estimar, de forma optimizada, um conjunto de gaussianas recorrendo a todo o material de fala disponível para o treino do reconhecedor. Estas gaussianas deverão permitir obter, por associação em misturas, uma descrição adequada das densidades da probabilidade de observação $f(o_t|s_t)$. De facto, é possível determinar equações de reestimação por aplicação do designado *método EM* (expectância-maximização) de forma semelhante ao efectuado para os modelos contínuos e utilizando todas as locuções destinadas à estimação de parâmetros do reconhecedor (Huang et al., 1990):

$$\bar{\mu}_j = \frac{\sum_{m=1}^{F}\sum_{n=1}^{r}\sum_{t=1}^{T_{mn}} \zeta_t^{mn}(j) o_t^{mn}}{\sum_{m=1}^{F}\sum_{n=1}^{r}\sum_{t=1}^{T_{mn}} \zeta_t^{mn}(j)}; \tag{2.31}$$

$$\bar{\Sigma}_j = \frac{\sum_{m=1}^{F}\sum_{n=1}^{r}\sum_{t=1}^{T_{mn}} \zeta_t^{mn}(j)(o_t^n - \bar{\mu}_j)(o_t^n - \bar{\mu}_j)^t}{\sum_{m=1}^{F}\sum_{n=1}^{r}\sum_{t=1}^{T_{mn}} \zeta_t^{mn}(j)}. \tag{2.32}$$

O índice m distingue os parâmetros e os dados o_t^{mn} destinados a cada um dos F modelos a serem estimados.

2.4.2 Redução do número de gaussianas por estado

Nos modelos semicontínuos, o algoritmo de reestimação é separado em dois ciclos de iterações. No primeiro ciclo são considerados todos os coeficientes c_{ij} que associam cada estado de Markov (i) às componentes (j) do dicionário de gaussianas. Verifica-se a posteriori que a maioria destes coeficientes apresenta um valor insignificante. Por forma a reduzir o número de parâmetros a estimar e o tempo de processamento, nomeadamente na fase de reconhecimento, (teste) assumiu-se que apenas seriam consideradas para cada estado as componentes com maior peso. Reescrevendo a equação 2.28 obtém-se:

$$f(o_t|i) = \sum_{j\in\mathcal{J}(i)}^{L} \mathcal{N}(o_t, \mu_j, \Sigma_j) c_{ij}, \tag{2.33}$$

em que $\mathcal{J}(i)$ é o conjunto dos índices das componentes seleccionadas para o estado i.

Ensaiou-se o seguinte critério de selecção: seleccionaram-se sucessivamente os valores mais elevados de c_{ij} até que a respectiva soma igualasse ou ultrapassasse o valor de 0,9. Com um dicionário com 128 entradas, obtêm-se deste modo em média cerca de 12 componentes gaussianas, ou seja, uma redução de uma ordem de grandeza no número total de coeficientes c_{ij}. Estes coeficientes são posteriormente normalizados

$$\hat{c}_{ij} = c_{ij} / \sum_{j \in \mathcal{J}(i)} c_{ij},$$

por forma a verificar-se

$$\sum_{j \in \mathcal{J}(i)} \hat{c}_{ij} = 1.$$

Após esta redução de componentes disponíveis para cada estado o processo de reestimação é retomado. Contudo, neste segundo ciclo o dicionário de gaussianas não é reestimado. Deste modo, obtiveram-se resultados muito semelhantes aos obtidos com a utilização de todas as componentes no dicionário. Realça-se o facto do número de componentes gaussianas determinadas para cada estado ser muito próximo dos números habitualmente considerados para os modelos contínuos.

Huang e os seus colegas sugeriram uma redução de componentes semelhante mas efectuada para cada observação $\mathcal{J}(o_t)$ (Huang et al., 1990). Em relação a esta proposta, o procedimento anterior apresenta a vantagem de se obterem modelos com um número de parâmetros c_{ij} reduzido a 10% do número inicial (assumindo-se os restantes como sendo nulos).

2.4.3 Inicialização do dicionário de gaussianas

Para iniciar o processo de reestimação de parâmetros dos modelos semicontínuos, é necessário dispor-se de estimativas iniciais que se aproximem, tanto quanto possível, dos valores óptimos. Para tal, utilizaram-se alguns dos procedimentos genéricos já descritos em relação aos modelos contínuos. Contudo, no caso dos modelos semicontínuos, é necessário inicializar uma estrutura diferente para as densidades de probabilidade de observação, a qual deve ser baseada em todas as observações disponíveis para o treino: o dicionário de gaussianas. Para tal, utilizou-se um procedimento relativamente simples que determina L partições de observações através de um procedimento de quantificação vectorial tradicional. A partir de cada partição, determina-se a gaussiana que melhor a descreve, de acordo com o princípio de máxima verosimilhança. O método de quantificação vectorial utilizado é uma extensão conhecida do método k-médias (subsecção 4.5.2):

o método LGB (Linde et al., 1980). A medida de distância utilizada foi a distância de Mahalanobis (Duda e Hart, 1973).

O método LGB implementado é descrito nos seguintes passos:

1. Inicialmente todas as observações são consideradas dentro da mesma partição $l = 1$. Determina-se a matriz de covariância com esta partição inicial que será utilizada na distância de Mahalanobis.
2. São calculados l centróides como sendo a média de todas as observações dentro de cada partição.
3. Cada centróide é perturbado em $\pm\epsilon$ por forma a gerar dois novos centróides: um deles resulta do produto por $1{,}0+\epsilon$ e o outro por $1{,}0-\epsilon$. Duplica-se assim o número l de partições. Utilizou-se $\epsilon = 0{,}01$.
4. De acordo com os novos centróides, classificam-se todas as observações determinando-se novas partições.
5. O processo regressa ao passo 2 até que $l = L$. De acordo com este método, L deverá ser uma potência de dois.
6. Para cada partição final e de acordo com o princípio de máxima verosimilhança determinam-se as respectivas gaussianas.

2.4.4 Vantagens e desvantagens dos modelos semicontínuos

O conceito de estimar alguns dos parâmetros associados a diferentes estruturas dos modelos HMM como sendo iguais, tem sido explorado com sucesso no reconhecimento automático de fala. As vantagens obtidas são essencialmente devidas a uma estimação mais eficaz destes parâmetros para o mesmo conjunto de observações disponíveis. Este conceito é conhecido pela designação de *ligação de parâmetros* ("parameter tying"). Os modelos semicontínuos podem ser analisados como uma consequência directa da aplicação deste conceito aos modelos contínuos, recebendo, de acordo com esta perspectiva a designação de *misturas ligadas* ("tied mixtures") (Bellegarda e Nahamoo, 1990). Existem actualmente várias aplicações deste conceito a outros parâmetros, por exemplo, os designados por *densidades ligadas* ("tied density HMMs") (Euler, 1990), as *transições ligadas* ("tied transitions") (Young e Woodland, 1993) ou, mais recentemente, os *trifones de estados ligados* ("tied-state triphones") (Young e Bloothooft, 1997).

De seguida, resumem-se algumas das vantagens dos modelos semicontínuos:

- Permitem a obtenção de distribuições sobrepostas e não particionadas como acontece com os modelos discretos.
- Não apresentam tantos parâmetros como os modelos contínuos porque as descrições das funções contínuas se encontram *ligadas* (subsecção 2.4.4) através do uso do dicionário de gaussianas.
- Ao contrário dos modelos discretos, dispõem de uma formulação que lhes permite reestimar conjuntamente o dicionário de gaussianas com os restantes parâmetros do modelo.

É possível identificar duas desvantagens remanescentes dos modelos contínuos e discretos, respectivamente: mantêm-se os pressupostos referidos a propósito dos modelos contínuos e não se garante o anulamento completo da existência de erros de quantificação.

No capítulos 4 e 6 serão apresentados alguns resultados obtidos com um reconhecedor de fala baseado em modelos semicontínuos. Este reconhecedor foi desenvolvido exclusivamente no âmbito do presente trabalho para a realização destas experiências.

2.5 Reconhecimento baseado em unidades subpalavra

A utilização prática do reconhecimento automático de fala é particularmente importante no desenvolvimento de interfaces homem-máquina para tarefas complexas (subsecção 1.1.1). Por forma a facilitar eficazmente esta interacção, é essencial o reconhecimento de fala contínua (subsecção 1.1.3). Entre as aplicações mais complexas actualmente procuradas destacam-se os sistemas de ditado, de tradução automática e o acesso a bases de dados. Estes sistemas deverão ser capazes de reconhecer dezenas ou mesmo centenas de milhares de palavras se se pretender processar fala espontânea. A recolha de corpora de fala representativos destes léxicos e das respectivas variações acústicas representa uma tarefa em geral impraticável. A utilização de modelos de fala baseados em unidades subpalavra é essencial para o reconhecimento de vocabulários desta dimensão.

2.5.1 Unidades elementares da fala

A análise da fala em unidades elementares é um problema crucial da linguística conforme se deduz da seguinte definição de língua: "A língua é um instrumento de comunicação

segundo o qual, de modo variável de comunidade para comunidade, se analisa a experiência humana em unidades providas de conteúdo semântico e de expressão fónica — os monemas; esta expressão fónica articula-se por sua vez em unidades distintivas e sucessivas — os fonemas —, de número fixo em cada língua e cuja natureza e relações mútuas também diferem de língua para língua" (Martinet, 1967). No processamento automático do sinal de fala estas unidades podem assumir um carácter diferenciado da linguística, uma vez que os meios de análise do sinal, anteriormente descritos, não permitem integrar o conhecimento associado ao contexto sintáctico e semântico. Este contexto é apreendido de forma natural pelo ouvinte comum e é portanto um dado acessível para o estudo por parte dos linguistas.

O conceito de unidade subpalavra deriva do conceito mais geral referido por Serralheiro de "entidade básica" (Serralheiro, 1990). Este conceito refere-se a "segmentos do sinal de fala, cujas características sejam determináveis e reprodutíveis". Ou seja, desta forma condiciona-se a definição dos referidos segmentos à existência de um modelo adequado para cada um deles.

O problema da determinação da transcrição em unidades subpalavra pode ser analisado a vários níveis. Na base do problema, está a definição do conjunto de unidades subpalavra a utilizar.

A escolha de um destes conjuntos de unidades subpalavra obedece essencialmente a um compromisso entre a duração média do segmento de sinal que cada unidade representa e a capacidade de incorporar aspectos relacionados com a coarticulação dessas mesmas unidades. Quanto maior for a duração média dos referidos segmentos tal significa, em geral, que mais unidades diferentes serão necessárias para representar o sinal de fala. Assim, numa aplicação para fala sem restrições, as unidades de menor duração, tais como os fones, são da ordem da meia centena, enquanto que, no caso limite de se considerar cada palavra uma unidade, são necessárias algumas centenas de milhar. Por sua vez, quanto maior o número de unidades, maiores são as dificuldades em termos do respectivo treino. Nomeadamente, é necessário dispor-se de grande quantidade de sinais de fala, nos quais todas as unidades se encontrem representadas através de diversas realizações repetidas em número significativo. Por outro lado, o número de parâmetros e a complexidade em geral dos modelos adoptados para cada unidade deverá aumentar com a duração média correspondente a cada unidade.

Em termos de capacidade de incorporar aspectos referentes ao contexto ou coarticulação entre estas unidades, esta será obviamente menor quanto menor for a duração das referidas unidades. Estes problemas não se colocam com os reconhecedores mais rudi-

mentares em que se modelam frases inteiras ou se permite apenas a locução de palavras isoladas com os correspondentes modelos de palavras inteiras. Surgem, no entanto, num reconhecedor que utilize estes modelos para fala ligada, uma vez que a coarticulação entre palavras (ou interpalavra) não se encontra modelada. A estes acrescem os problemas de coarticulação intra-palavra que surgem com a utilização de unidades subpalavra e que se tornam particularmente evidentes para as unidades mais pequenas, tais como as baseadas em fones. Na secção 2.5.3, descrever-se-á uma estratégia utilizada para ultrapassar este problema.

A escolha de um conjunto de unidades elementares da fala, obedecendo a uma perspectiva estritamente linguística ou fonética, requer invariavelmente a supervisão de um operador humano ou, eventualmente, de um sistema pericial baseado em regras. Do ponto de vista acústico, é possível utilizar métodos automatizados baseados nas técnicas das áreas do processamento de sinal e do reconhecimento de padrões. Os métodos mais conhecidos para o reconhecimento de fala também derivam dessas áreas. Para além da economia de meios (humanos) na preparação dos corpora de fala, uma escolha de unidades subpalavra mais próxima da representação acústica pode contribuir para um melhor desempenho no reconhecimento automático de fala (Haeb-Umbach et al., 1995).

Enumeram-se algumas escolhas possíveis de unidades subpalavra que podem ser utilizadas na modelização do sinal da fala (Rabiner e Juang, 1993):

- unidades baseadas em fonemas;
- unidades baseadas em sílabas;
- unidades baseadas em semi-sílabas;
- unidades acústicas.

A escolha de unidades subpalavra baseadas em **fonemas** é justificada pelo facto de estes representarem segmentos mínimos, distintivos e sucessivos que servem para distinguir significantes (Barbosa, 1994). Os fonemas são entidades abstractas, definidas de acordo com critérios preponderantemente linguísticos. No entanto, a caracterização acessível aos métodos analíticos e consequentemente aos meios automáticos é, essencialmente, baseada em medidas de semelhança do tipo acústico. No caso em que as caracterizações fonéticas e acústicas se aproximam, como é o caso das vogais acentuadas, a sua utilização como unidades subpalavra poderá ser adequada. No processamento de fala, o conceito de fonema é, em geral, substituído pelo conceito de fone. O fone é a realização acústica

do fonema quando este é articulado. Quando existem diferentes realizações do mesmo fonema, denominam-se de alofones (Mateus et al., 1990).

Uma vez que o tracto vocal não é discreto mas pode variar de infinitas maneiras, um número infinito de fones pode corresponder a um único fonema. Contudo, no reconhecimento automático de fala, considera-se um número relativamente pequeno de fones, normalmente entre 40 a 50 unidades, quer no inglês quer no português. Assim, prevê-se que os aspectos relativos à variabilidade acústica sejam descritos pelos modelos considerados para cada um destes fones. Com este número reduzido de fones, facilita-se a obtenção de um vocabulário de treino que inclua a maioria ou mesmo a totalidade dos fones existentes numa língua.

Muita da informação que permite identificar os fones no sinal de fala encontra-se nas zonas de fronteira ou de transição entre fones, em particular entre vogais e consoantes e vice-versa. Uma outra entidade linguística que abrange algumas destas fronteiras é a **sílaba**. Esta encontra-se centrada numa vogal à qual se juntam consoantes iniciais ou finais. A utilização de unidades baseadas em sílabas está, tal como o fonema, condicionada desde logo pela dissociação entre a definição linguística e a respectiva realização acústica. Já no que se refere aos aspectos de coarticulação representa um melhoramento em relação ao fone. Uma consequência imediata disto e que se poderá confirmar com as restantes descrições de unidades subpalavra é o aumento exponencial do número de unidades a utilizar. No caso do inglês são cerca de 10.000, contudo, é possível considerar um número muito inferior de unidades semelhantes, designadas por *tipo-sílaba* (Hu et al., 1996).

O conceito de **semi-sílaba** está ligado ao de difone (secção 2.5.3) (Serralheiro, 1990). A utilização de unidades baseadas em semi-sílabas no processamento da fala surge essencialmente como forma de diminuir o número de unidades em relação às baseadas na sílaba, mantendo a vantagem de preservar as zonas de maior coarticulação. Efectivamente, o número de semi-sílabas no inglês é da ordem das 2.000. As semi-sílabas podem ser de dois tipos em alternativa: um grupo inicial de consoantes (ataque silábico) e parte do núcleo de uma vogal ou este último seguido de um grupo de consoantes final. Em geral, a cada sílaba correspondem duas semi-sílabas. Estas unidades foram identificadas como particularmente interessantes para o modelamento da língua espanhola no reconhecimento automático de fala. De facto, o inventário de semi-sílabas na língua espanhola é relativamente pequeno (menos de 750 unidades) (Mariño et al., 1990). No caso da língua alemã identificaram-se 344 grupos de semi-sílabas (54 de consoantes iniciais, 160 de consoantes finais e 130 referentes a vogais. Associando de forma automática estas semi-sílabas em unidades com duração superior, de acordo com um critério baseado na frequência de ocorrência das unidades escolhidas, obtiveram-se resultados de reconhecimento superiores,

em cerca de 10%, aos obtidos com fonemas (Pfau et al., 1997).

As **unidades acústicas** são definidas a partir do agrupamento de segmentos do sinal de fala. A segmentação é feita de acordo com um critério objectivo pré-definido. Uma das primeiras aplicações deste conceito no reconhecimento de fala (Lee et al., 1988) utiliza o critério de máxima verosimilhança. Recorrendo à quantificação vectorial tradicional (Linde et al., 1980) considera 128 unidades diferentes, obtendo desempenhos comparáveis aos obtidos com modelos de palavra. Outro estudo importante refere-se a unidades designadas por *fenónicas* ("fenonic") ou *fenones* (a mesma palavra em inglês) (Bahl et al., 1993). Estas unidades têm por base uma etiquetagem fonética com o intuito exclusivo de darem alguma indicação para uma inspecção humana directa. De resto, são determinadas de forma inteiramente não supervisionada, por um algoritmo de quantificação vectorial tradicional, com base na representação acústica do sinal. Utilizaram-se 200 fenones no sistema Tangora da IBM. Em geral, para um vocabulário de 1.000 palavras, utilizam-se 256 unidades acústicas e para a fala contínua corrente calcula-se que 1.024 unidades serão suficientes (Lee et al., 1990a).

Podem ser consideradas duas filosofias de base no reconhecimento de fala: a da acústico-fonética e a fonémica baseada no reconhecimento de padrões. Na primeira, o pressuposto de base é o de que a fala contínua pode ser segmentada em regiões bem definidas às quais podem ser atribuídas diversas etiquetas fonéticas baseadas em medidas sobre as características de cada segmento. Assume-se, portanto, que se pode encontrar uma caracterização universal dos parâmetros de unidades básicas de fala e que a fala pode ser etiquetada num encadeamento dessas unidades.

Na perspectiva designada por *fonémica* e baseada no reconhecimento de padrões, as unidades básicas da fala são modeladas acusticamente mas baseando-se na descrição lexical das palavras constantes no vocabulário. Não existe nenhum pressuposto, a priori, sobre o mapeamento entre as medidas acústicas e os fonemas. Esse mapeamento é aprendido a partir de um conjunto finito de frases de treino. As unidades obtidas são designadas por *tipo-fone* ("phone-like units" — PLUs). De modo semelhante definem-se as unidades *tipo-sílaba* (Hu et al., 1996). São essencialmente descrições acústicas de unidades linguísticas, tal como estão representadas nas palavras que ocorrem num dado conjunto de treino (Lee et al., 1990a).

Estas duas formas de abordagem do reconhecimento de fala têm sido exaustivamente estudadas em várias aplicações. A abordagem acústico-fonética foi estudada durante muito tempo, tendo sido utilizada no reconhecimento de vocabulários extensos, tal como nos sistemas Summit do MIT (Zue et al., 1989; Manos e Zue, 1997) e APHMM ("acoustic-

phonetic HMM") dos laboratórios da Bell (Levinson et al., 1989). De acordo com alguns investigadores, é considerada particularmente indicada para o reconhecimento de fala espontânea (Fukada et al., 1996).

Também a abordagem fonémica baseada no reconhecimento de padrões tem sido exaustivamente estudada e entre as suas aplicações mais representativas mencionam-se os sistemas da IBM (Jelinek, 1985), o sistema SPHINX da CMU (Lee, 1989; Lee et al., 1990b) e o sistema BYBLOS da BBN (Schwartz et al., 1989). Os desenvolvimentos mais recentes no âmbito da abordagem fonémica têm sido baseados em modelos que abrangem um contexto temporal mais alargado, envolvendo geralmente várias unidades tipo-fone (subsecção 2.5.3).

Embora ambas as abordagens apresentem vantagens e desvantagens distintas, a abordagem fonémica baseada no reconhecimento de padrões tem produzido consistentemente os melhores resultados de reconhecimento. Nas experiências de reconhecimento independentes do vocabulário descritas no presente dissertação, utilizaram-se unidades tipo-fone. Para maior simplicidade, estas unidades serão designadas daqui por diante apenas por fone.

2.5.2 Selecção de um inventário de fones

Na tabela 2.1 representa-se a lista de fones utilizados no presente trabalho. Nesta tabela procurou-se ainda estabelecer uma correspondência, ainda que aproximada, entre os fones escolhidos e os fonemas de um inventário de fonemas para a língua inglesa (O'Shaughnessy, 1987). Cada fonema encontra-se classificado em termos do modo de articulação e de vozeamento (Barbosa, 1994), tendo-se incluído ainda um exemplo de uma palavra contendo o respectivos fone (Lee, 1989).

Para a comparação de diversos inventários de fones, utilizaram-se os resultados de reconhecimento e a audição dos respectivos sinais. Para estes efeitos considerou-se exclusivamente o corpus disponível para oradores nativos (secção 3.2). A escolha das unidades representadas na tabela 2.1 seguiu um percurso baseado naquele que conduziu ao desenvolvimento do sistema SPHINX. Os seus autores começaram por adoptar um conjunto de fones muito semelhante ao utilizado no corpus TIMIT (Keating et al., 1994). Neste corpus foram atribuídas etiquetas diferentes ao mesmo fonema quando as propriedades espectrais eram diferentes. Por exemplo, o **r** existente nas palavras "bi**r**d" (/er/) e "dine**r**" ou "butte**r**" (/axr/) correspondem a fones diferentes. O mesmo acontece nas palavras: "boot" (/uw/) e "beauty" (/ux/); "mom" (/m/) e "yes'em" (/em/); "non", (/n/) "but-

fone	aprox.	articulação	exemplo
iy	i		beat
ih	I		bit
ix			roses
ey	e		bait
eh	ϵ		bet
ae	æ		bat
aa	α	vogal	cot
ao	$\supset$		bought
ow	o		boat
uh	U		book
uw	u		boot
ah	$\wedge$		butt
ax	$\Im$		the
ay	αj		bite
oy	$\supset$j	ditongo	boy
aw	αw		about
y	j	semi-	yet
w	w	consoante	wet
l	l		led
el		consoante	bottle
r	r	líquida	red
er			bird
sil		silêncio	—

fone	aprox.	articulação	vozea/	exemplo
m	m		sonora	mom
n	n	nasal	sonora	non
ng	η		sonora	sing
f	f		surda	fief
v	v		sonora	very
th	θ		surda	thief
dh	δ		sonora	they
s	s	fricativa	surda	sass
z	z		sonora	zoo
sh	$\int$		surda	shoe
hh	h		surda	hay
p	p		surda	pop
pd				ship
b	b		sonora	bob
t	t		surda	tot
td		oclusiva		set
d	d		sonora	dad
dd				deleted
k	k		surda	kick
kd				comic
g	g		sonora	gag
ch	$\check{c}$	semioclusiva	surda	church
jh	$\check{z}$	ou africada	sonora	judge

Tabela 2.1: Lista de unidades tipo fone para o inglês utilizadas nas experiências independentes do vocabulário. Para cada fone apresenta-se igualmente a representação fonológica mais próxima de cada fone, o modo correspondente de articulação, o vozeamento e uma palavra cuja realização acústica contém normalmente esse fone. Na coluna da esquerda não se assinalou o vozeamento porque todas as unidades são sonoras.

tom" (/en/) e "winner" (/nx/); "sing" (/ng/) e "Washington" (/eng/); "hay" (/hh/) e "Leheight" ou "ahead" (/hv/). Contudo, alguns destes fones são relativamente raros, não permitindo um treino adequado dos respectivos modelos. Assim, estes alofones do mesmo fonema foram considerados num único fone.

Consideram-se fones adicionais para as oclusivas surdas (/p/,/t/ e /k/) e para a oclusiva sonora /d/. Verificou-se que para estes fones a oclusão estava sempre presente mas que a explosão[9] nem sempre lhe sucedia. Para modelar as situações (transcrições) em que era previsível a ocorrência de explosões mais curtas ou a inexistência destas considerou-se o uso de quatro novos fones: /pd/, /td/, /kd/ e /dd/. Este tipo de modelamento é considerado uma forma implícita de considerar a inserção e a supressão de elementos subpalavra. Tal foi inicialmente proposto por Lee (Lee, 1989) em conjugação com uma topologia de modelos HMM apropriada. O uso destes fones nas experiências do presente trabalho, embora não implique melhorias significativas nas taxas de reconhecimento globais, determinou acréscimos substanciais destas taxas nas palavras em que ocorre. Tal facto determina a sua presença na tabela 2.1.

Na TIMIT é ainda descrito o fenómeno designado por "flapping", o qual transforma /t/ ou /d/ num fone que se poderia considerar intermédio (/dx/). Este fenómeno é tipicamente norte-americano e não foi considerado no inventário da tabela 2.1. Afecta palavras tais como "better","butter", etc.

A fricativa existente na palavra "measure" (/zh/) é relativamente pouco frequente e foi substituída pela fricativa mais próxima (/sh/) que existe em palavras como "shoe".

Embora os fonemas do inglês se encontrem bem definidos, existem muitos sons que são frequentemente utilizados e que são considerados não fonémicos. Tal é o caso da maioria dos pares oclusão-fricativa, que são na realidade muito diferentes do par de fonemas concatenados, assemelhando-se mais a uma espécie de africativas (por exemplo, no vocábulo "its"). No SPHINX considerou-se um fone para o par /t/-/s/ (/ts/) o qual foi também inicialmente utilizado no presente trabalho. Contudo, a existência deste fone não produziu qualquer acréscimo no desempenho do reconhecedor com o corpus estudado. Tal pode ser devido ao facto deste corpus ser de palavras isoladas e dos respectivos oradores serem europeus (secção 3.2).

[9] A explosão ("release", "burst" ou "explosion") sucede logo que a oclusão termina libertando a pressão de ar acumulada (cerca de 6cm H_2O devido a pressão continuada dos pulmões) numa fuga de ar que conduz a uma explosão acústica de cerca de 10ms (O'Shaughnessy, 1987).

2.5.3 Modelos dependentes do contexto

As unidades de fala, tal como foram descritas nos parágrafos anteriores, não permitem incorporar diferenças referentes ao contexto em que ocorrem. Ou seja, os efeitos de coarticulação com as unidades envolventes, antes e depois do segmento em análise não são contemplados por definição. Naturalmente, as consequências são tanto mais negativas quanto menor for a duração média destas unidades. No limite, se a unidade a modelar for a própria frase, o problema torna-se praticamente inexistente. Dados os inconvenientes anteriormente referidos para a utilização de unidades demasiado grandes, utilizam-se alternativas para as unidades mais pequenas, de modo a integrar a informação relevante do contexto em que se inserem. Surge desta forma uma nova variedade de unidades de fala designadas por *unidades dependentes do contexto*. Deste modo, as unidades referidas nos parágrafos anteriores passam a ser também designadas por *unidades independentes do contexto*.

Entre as unidades dependentes do contexto mais divulgadas, contam-se os *trifones* e os *difones*. Cada trifone classifica de forma diferenciada a ocorrência de um fone de acordo com o fone imediatamente anterior e do fone seguinte, também designados por *contexto à esquerda* e por *contexto à direita*, respectivamente. Ou seja, para um inventário de N_s fones poderiam existir N_s^3 trifones (97.336 para $N_s = 46$). Na realidade, numa dada língua, a maioria destes trifones não existem ou raramente ocorrem. O difone, por sua vez classifica a ocorrência de um fone de acordo com apenas um dos contextos, normalmente o referente ao fone anterior. Actualmente utilizam-se com sucesso modelos com dois ou mais fones à esquerda e à direita do fone a ser considerado. Estes modelos são designados por *polifones* (Finke e Rogina, 1997). Apesar de utilizarem a mesma designação, não devem ser confundidos com o conceito de polifone definido no âmbito da multilingualidade e descrito no capítulo 5.

Os inconvenientes decorrentes do uso destas unidades são essencialmente devidos ao seu número extremamente elevado. Assim, o treino dos modelos correspondentes necessita de um corpus de fala de dimensão elevada. O tempo de cálculo necessário para este treino e a memória necessária para o armazenamento computacionalmente eficaz destes modelos tornam-se, também, exageradamente elevados. Existem algumas estratégias para diminuir o número total destas unidades. Por exemplo, o conceito de *trifones generalizados*, baseado no agrupamento criterioso de trifones, determinou cerca de 500 unidades deste tipo em vez dos 2.381 trifones exigidos para uma tarefa em inglês com mil palavras (Lee, 1989). Para o problema dos dados insuficientes para um treino adequado dos modelos, existem várias soluções, entre as quais se destaca a interpolamento da supressão

("deleted interpolation") (Lee, 1989) e o conceito de ligação de parâmetros apresentado na subsecção 2.4.4.

Existem ainda outras formas de integrar contexto nas unidades subpalavra para o problema, tal como o uso de fones dependentes da palavra. Com esta solução atinge-se uma situação de fronteira com o reconhecimento baseado em palavras isoladas.

No decorrer deste trabalho fizeram-se algumas experiências com modelos de trifones. Dado o tamanho relativamente pequeno do vocabulário não foi possível identificar vantagens significativas com o uso destas unidades quando comparadas com o uso de modelos do tipo fone. A justificação deve-se ao facto do número de contextos por fones ser em média relativamente baixo. Assim, os modelos de fones com várias componentes gaussianas e com mais material de treino disponível permitem a obtenção de resultados dificilmente ultrapassáveis.

2.5.4 Anotação do sinal de fala

Para que os sinais acústicos da fala possam ser estudados e utilizados, é em geral indispensável associar-lhes alguma forma de transcrição. A transcrição consiste na atribuição ao sinal, ou a segmentos deste, de uma ou mais etiquetas. O processo de transcrição é também por vezes designado por *etiquetagem* ou por *anotação*. A transcrição pode ou não ser complementada com a descrição das durações dos referidos segmentos. Estes dados e o processo de os obter, são conhecidos pelo nome de segmentação. Cada etiqueta possui um significado pré-estabelecido e que se opõe ao significado das restantes etiquetas. Deste modo o conjunto de etiquetas utilizado constitui uma espécie de código para um determinado fim. Neste sentido mais amplo, as transcrições baseadas no sinal de fala são usadas em muitos campos reconhecidos da linguística, incluindo a fonética, a fonologia, a sociolinguística, a psicolinguística, a patologia da fala, a dialectologia e o ensino de línguas. São também utilizadas em disciplinas como as da psicologia, a antropologia e a sociologia (Gibbon et al., 1997). Naturalmente, o significado das etiquetas a atribuir em cada uma destas áreas pode ser muito diferente. Algumas componentes essenciais do presente trabalho podem ser descritas, de forma resumida, sob o prisma da implementação automática de procedimentos para a obtenção de transcrições:

- no caso do reconhecimento de fala pretende-se obter as transcrições ortográficas do que foi dito;
- no caso de se incluir a capacidade de rejeição de palavras existe uma etiqueta adicional que designa genericamente todas as palavras que não pertencem ao vocabulário

pré-definido (capítulo 4);

- no caso da identificação do sexo do orador pretende-se atribuir a determinada locução uma das duas etiquetas possíveis, referentes a cada sexo (capítulo 6). Em algumas circunstâncias esta decisão pode ser feita com base em múltiplas locuções do mesmo orador;

- o caso da identificação do sotaque é muito semelhante ao anterior, podendo haver apenas duas etiquetas para distinção do sotaque estrangeiro do sotaque nativo, ou uma etiqueta para cada grupo de oradores que partilhem a mesma língua materna (capítulo 6).

Estas tarefas de reconhecimento de fala e de identificação de dados do orador necessitam em geral de modelos associados a cada etiqueta. O treino destes modelos depende da existência de um corpus de sinais previamente transcrito com o mesmo código de etiquetas daquele que se pretende descodificar automaticamente. De igual modo, a verificação de resultados depende da existência de um corpus com características semelhantes ao de treino que permita testar os procedimentos de reconhecimento e de identificação. É necessário confrontar os resultados obtidos por via automática com os obtidos por via da designada *transcrição manual*: a obtida directamente a partir de um operador humano por inspecção auditiva e eventualmente com o apoio da inspecção visual.

As condições da recolha do corpus utilizado no presente trabalho permitiram que todos os sinais de fala fossem ortograficamente descritos, bem como as características relevantes dos respectivos oradores, isto é, a sua nacionalidade e género (secção 3.2). Estas transcrições são as necessárias e suficientes para as experiências com modelos de palavras inteiras, uma vez que se trata de um corpus de palavras isoladas. Contudo, havia a necessidade de aprofundar o estudo dos aspectos relativos ao sotaque estrangeiro e a de, simultaneamente, prever a utilização de reconhecedores da última geração para fala contínua. Consequentemente, era indispensável o treino de modelos de unidades subpalavra. No caso presente, tal como já foi explicado, as unidades escolhidas foram do tipo fone. Torna-se assim necessário dispor de uma forma de transcrição adequada que permita identificar quais os segmentos de sinal a utilizar no treino dos modelos de fones. De facto, para o treino destes modelos não é essencial dispor da segmentação do sinal. Os procedimentos de inicialização e de reestimação dos modelos HMM permitem realizar esta segmentação de forma implícita a partir de um léxico de pronúncia.

Distinguem-se os seguintes níveis de transcrição para um sinal de fala (Barry e Fourcin, 1992; Gibbon et al., 1997):

ortográfico A transcrição ortográfica é relativamente simples de obter a partir da audição simples do sinal de fala. Conforme se referiu esta transcrição foi gerada durante a recolha do corpus utilizado (secção 3.2);

morfo-sintáctico Este nível é particularmente interessante para aqueles que pretendem estudar a relação entre a prosódia e a sintaxe;

forma de citação ("citation phonemic") A forma de citação contém etiquetas que representam as unidades de som funcionalmente distintas de uma língua. A forma de citação de uma palavra é uma construção analítica, resultante da comparação de palavras pronunciadas isoladamente e de forma cuidada. Os sons distintivos que formam a pronúncia "ideal" de uma palavra podem não ser efectivamente pronunciados por um dado orador numa determinada circunstância. Contudo, este tipo de transcrição é importante no estabelecimento de relações entre o sinal e os restantes níveis linguísticos e como fonte de conhecimento para o desenvolvimento de regras fonológicas.

É possível obter transcrições deste tipo, a partir de transcrições ortográficas, utilizando um conversor automático de grafemas. Este processo é muito utilizado pelos sintetizadores automáticos de fala. A outra forma de obter formas de citações é a consulta de um léxico de pronúncia já existente. Para o reconhecimento de fala este tipo de transcrição é normalmente considerado demasiado geral, demasiado dependente dos contrastes linguísticos e não tanto dos aspectos acústicos passíveis de parametrização.

O IPA ("International Phonetic Association") é o sistema de transcrição mais utilizado para este fim. O sistema SAMPA é considerado a versão do IPA para as representações em computador uma vez que é baseado em caracteres ASCII mas foi essencialmente concebido para transcrever as principais línguas europeias.

Alguns autores referem também a forma de citação como *transcrição fonémica* (Barry e Fourcin, 1992);

fonético largo A *transcrição fonética larga* também é por vezes designada por *transcrição fonémica* ou ainda por *fonotípica*. Corresponde a diversos tipos de transcrição num nível intermédio entre o nível fonético estreito e a forma de citação. Emprega símbolos com significado fonémico mas utiliza-os para indicar fenómenos não fonémicos tais como os que ocorrem na fala contínua como os de assimilação, redução de vogais, supressão de consoantes, etc. O objectivo é o de capturar a sequência de sons efectivamente produzida em termos das categorias utilizadas para representar palavras no léxico. Reduz assim o nível de abstracção dos símbolos utilizados na

forma de citação. Assim, exige um nível de detalhe superior ao da forma de citação mas pode, em princípio, ser gerado a partir desta através de regras fonológicas. É o tipo de transcrição habitualmente utilizado na etiquetagem dos corpora de fala (capítulo 3). Um procedimento habitual é o de segmentar manualmente uma pequena parte de um corpus que é utilizado para treinar modelos HMM. Alternativamente podem utilizar-se as formas de citação para se criar uma segmentação uniforme, conforme se explicou na secção 2.3.5. Estes modelos permitem posteriormente obter uma segmentação da totalidade do corpus, a qual poderá ainda ser corrigida manualmente. Existem programas ("Waves", marca registada da Entropics) que facilitam consideravelmente esta tarefa, fornecendo alguns deles aproximações automáticas adicionais (Zue et al., 1989; Andersen e Dalsgaard, 1992).

A utilização de unidades tipo-fone é geralmente associada a este nível de transcrição. Os modelos de transcrição descritos no capítulo 5 permitem obter descrições que podem ser consideradas deste tipo;

fonético estreito A designação de *transcrição fonética estreita* é também por vezes usada para a transcrição alofónica. A transcrição alofónica reflecte o modo como o material de fala revela diferenças relacionadas com o contexto estrutural e ambiental. A transcrição fonética estreita é a mais específica em termos linguísticos, uma vez que pretende ser o mais representativa possível de uma realização individual contida num determinado sinal de fala. É a única que não dispensa a audição do sinal e a visualização da forma de onda e do espectrograma. Inclui marcas ou sinais diacríticos, tais como os de labialização, nasalação, articulação síncrona, etc. para indicar a forma exacta como foi pronunciado. Este tipo de transcrição deverá ser ideal para o treino de modelos HMM mas exige muito tempo de trabalho por parte de técnicos especializados. A transcrição fonética larga representa um compromisso aceitável para o treino de reconhecedores automáticos.

acústico-fonético Este nível distingue toda a porção de fala que é reconhecível como um segmento separado, quer na forma de onda acústica, quer no espectrograma: a oclusão, a explosão, a aspiração, as partes sonoras e surdas de uma fricativa, nasal ou oclusiva, etc. O reconhecimento destas categorias articulatórias permite estabelecer relações com outros níveis linguísticos mais elaborados. As fronteiras de alguns segmentos são, até certo ponto, arbitrárias sendo necessário estabelecer critérios de homogeneidade;

físico É o nível de maior detalhe e pode-se basear, para além do registo acústico, noutras tipos de registos, tais como em detectores da transmissão nasal e no registo palatográfico ("palatography"). Obtêm-se desta forma várias camadas de descrições uma

vez que estes eventos físicos podem sobrepor-se no tempo. Contudo, os parâmetros acústicos são os mais frequentemente utilizados sendo escolhidos de acordo com aplicações específicas: energias resultantes de bancos de filtros, frequências das formantes, coeficientes autorregressivos e cepstrais, frequência fundamental, etc.;

fenómenos não linguísticos Inclui ruídos do orador tal como tosse, risos, estalidos ou cliques no descolar dos lábios, ("lip smacking") bem como ruídos do meio onde decorre a recolha do sinal. Este nível pode incluir igualmente informação paralinguística tal como a referente ao preenchimento de pausas, algumas interjeições, etc.

A transcrição prosódica define aspectos de uma locução que vão além das fronteiras de um único segmento. Por exemplo: a acentuação relativa de uma sílaba, o padrão de entoação de uma ou mais sílabas, o ritmo de uma frase ou parte de frase devido à distribuição das pausas e da acentuação. Uma vez que pode ser representada na maior parte dos níveis de transcrição cima descritos, como uma espécie de camada adicional, não se considera este tipo de transcrição como um único nível de transcrição (Barry e Fourcin, 1992). Os padrões de entoação são elementos importantes no estudo da fala dos oradores estrangeiros (Arslan, 1996). Contudo, estes padrões não são habitualmente utilizados, pelo menos de forma explícita, no reconhecimento automático de fala.

2.5.5 Obtenção do léxico de pronúncia

As transcrições fixas determinadas para o presente trabalho, podem ser consideradas do tipo fonético largo e encontram-se no apêndice A. Embora relativamente pequeno, não foi possível construir este léxico com base num único léxico de pronúncia existente. Tal facto deve-se em parte à existência de algumas palavras menos comuns, mas também às taxas de reconhecimento extremamente baixas obtidas para algumas das palavras. Este último aspecto obrigou à procura de transcrições mais adequadas ao corpus estudado. A maioria dos léxicos de pronúncia para a língua inglesa em suporte informático são de origem americana, enquanto que os testes de referência foram realizados com um corpus europeu. Tal poderá também justificar algumas diferenças em termos de transcrições. Assim, o léxico de pronúncia utilizado baseou-se essencialmente, nos seguintes elementos:

- léxico de pronúncia que acompanha o corpus de fala TIMIT;
- transcrições fonéticas estreitas associadas a cada locução no corpus TIMIT;

- léxico de pronúncia do sistema SPHINX para a tarefa "Resource Management Task" (Lee, 1989);
- repetidas audições de parte das palavras proferidas pelos oradores ingleses no corpus SUNSTAR (secção 3.2).

Para as palavras menos comuns e que apresentaram taxas de reconhecimento mais baixas, consultaram-se ainda:

- léxico de pronúncia de Carnegie Mellon (1993) — da Universidade com o mesmo nome (CMU) com transcrições alternativas[10];
- léxicos de pronúncia do "Wall Street Journal" (WSJ) de LIMSI. Inclui um subconjunto com mais de 20.000 palavras com pronúncias múltiplas;
- léxico de pronúncia "Moby Pronunciator"[11];
- resultados do conversor automático de texto em inglês para fonemas de John A. Wasser (1985[12]);
- dicionários convencionais que incluem chaves de pronúncia (Patterson e Litt, 1996).

De referir ainda que o léxico de CMU é gerado a partir das seguintes fontes independentes:

- léxico de inglês de mais de 20.000 palavras construído manualmente na CMU e exaustivamente testado;
- léxico de mais de 200.000 palavras da UCLA — verificado;
- subconjunto de 32.000 palavras do dicionário da Dragon Sytems;
- dicionário de nomes próprios com mais de 53.000 palavras gerado a partir de um sintetizador — não verificado;
- léxico de 200.000 palavras gerado pelo programa Orator — não verificado;
- léxico de 200.000 palavras gerado pelo programa Mitalk — não verificado.

Contudo, as entradas que ocorrem exclusivamente em fontes com direitos de autor não foram incluídas neste léxico. Tal é o caso do léxico da Dragon.

[10] "URL: http://www.speech.cs.cmu.edu/cgi-bin/cmudict".
[11] "URL: http://www.dcs.shef.ac.uk/research/ilash/Moby/".
[12] "URL: http://www.tiac.net/users/wasser/speak/".

2.5.6 Adaptação de alguns algoritmos

Os algoritmos referidos na subsecção 2.3.4 necessitam de algumas adaptações por forma a permitirem o uso de unidades subpalavra, léxico de pronúncia, modelo linguístico, (secção 2.6) etc. O Prof. Steve Young e os seus colegas implementaram algumas destas adaptações a partir de dois conceitos, um para o treino de modelos, outro para o teste, respectivamente, o *treino embutido* e o modelo de *token passing* (Young et al., 1991; Young et al., 1996).

Reestimação embutida

Para o treino de cada modelo de palavra dispõe-se em geral de múltiplas locuções devidamente segmentadas, contendo, cada uma delas, exclusivamente, o sinal que deve afectar o treino desse modelo. Deste modo o procedimento de treino pode ser totalmente independente para cada modelo de palavra. O conceito de *treino embutido* ("embedded training") é particularmente importante quando se pretende utilizar modelos subpalavra. Conforme se descreveu na subsecção 2.5.4, para o treino dos modelos subpalavra, dispõe-se usualmente de uma transcrição mas não de uma segmentação ao nível das unidades subpalavra. O conceito de treino embutido permite realizar essa segmentação em cada locução de forma implícita e tendo em consideração, em simultâneo, todas as outras realizações de cada unidade nas restantes locuções. Para cada locução é construído um modelo em que os modelos das unidades subpalavra são concatenados de acordo com a respectiva transcrição. Na perspectiva da ligação de parâmetros (subsecção 2.4.4) pode-se afirmar que se pretende ligar os parâmetros dos modelos referentes à mesma unidade subpalavra que se encontram embutidos nos modelos de cada locução. Assim, o conceito de treino embutido é utilizado no algoritmo de reestimação de modo a utilizar todos os dados de treino e a afectar os parâmetros de todos os modelos numa única iteração (Young et al., 1996).

Adaptação do algoritmo de Viterbi

O algoritmo designado por *token passing*[13] foi inicialmente utilizado num reconhecedor de palavras ligadas (VODIS) em que o reconhecimento propriamente dito era efectuado através de um algoritmo do tipo DTW (Young et al., 1991). A formulação deste algoritmo permitiu incorporar as restrições de uma gramática de contexto livre (secção 2.6) directa-

[13] O significado de "token passing" é o de passagem ou transmissão de uma determinada estrutura de informação.

mente no reconhecedor de palavras isoladas, processando redes de palavras alternativas em lugar de uma única sequência óptima. O algoritmo foi posteriormente utilizado com modelos HMM subpalavra num sistema que se revelou, principalmente nos meios académicos, como um padrão de referência no reconhecimento automático de fala: o HTK ("Hidden Markov Model Toolkit") (Young et al., 1996). Este mesmo reconhecedor foi usado na obtenção de muitos dos resultados apresentados na presente dissertação.

A descrição que se segue, tal como no sistema VODIS, refere a palavra como a unidade elementar da fala utilizada. Contudo esta unidade pode ser substituída por qualquer tipo de unidade subpalavra, sem necessidade de outra alteração significativa nesta descrição.

O algoritmo *token passing* pode ser apresentado como uma formulação alternativa do algoritmo de Viterbi que explicita o trajecto de alinhamento dos estados. A designação de *token* pode ser traduzida como um conjunto de informação organizada numa estrutura. No caso presente representa um acerto parcial da sequência de observações até ao instante t sob a restrição de se considerar o estado j nesse instante.

Para cada estado j de um HMM existe no instante t um único *token* que contém, entre outras informações, o logaritmo da probabilidade parcial $\delta_t(j)$. Assim, a expressão da recursão de Viterbi (equação 2.19) é substituída pelo algoritmo equivalente de *token passing* que é executado para cada trama (cada instante t). Os passos essenciais do algoritmo são os seguintes:

1. Todos os *tokens* no estado i são copiados para todos os estados j que sucedem a i incrementando o logaritmo da probabilidade do valor de $\log a_{ij} + \log b_j(o_t)$.

2. Examinar os *tokens* existentes em cada estado e eliminar todos menos o que apresentar a maior probabilidade para $\log \delta_t(j)$.

O trajecto de cada *token*, pode ser registado eficientemente se este incluir, para além do valor de $\log \delta_t(j)$, uma referência ao HMM anterior que regista sempre que atingir um estado de saída. Como os *tokens* anteriores podem ter sido eliminados no passo 2 do algoritmo acima descrito, esta informação tem de ser registada noutra estrutura. Cada vez que um *token* é copiado de um estado de saída para um de entrada e uma vez que esta transição corresponde potencialmente a uma transição de modelo, regista-se o valor na referência do *token* conjuntamente com a identidade do modelo de origem. O valor na referência do *token* é depois substituído por uma referência para este novo registo.

A referência contida no melhor *token* permite posteriormente identificar recursivamente todos os registos que lhe dizem respeito, ou seja, este procedimento permite obter todas

as fronteiras entre palavras.

2.6 Modelo linguístico

Na introdução do presente capítulo apresentou-se $Pr(W)$ na equação 2.1, como a probabilidade a priori de observar a palavra ou sequência de palavras $W = \{w_1, w_2, \ldots, w_Q\}$ independentemente da sua realização acústica e que deve ser determinada a partir de um modelo linguístico[14]. Embora a definição de unidades subpalavra e do léxico também envolvam conceitos da linguística, os modelos aqui referidos de linguísticos referem-se aos níveis de construção da língua imediatamente superiores a estes: a sintaxe, a semântica e a pragmática. Nas aplicações de reconhecimento automático de fala nem sempre é possível separar o uso destes diferentes níveis de conhecimento, em particular nas aplicações menos complexas.

Os modelos linguísticos podem ser utilizados para discriminar diferentes línguas, nomeadamente aquelas que derivam de famílias diferentes apresentando estruturas sintácticas diferenciadas. Contudo, este tipo de modelos não são, por si só, de utilidade evidente para o problema da identificação do sotaque. No presente trabalho, a abordagem deste problema centra-se nas diferenças aos níveis acústico e fonético. No entanto, as alterações na língua (subsecção 3.1.1) introduzidas pelo orador estrangeiro podem ser extensíveis a outros níveis, nomeadamente ao nível sintáctico.

No reconhecimento de fala procuram-se formalismos relativamente simples para o modelo linguístico, uma vez que este deve ser integrado nos algoritmos de descodificação. Pelo contrário, no processamento da língua natural são utilizados modelos mais complexos. Nomeadamente, referem-se os modelos baseados nas gramáticas formais (Ney, 1990), tais como as designadas por *dependentes do contexto* e as *livres do contexto* ("context free grammars") (Jelinek et al., 1990).

Um dos modelos de maior sucesso para a integração com o reconhecimento automático de fala é conhecido pela designação de *N-gramas* ("N-grams"). Estes modelos baseiam-se no cálculo aproximado da probabilidade

$$Pr(W) = Pr(w_1)Pr(w_2|w_1)Pr(w_3|w_1w_2)\cdots Pr(w_Q|w_1w_2\cdots w_{Q-1}). \tag{2.34}$$

De facto, é impossível calcular de forma segura $Pr(W)$ para todas as palavras e valores Q numa determinada língua. O mesmo acontece com as probabilidades condicionais

[14] Tem sido igualmente utilizada a designação de *modelo de linguagem* na tradução da expressão inglesa "language model" (Martins, 1998a).

$Pr(w_j|w_1 \cdots w_{j-1})$ da expressão 2.34. Nos modelos *N-gramas*, estas probabilidades são aproximadas por forma a basearem-se exclusivamente nas $N-1$ palavras anteriores a w_j:

$$Pr(w_j|w_1 \cdots w_{j-1}) \approx Pr(w_j|w_{j-N+1} \cdots w_{j-1}).$$

Os exemplos mais conhecidos são os bigramas (N=2) e os trigramas (N=3) (Jelinek, 1985), mas este tipo de formulação abrange ainda outros modelos mais simples, utilizados em muitos dos reconhecedores actuais. Tal é o caso dos modelos designados por:

par de palavras ("word pair"). Atribui valores binários às probabilidades condicionais com N=2. Este modelo é particularmente útil quando não é possível estimar convenientemente bigramas. Deste modo atribui-se $Pr(w_j|w_{j-1})$=1 sempre que se conhece numa determinada língua a sequência de palavras $w_{j-1}w_j$ e zero às restantes probabilidades condicionais;

unigramas (N=1). Baseia-se exclusivamente na frequência relativa de cada palavra numa dada língua e é geralmente menos eficaz do que o anterior;

sem gramática ou de *gramática nula* com N=2 mas com $Pr(w_j|w_{j-1})$=1 qualquer que sejam as palavras w_j e w_{j-1}. Adopta-se este tipo de modelo quando não é possível ou necessário utilizar um modelo linguístico.

Embora menos frequentes, existem trabalhos na área do reconhecimento em que se utilizou valores de N superiores a três (Hetherington, 1995).

Uma outra alternativa aos modelos probabilísticos do tipo *N-gramas* são os modelos baseados em *gramáticas com número finito de estados* ("finite-state grammars"). De acordo com a hierarquia das gramáticas de Chomsky estas apresentam um poder de modelação imediatamente inferior às já referidas *gramáticas livres do contexto* (Miller e Levinson, 1988). Este tipo de modelos linguísticos são simples de obter para aplicações de pouca complexidade, como a que foi experimentada na secção 4.9.

Em termos de processamento de fala torna-se, em geral, difícil a comparação de resultados, nomeadamente os obtidos a partir de diferentes investigadores. Um dado importante para a comparação consiste na avaliação da complexidade da tarefa que determinado estudo pretende resolver.

No reconhecimento de fala automático é habitual considerarem-se domínios restritos ou subconjuntos de uma língua, que podem ser descritos como limitações em termos dos modelos linguísticos. Quanto menores forem estas limitações maior será a complexidade

da tarefa de reconhecimento de fala. É possível determinar diversas medidas objectivas da qualidade de um modelo linguístico (Jelinek et al., 1991) ou da complexidade associada à escolha de determinado subconjunto L de uma língua. Uma medida utilizada para este fim é a designada por *perplexidade* e é dada por $2^{H(L)}$, em que $H(L)$ é a entropia associada a L. Se se pensar em L como uma fonte de informação gerando na saída uma sequência de k palavras $p^k = \{p_1, p_2, \ldots, p_k\}$, então $H(L)$ representa a quantidade de informação média por palavra e é dada por

$$H(L) = -\sum_{p^k} \frac{1}{k} Pr(p^k) \log Pr(p^k).$$

A perplexidade é uma medida aproximada do factor de ramificação médio ("average branching factor"). Num modelo linguístico implementado com uma máquina probabilística com número finito de estados, este factor é facilmente calculado como a divisão do número de ramos ou transições que emergem dos estados, pelo número total de estados excluindo os estados finais (Huang et al., 1990).

Com excepção das experiências descritas na secção 4.9, considera-se no presente trabalho o modelo linguístico mais simples, no qual estas sequências contêm apenas uma única palavra de um vocabulário de P palavras. Além disso, considera-se igual probabilidade de ocorrência para todas as palavras ($= \frac{1}{P}$). Deste modo, obtém-se $H(L) = \log P$, sendo a perplexidade igual à dimensão do vocabulário. Por exemplo, a perplexidade de uma tarefa de reconhecimento de dígitos é igual a dez.

2.7 Avaliação de resultados

A avaliação dos resultados obtidos com reconhecedores automáticos de fala baseia-se na determinação de alguns valores numéricos que permitem verificar dois aspectos diferentes: a eficácia do reconhecimento e a comparação entre reconhecedores. No primeiro aspecto, avaliam-se essencialmente aplicações do reconhecimento, por vezes já em testes de campo. Procuram-se melhorar detalhes da interacção com o utilizador ou obter especificações ou certificações da aplicação. A comparação entre reconhecedores é particularmente importante nas áreas de investigação e desenvolvimento onde são ensaiados novos algoritmos.

A medida mais comum para avaliar um reconhecedor de palavras isoladas é dada pela taxa de reconhecimento:

$$R_c = 100\% \frac{\text{\# palavras correctamente reconhecidas}}{\text{\# total de palavras proferidas}}.$$

Uma medida semelhante é a designada taxa de substituição:

$$S = 100\% \frac{\#\text{ palavras do vocabulário que foram substituídas por outras}}{\#\text{ total de palavras proferidas}}.$$

Numa tarefa simples de reconhecimento obtém-se $S = 100\% - R_c$. Pode ser ainda útil contabilizar as palavras que eventualmente se perdem no detector de início e fim de palavra, ou que sejam rejeitadas pelo próprio reconhecedor, no caso de este dispor de um mecanismo que lhe permita identificar quais são as palavras que pertencem ao respectivo vocabulário (capítulo 4). Neste caso, determina-se uma taxa de supressão ("deletion rate"):

$$D = 100\% \frac{\#\text{ palavras do vocabulário que não produziram qualquer resultado no reconhecedor}}{\#\text{ total de palavras proferidas}}.$$

O reconhecedor pode ainda emitir um resultado de reconhecimento quando não o deveria de fazer, situação por vezes designada por falso alarme. Tal pode ser devido a um ruído, ao uso de palavras que não pertencem ao vocabulário, etc. Desta forma, é ainda possível determinar uma taxa de inserção I:

$$I = 100\% \frac{\#\text{ resultados do reconhecedor que não correspondem à locução de nenhuma palavra do vocabulário}}{\#\text{ total de palavras proferidas}}.$$

A taxa de erro E associa as medidas de insucesso anteriormente referidas: $E = S + D + I$. Define-se ainda a *taxa de exactidão* ("word accuracy") com o valor de $100\% - E$.

Este tipo de taxas é habitualmente considerado, no reconhecimento de palavras isoladas, em relação a cada uma das palavras do vocabulário a reconhecer. Este tipo de análise de resultados é normalmente efectuado a partir da designada *matriz de confusão*. Trata-se de uma matriz quadrada, de dimensão P igual à do vocabulário a reconhecer. Cada elemento $p_{i,j}$ desta matriz representa a probabilidade de uma locução de uma palavra i ser reconhecida como sendo a palavra j. O conceito de matriz de confusão é frequentemente utilizado na classificação supervisionada de qualquer tipo de dados. Neste âmbito, o valor de P corresponde ao número de classes a considerar e cada elemento $p_{i,j}$ representa a probabilidade de um padrão gerado pela classe i ser classificado na classe j. Assim, também se determinaram estimativas das matrizes de confusão no âmbito da identificação do sexo e da identificação do sotaque (capítulo 6).

A determinação das matrizes de confusão revelou-se particularmente importante no reconhecimento de palavras isoladas com unidades subpalavra realizado no presente estudo. De facto, com base nas palavras com piores taxas de reconhecimento e na respectiva substituição por outras palavras de alguma forma semelhantes, foram introduzidos alguns

dos melhoramentos no léxico de pronúncia. A análise da matriz de confusão permitiu igualmente detectar alguns dos problemas relacionados com os sotaques de cada orador ou grupo de oradores.

A matriz de confusão $\{p_{ij}\}$ é uma *matriz estocástica* uma vez que verifica as condições:

$$p_{ij} \geq 0 \qquad \sum_{j=1}^{P} p_{ij} = 1$$
$$1 \leq i, j \leq P.$$

Deste modo, verifica também as propriedades de uma matriz de Markov (equações 2.14).

Num classificador ideal, em que não existam erros de classificação, a matriz de confusão é igual à matriz identidade. Na maioria dos casos, não é possível o cálculo analítico da matriz de confusão, recorrendo-se por isso aos resultados experimentais obtidos a partir de um conjunto de padrões de teste. Neste caso, contabiliza-se o número de vezes n_{ij} em que ocorre a saída j do classificador para um padrão i. Normalizando com o número n_i de padrões de teste da classe i, obtêm-se uma estimativa de p_{ij} com base na frequência relativa

$$\hat{p}_{ij} = \frac{n_{ij}}{n_i}.$$

Se n_i for considerado conhecido, $\hat{p}_{ij}$ tem uma distribuição binomial com média $E[\hat{p}_{ij}] = p_{ij}$ e variância dada por

$$E[(\hat{p}_{ij} - p_{ij})^2] = \frac{1}{n_{ij}} p_{ij}(1 - p_{ij}).$$

Assim, o estimador é centrado e a respectiva variância tende para zero quando n_i tende para infinito (com $1/n_i$) (Marques, 1994). Deste modo, compreende-se que o valor desta estimativa, ou o seu significado estatístico esteja dependente do valor de n_i. O mesmo acontece para as taxas anteriormente definidas em relação aos números que apresentam no denominador.

Considere-se, por exemplo, a taxa de reconhecimento R_c. É possível assumir que a distribuição do erro na determinação do respectivo numerador seja binomial (Chollet e Montacie, 1987). Assim, considera-se n como sendo o número de experiências de Bernoulli, ou seja, o número de padrões no conjunto de teste e h a probabilidade de uma classificação individual estar errada. Para valores elevados de n o teorema de DeMoivre-Laplace garante uma aproximação da distribuição binomial por uma distribuição normal para R_c, $\mathcal{N}(\mu, \sigma)$. Deste modo, é possível determinar o designado *intervalo de confiança* para a estimativa R_c, $L_1 \leq R_c \leq L_2$ em que

$$L_{1,2} = \frac{hn + \frac{z_x^2}{2} \pm z_x \sqrt{h(1-h)n + \frac{z_x^2}{4}}}{n + z_x^2}, \tag{2.35}$$

e em que x é o designado *nível de confiança*. O seu valor indica qual a probabilidade de o intervalo aleatório

$$[E\{R_c\} - \frac{z_x\sigma}{\sqrt{n}}, E\{R_c\} + \frac{z_x\sigma}{\sqrt{n}}], \tag{2.36}$$

conter o valor de R_c (Meyer, 1965). Para um nível de confiança a 90%, a 95% e a 99% vem, respectivamente $z_{90\%} = 1,645$, $z_{95\%} = 1,960$ e $z_{99\%} = 2,576$ (Kreyszig, 1970).

No reconhecimento de palavras isoladas, assume-se a existência de um mecanismo capaz de segmentar explicitamente no sinal de fala o padrão que será comparado com os modelos de palavra do vocabulário activo: um detector de início e fim de palavra (subsecção 2.2.3). No reconhecimento de fala contínua não existe um mecanismo equivalente. Esta tarefa é agora dificultada pelo facto de existirem uma sucessão de unidades subpalavra encadeadas no tempo que, devido a inserções e supressões, não permitem estabelecer uma relação directa com as unidades de alguma das transcrições previstas no léxico de pronúncia. O mesmo se passa ao nível sintáctico e, nomeadamente, com o reconhecimento de fala ligada (secção 4.9) quando se pretende comparar uma sucessão de palavras com algumas das frases previstas para a aplicação. Este problema não se coloca se apenas se pretender determinar qual a frase mais provável. Contudo, na avaliação de resultados pode ser importante obter, para além de uma taxa de frases correctas, a taxa de reconhecimento em termos de palavras correctas ou, no caso do reconhecimento com unidades subpalavra, a taxa de unidades deste tipo correctamente identificadas. Neste caso é necessário efectuar um emparelhamento óptimo ("optimal string match") das sequências de referência com as sequências de teste, através de um procedimento de programação dinâmica (Tavares e Correia, 1986). O procedimento utilizado no presente trabalho (programa NIST) associa um valor de custo para cada emparelhamento em relação à sequência de referência (Young et al., 1996): duas etiquetas (palavras ou unidades subpalavra) iguais adicionam zero unidades nesse valor; a inserção ou a supressão de uma etiqueta adiciona-lhe três unidades; a substituição de uma etiqueta por outra adiciona-lhe quatro unidades. A penalização da substituição é muito superior à da inserção e à da supressão, uma vez que pode ter consequências mais graves em termos da aplicação do reconhecimento. A sequência que apresentar um valor mais baixo para o emparelhamento é considerada a sequência óptima.

2.8 Conclusões

Neste capítulo descreveram-se os fundamentos dos modelos e dos métodos utilizados no âmbito da presente dissertação para o desenvolvimento de reconhecedores au-

tomáticos de fala. Analisou-se a construção do conjunto de unidades subpalavra e do respectivo léxico de pronúncia usado no reconhecimento independente do vocabulário. Descreveram-se alguns modelos linguísticos que serão referidos nos capítulos seguintes. Por fim, apresentaram-se parâmetros para a avaliação dos resultados do reconhecimento.

Capítulo 3

Corpus de fala multissotaque

3.1 Introdução

O problema do reconhecimento automático de fala independente do orador assume aspectos particulares quando a língua a ser reconhecida é utilizada por populações numerosas e heterogéneas. Tal como as próprias línguas, estas heterogeneidades têm em geral uma identificação clara em termos geográficos. Assim, as fronteiras geográficas e políticas delimitam também zonas de hábitos linguísticos diferentes.

No presente capítulo descreve-se o corpus de sinais de fala utilizado nas experiências efectuados nos capítulos seguintes. Este corpus inclui alguns sotaques dominantes na União Europeia quando se pronuncia um conjunto determinado de palavras em inglês.

3.1.1 Variação linguística

No texto seguinte designam-se por *oradores nativos* aqueles que em determinada circunstância falam a sua língua materna ou primeira língua, também designada por L1 nos estudos interlíngua. É no universo dos oradores nativos que surgem os designados *dialectos*. A partir da designada linguística diacrónica, são conhecidas algumas propostas de árvores de classificação para as línguas mais representativas, as quais são organizadas em famílias e subfamílias de línguas mais antigas (Akmajian et al., 1990a). Os dialectos associados a uma língua multiplicam-se devido a alguns factores semelhantes aos que determinaram a génese de línguas pertencentes à mesma família: factores geográficos, tais como a distância entre regiões habitáveis e factores políticos que determinam a adopção de línguas oficiais de outros países. O rigor climático também é um factor de diversi-

dade, nomeadamente nas regiões mais frias. Os hábitos sedentários em virtude destas barreiras favorecem a não propagação de inovações idiomáticas que se acumulam assim de forma localizada. Parafraseando Ferdinand de Saussure: "há tantos dialectos como lugares" (Saussure, 1916). As diferenças em relação à língua original são ainda, muitas vezes, orgulhosamente salientadas pelos representantes de alguns dialectos. Este aspecto enquadra-se num conflito mais geral, designado por Saussure: entre a comunicação e o "espírito da capelinha".

As línguas sofrem ainda variações impostas por outros utilizadores que não são os seus oradores nativos. Por vários motivos, grupos de pessoas falando diversas línguas têm, em muitas circunstâncias, necessidade de contactos sociais. Quando isto acontece, torna-se por sua vez necessário encontrar uma linguagem comum que sirva como meio de comunicação.

Historicamente, existe um tipo de situação na qual as pessoas estabelecem contacto, sem partilharem uma língua comum: nomeadamente, quando um grupo se torna politicamente e economicamente dominante sobre outro. Este é o caso típico dos territórios coloniais, onde ocorrem diversas facetas da designada *aculturação* (Bűhler, 1962) predominantemente no sentido do grupo mais forte para o mais fraco. Nestas circunstâncias, surgem as designadas línguas *pidgin*, baseadas em características linguísticas de uma ou mais línguas. Um *pidgin* é uma língua simplificada com um vocabulário e uma gramática reduzida, usada como meio de comunicação entre pessoas com diferentes línguas nativas. Não existem oradores nativos para o *pidgin*. Conhecem-se *pidgins* baseados no inglês, francês, holandês, espanhol, português, árabe e swahili, entre outros (Akmajian et al., 1990b). As línguas *pidgin* são por vezes designadas por línguas de contacto (o que reflecte o facto de estas línguas surgirem habitualmente quando há contacto entre determinados grupos sociais) ou de línguas marginais (sublinhando as reduções no vocabulário e na gramática).

O próprio termo *pidgin* diz-se ser derivado da palavra inglesa "business" para o correspondente *pidgin* chinês. O vocabulário reduzido das línguas *pidgin* é geralmente derivado da língua "dominante". As respectivas gramáticas carecem habitualmente de morfemas inflexionais (os substantivos não têm terminações para indicar o plural e os verbos não têm terminações para as conjugações de tempos e de pessoas). As formas do verbo "ser" são em geral, inteiramente inexistentes nos *pidgins* e as preposições são também limitadas a um conjunto reduzido e com funções múltiplas.

O *pidgin* representa um código extremamente flexível que procura referências em duas ou mais línguas e que pressupõe não haver fluência ou conhecimento profundo de nenhuma

delas que seja comum a todos os interlocutores.

Noutras circunstâncias, quando existe uma determinada língua (não necessariamente a língua nativa de algum dos presentes) conhecida por todos os participantes, decide-se por acordo a sua utilização. A linguagem usada deste modo é designada por *língua franca*. Este termo deriva de uma linguagem comercial usada nos portos Mediterrânicos nos tempos medievais, consistindo essencialmente no italiano com elementos do francês, espanhol, grego e árabe. Até meados do século dezoito, as escolas europeias usavam o latim como língua franca - uma língua comum para os documentos científicos e outros assuntos escolares (Akmajian et al., 1990b). No mundo contemporâneo, o inglês serve de língua franca em numerosas publicações e encontros sociais, científicos e políticos, onde existe a necessidade de utilizar uma língua comum.

A intensificação das relações internacionais reflecte-se actualmente na massificação do uso de uma segunda ou terceira língua, em geral uma língua franca. A excepção natural a esta tendência é representada pelos oradores cuja primeira língua é uma língua franca. Num sentido lato, as línguas faladas por oradores não nativos, também aqui designados por oradores estrangeiros, recebem a designação de segunda língua (L2) ou de língua estrangeira sem distinção entre uma e a outra[1].

A fala produzida pela maioria dos oradores estrangeiros apresenta características que permitem a um ouvinte nativo, distingui-los de um orador nativo. Estas características distintas configuram-se no que é habitualmente designado por sotaque estrangeiro, ou apenas por sotaque. Os ouvintes mais treinados, conseguem mesmo distinguir, de forma aproximada, qual é a primeira língua de alguns oradores estrangeiros. Surgem assim designações de um sotaque alemão, espanhol, italiano, etc. Em geral, é possível distinguir os oradores estrangeiros, com excepção de alguns particularmente dotados para línguas estrangeiras, nomeadamente com capacidades excepcionais de imitação (Markham e Nagano-Madsen, 1996) ou uma experiência intensiva e duradoura com uma língua estrangeira. Estes últimos chegam, por vezes, a uma situação em que é difícil decidir qual é a primeira e qual é a segunda língua, uma vez que o contacto com a segunda língua afectou de forma significativa o desempenho com a primeira língua. Nos oradores considerados bilíngues deverá ser difícil descriminar qual das duas línguas é mais ou menos afectada pela outra (Flege, 1998). Em geral, considera-se que existe uma transposição de características entre as línguas que um dado orador é capaz de falar. Parte da presente dissertação foi desenvolvida com base neste pressuposto.

[1] "As noções de Língua Estrangeira (LE) e de Segunda Língua (L2) opõem-se, uma e outra, mas de maneira diferentes , à de Língua Materna (LM) ou Primeira Língua (L1)" (Frias, 1992).

O inglês, na qualidade da língua franca com maior disseminação na actualidade, é a língua estrangeira que será objecto de estudo neste trabalho. O termo de comparação ou referência será feito com o inglês europeu ou britânico padrão, ou seja, o mais próximo do que é leccionado nos institutos britânicos e outras escolas internacionais na Europa.

3.1.2 Motivações para a recolha de corpora de sinais de fala

A fala tem sido registada ao longo da história humana sob a forma escrita. Existem actualmente disponíveis numerosos corpora de texto digitalizado. Eliminando-se os aspectos referentes à caligrafia, estes corpora permitem o estudo sistemático das línguas ao nível linguístico, semântico ou pragmático. Estes estudos são importantes para o desenvolvimento dos reconhecedores de fala do futuro.

Contudo, a informação envolvida no processo da fala oral não se confina ao conteúdo textual da mensagem que lhe está subjacente. As características específicas do orador e inúmeras condicionantes circunstanciais são dados essenciais a ter em consideração no estudo do sinal de fala e do seu reconhecimento automático. Para o seu estudo, os sinais acústicos da fala têm sido registados sistematicamente em suportes analógicos e, actualmente, quase exclusivamente em suportes digitais. Estes registos, quando anotados e coleccionados em quantidades significativas e acompanhados de documentação suficiente que permita o seu uso para fins científicos, são designados por corpora de fala.

Existem actualmente disponíveis corpora de sinais acústicos de fala que cobrem grande variedade de situações, por vezes incluindo alguns resultados de pós-processamento manual ou automático de grande utilidade, como é o caso das transcrições ortográficas e fonéticas.

A recolha de um corpus representativo de todos os sinais de fala conhecidos, mesmo que referente a uma só língua ou outra limitação de carácter geral, é uma tarefa impossível uma vez que necessita de infinitos recursos tanto na aquisição como no respectivo processamento. Daí que a maioria dos estudos realizados procedam a uma recolha controlada de corpora de fala que ilustrem isoladamente os aspectos específicos a serem estudados.

O conteúdo de um corpus é habitualmente especificado em termos do ambiente acústico e emocional em que é efectuada a recolha, do equipamento a utilizar, do conteúdo do material linguístico a ser recolhido, da respectiva anotação e da escolha criteriosa dos oradores. Em termos destes últimos, consideram-se aspectos da variabilidade interorador tais como o sexo, a idade, o peso, a educação, o contacto com outras culturas, o hábito de fumar, etc. Estes dados são anotados e utilizados na selecção dos oradores.

As especificações da recolha de um corpus de fala são pré-definidas por forma a serem atingidos determinados objectivos. Assim, para uma dada aplicação de reconhecimento de fala, é comum proceder-se à recolha de um corpus no ambiente típico de utilização (por exemplo: na cabina de um automóvel). Nos estudos mais académicos, prefere-se a realização da recolha numa câmara surda ou mesmo anecóica, eliminado-se desta forma os efeitos de ruído e reverberação. Se, por outro lado, se pretende eliminar os efeitos da coarticulação interpalavra, recolhem-se palavras isoladas com a vantagem do início e o fim das palavras poderem ser determinados automaticamente com maior precisão.

Os reconhecedores do futuro deverão ser capazes de manter uma interacção com o respectivo utilizador tão próxima quanto possível de um diálogo convencional e utilizando fala natural. Contudo, os reconhecedores actuais não são capazes, em geral, de reconhecer este tipo de fala. Por forma a garantir a obtenção de material de fala mais controlado, exige-se frequentemente a leitura de textos.

A ausência de contacto visual com um interlocutor é uma situação comum devido à utilização muito frequente do telefone na comunicação verbal. Não só o interlocutor como também a ausência de referências em relação ao ambiente do outro lado do canal, alteram profundamente o processo de comunicação. A recolha de corpora de fala por via telefónica permite capturar as especificidades deste canal e deste tipo de interacção para o desenvolvimento de aplicações na rede telefónica. Além disso, a recolha é efectuada numa situação que é familiar ao orador em contraste com a recolha efectuada no laboratório. O advento dos telefones com imagem poderá alterar este cenário e faz prever um interesse crescente na recolha dos já designados *corpora multimodais*. Neste caso, pode incluir-se nos corpora a imagem que cada interlocutor tem do outro, ou pelo menos, alguns parâmetros que caracterizem essa imagem.

Recentemente têm sido recolhidos alguns corpora, designados por *corpus multilíngua*, cujo vocabulário é repetido em diversas línguas. Estes corpora possuem especificações comuns para todas as gravações, as quais são frequentemente efectuadas no país onde cada língua é a língua oficial. São importantes para os estudos interlíngua, para o desenvolvimento de sistemas de identificação automática da língua (secção 6.1.2) e de aplicações de reconhecimento de âmbito internacional.

Existe um outro tipo de corpus, menos comum, em que, embora gravado numa única língua, se cruzam de algum modo várias línguas. Estes corpora utilizam vários grupos de oradores com diferentes línguas maternas, pelo que, para a maioria desses grupos, a língua gravada representa uma segunda língua. Estes corpora são aqui designados por corpora *multissotaque*.

3.2 Corpus SUNSTAR multissotaque

Nesta secção descreve-se o corpus multissotaque que serviu de base de trabalho para os estudos efectuados nos capítulos seguintes. Este corpus inclui alguns sotaques dominantes da União Europeia na pronúncia de vocabulários em inglês. Os vocabulários considerados referem-se a aplicações típicas do reconhecimento de fala.

A selecção deste corpus deve-se ao facto de ser o único corpus de fala com sotaque estrangeiro, disponível à data do início deste trabalho.

3.2.1 Motivações da criação do corpus

O corpus aqui designado por *SUNSTAR multissotaque*, surgiu na sequência do trabalho desenvolvido no projecto Esprit 2094 "Integration and Design of Speech Understanding Interfaces" (SUNSTAR) que decorreu entre Junho de 1989 e Setembro de 1992. No consórcio SUNSTAR participaram importantes companhias europeias no sector das telecomunicações: a "Jydsk Telefon" (actual TeleDanmark) da Dinamarca, o Instituto Fraunhofer, a "AEG Eletrocom GmbH" e a "Telefunken System Technik" da Alemanha, a "Telefónica Investigación y Desarrollo" de Espanha e a "Alcatel FACE" de Itália. Para além destes parceiros considerados industriais foram subcontratados três parceiros associados a universidades, o "Speech Techonology Centre" (actual "Center for Person-Kommunication") da Universidade de Aalborg, a Universidade de Estugarda e o INESC de Lisboa. Cada um dos parceiros industriais pretendeu demonstrar as vantagens do reconhecimento automático de fala com uma aplicação própria, baseada num vocabulário com algumas dezenas de palavras isoladas (excepção para a aplicação italiana que além de palavras isoladas também utilizava algumas frases completas). Cada parceiro gravou no seu próprio país e na respectiva língua oficial, o vocabulário correspondente à sua aplicação. Para tal, seleccionou cerca de cem oradores nativos que repetiram duas vezes cada uma das palavras do vocabulário nacional. Obteve-se deste modo um corpus que, com alguma dificuldade, se poderia considerar multilíngua, já que os vocabulários de cada aplicação apenas partilhavam algumas palavras com significado comum (por exemplo: os dígitos).

O carácter demonstrativo destes sistemas e o enquadramento europeu inerente ao projecto, determinou que todos os demonstradores deveriam ter, igualmente, réplicas com a língua inglesa. De facto, os revisores do projecto, nomeados pela Comunidade Económica Europeia, exigiram uma avaliação oficial dos demonstradores em inglês. Estes demonstradores foram também essenciais em apresentações posteriores, em reuniões e conferências

internacionais cujas audiências são, em termos de nacionalidades, particularmente heterogéneas e onde a língua oficial é invariavelmente o inglês. Por forma a garantir um desempenho tão bom quanto possível para estes reconhecedores, pretendeu-se dispor de material de fala representativo para o treino e teste de modelos (HMMs, à excepção do demonstrador italiano que era baseado na técnica DTW). Para tal, realizaram-se gravações em todos os países representados no projecto, com alguns dos oradores anteriormente utilizados, dos vocabulários traduzidos em inglês de todas as aplicações. Considerou-se igualmente importante a gravação de oradores nativos do inglês europeu. Cada um destes cinco sotaques europeus do inglês ficou representado, em média, por cerca de dez homens e dez mulheres, num total de cem oradores. Cada orador repetiu duas vezes todos os vocabulários. Uma vez que existiam palavras idênticas nestes vocabulários, o vocabulário resultante apresenta algumas palavras com mais do que duas repetições por orador.

Posteriormente, no âmbito deste trabalho, gravaram-se em Lisboa as palavras isoladas destes vocabulários, também traduzidas em inglês. Utilizaram-se onze oradores nativos do português de cada sexo que repetiram duas vezes todo o vocabulário. Foi portanto a este conjunto de gravações realizadas exclusivamente em língua inglesa que se convencionou aqui designar por *corpus SUNSTAR multissotaque*. As limitações mais notórias da representatividade deste corpus são o tamanho do vocabulário e o número de oradores.

3.2.2 Descrição dos oradores

Seguidamente descrevem-se os aspectos conhecidos dos grupos de oradores de cada país que colaboraram na recolha do corpus SUNSTAR multissotaque.

Alemães (de) As gravações dos oradores alemães foram efectuadas nos laboratórios do Instituto Fraunhofer em Estugarda. As respectivas pronúncias devem apresentar características típicas da zona do Bade Vurtemberga.

Dinamarqueses (da) Os oradores dinamarqueses em questão habitam na região central da Jutlândia, nas imediações de segunda cidade mais importante da Dinamarca, Aarhus. Os dinamarqueses costumam distinguir alguns sotaques regionais, nomeadamente as diferenças entre o dinamarquês falado na zona continental (Jutlândia) do falado nas zonas insulares (Odense e Copenhaga).

Espanhóis (es) Os laboratórios do operador público "Telefónica Investigación y Desarrollo" de Madrid assumiram a responsabilidade pela recolha da parte do corpus com sotaque espanhol. O sotaque predominante deverá apresentar características

típicas do castelhano. Devido à dificuldade em seleccionar oradores do sexo feminino, utilizaram-se oradores do sexo masculino por forma a obter-se um número de repetições igual ao exigido para cada sotaque. Assim, o número de locuções recolhidas de oradores do sexo masculino é aproximadamente o triplo do obtido com oradores do sexo feminino.

Ingleses (en) A parte do corpus com sotaque nativo foi encomendada ao "University College of London". Os oradores utilizados na recolha do corpus apresentam o sotaque padrão do inglês europeu.

Italianos (it) A empresa italiana de telecomunicações "Alcatel FACE" foi a responsável pela recolha da parte do corpus que deveria representar o sotaque italiano. Para tal, recrutou habitantes dos arredores de Roma que frequentavam um curso avançado de inglês. Tal como aconteceu na recolha do sotaque espanhol, também se verificaram dificuldades na selecção de oradores do sexo feminino. Assim, o número de locuções recolhidas de oradores do sexo masculino é aproximadamente o dobro do obtido com oradores do sexo feminino.

Portugueses (pt) Os oradores portugueses eram funcionários e investigadores do INESC de Lisboa, com idades compreendidas entre os 20 e os 40 anos. Na sua maioria habitam em Lisboa e arredores desde a infância. Apenas dois oradores do sexo masculino e dois do feminino eram fumadores regulares (aproximadamente 20% do corpus dos portugueses).

É importante realçar que apesar de em cada país os oradores serem geralmente recrutados no mesmo departamento de pessoal, as competências de leitura e de pronúncia apresentadas pelos oradores não nativos variam consideravelmente. Os erros de leitura variam entre 5 a 10% e afectam principalmente as vogais. Por exemplo a vogal *i* é por vezes pronunciada erroneamente como um ditongo quando deveria ser pronunciada como /i/ e vice-versa. Os erros de pronúncia são neste caso principalmente devidos a diferenças nos inventários fonémicos entre a língua materna do orador e o inglês. Por exemplo o som /θ/, que não ocorre em português, é frequentemente aproximado pelo som nativo /s/ (Teixeira et al., 1996).

Embora se tivesse obtido uma relativa homogeneidade nos oradores recrutados em cada país, tal facto implica, por outro lado, que esse grupo nunca pode representar em termos gerais a população desse país. Mesmo na Dinamarca, o país mais pequeno em área e número de oradores no grupo de países representados, é fácil distinguir o sotaque no inglês de um orador oriundo da Jutlândia de um outro de Copenhaga. Em relação

aos oradores seleccionados para representar o inglês nativo europeu, este aspecto é ainda mais flagrante. Em geral, um ouvinte estrangeiro consegue distinguir facilmente várias pronúncias na própria Inglaterra. As diferenças em relação à Escócia, País de Gales, Irlanda do Norte e Irlanda são igualmente evidentes.

As línguas germânicas e românicas representam duas das subfamílias mais importantes no ramo europeu da família de línguas indo-europeias (Akmajian et al., 1990a). O corpus SUNSTAR multissotaque utilizou aproximadamente o mesmo número de oradores com línguas maternas pertencentes a cada uma destas subfamílias.

3.2.3 Meios e procedimentos de recolha do corpus

A recolha do corpus SUNSTAR multissotaque seguiu com rigor as especificações do projecto SAM (Tomlinson, 1990; Ribeiro et al., 1993). De acordo com estas especificações, a sala onde é instalado o orador deve possuir condições especiais de atenuação de reverberações e do ruído exterior, tal como acontece nas designadas câmaras surdas e anecóicas. O orador fica sentado numa cadeira com apoio de cabeça por forma a manter com mais eficácia a distância em relação ao microfone. Dispõe de um microfone B&K 4155 de meia polegada e de um monitor de vídeo onde são apresentadas mensagens com o que deve ser dito em cada momento. Além disso, coloca uns auscultadores de onde recebe indicações adicionais do operador instalado numa sala exterior. Nessa sala, o sinal analógico é calibrado num pré-amplificador (B&K 2230 Sound Level Meter) e colectado por uma estação de trabalho que inclui uma placa de aquisição OROS AU22 instalada num computador pessoal 486/33. O sinal analógico é amostrado a 20 kHz e quantificado a 16 bits. A estação regista o sinal digitalizado num disco rígido donde é posteriormente transferido em "off-line" para uma unidade de fita magnética de 8mm (Python DAT). O programa que opera esta estação (Europec versão 4.0) permite instruir o orador, através do monitor no interior da câmara, com a sequência de locuções pretendida actuando simultaneamente um detector de início e fim de palavra (subsecção 2.2.3). Este detector rudimentar "on-line" é baseado num único limiar de energia, garantindo margens de silêncio relativamente grandes antes e depois da locução (0,2 e 2 segundos respectivamente). As sessões de recolha realizadas com cada orador ficam simultaneamente registadas, na íntegra, em cassetes DAT (gravador Sony DTC-57ES). A transcrição ortográfica com que se instrui o orador é utilizada na anotação da respectiva locução. Esta é única forma de anotação disponível no corpus, contudo, a respectiva segmentação foi ainda sujeita a melhoramentos "off-line", conforme se descreverá mais adiante.

Os dados armazenados em cerca de uma dúzia de fitas magnéticas de 8mm são pos-

teriormente de novo transferidos para discos rígidos instalados em estações de trabalho. Nestas estações, dispõe-se de um conjunto de programas especificamente desenvolvidos para o processamento de sinal de fala (SIRtrain, 1991; Jacobsen, 1992) que permitem nomeadamente, a reamostragem do sinal e a detecção de início e fim de palavra. Uma das vantagens mais significativas deste conjunto de programas, reside no facto de todos os dados de entrada e de saída obedecerem aos formatos especificados pelo projecto SAM. Constitui assim um elo totalmente compatível entre o software de recolha do sinal (Europec) e o reconhecedor inicialmente utilizado, que reconhecem os mesmos formatos.

O sinal inicialmente amostrado a 20 kHz, foi reamostrado ("downsampling") para 16 kHz, sendo aplicado um filtro de resposta impulsiva finita com 131 coeficientes e com uma largura de banda -3 dB de 6,4 kHz. O sinal foi posteriormente de novo reamostrado, por forma a obter-se um sinal a 8 kHz. Nesta reamostragem é utilizado um filtro de resposta impulsiva finita com 69 coeficientes e uma largura de banda a -3 dB de 3,4 kHz.

O sinal amostrado a 8 kHz é posteriormente segmentado com um programa de detecção de início e fim de palavra. O método implementado neste programa foi descrito na subsecção 2.2.3. Na tabela 3.1, apresentam-se as durações médias das palavras existentes neste corpus de acordo com os limites determinados com este programa, à excepção dos oradores alemães e italianos. Durante a recolha do corpus referente a estes dois grupos de oradores, o programa de recolha (Europec) terá sido ajustado por forma a minimizar os intervalos de silêncio antes e depois de cada locução. Este ajuste determinou a eliminação de uma duração significativa destes silêncios. Nestas condições, o método descrito na subsecção 2.2.3 elimina posteriormente, quase sempre, a totalidade do sinal. Assim, este método não foi aplicado nesta parte do corpus, retendo-se as segmentações originais produzidas pelo Europec. Como consequência, os referidos silêncios ficaram com uma duração ligeiramente superior à restante parte do corpus. É assim possível comparar os valores descritos na tabela 3.1 que foram obtidos com os restantes grupos de oradores, uma vez que decorrem de condições idênticas. Entre estes, verificam-se algumas tendências de forma mais ou menos consistente e independente do sexo: os oradores espanhóis têm tendência a articular as locuções mais rapidamente, seguidos dos dinamarqueses; os portugueses são os que demoram mais tempo a pronunciar esta locuções embora os respectivos oradores do sexo feminino apresentem tempos médios praticamente iguais aos obtidos com os oradores britânicos do mesmo sexo.

Após a determinação automática dos segmentos de sinal de fala a serem utilizados, procedeu-se a uma verificação manual em que todos estes segmentos foram ouvidos por forma a serem detectados e eliminados alguns segmentos defeituosos que pudessem desequilibrar o tipo de estudo aqui proposto. Como segmentos defeituosos foram considerados

Sotaque	Género	
	feminino	masculino
da	679	640
de	1679	1706
en	746	825
es	618	596
it	992	1021
pt	895	826

Tabela 3.1: Durações das locuções (em milissegundos). Cada estimativa foi calculada com base em todo o material disponível (no corpus utilizado neste trabalho) para cada sexo e sotaque dos oradores.

todos aqueles que incluiam ruídos ocasionais provocados pelo orador, pequenas hesitações ou falhas involuntárias de articulação.

Posteriormente e uma vez removidos a quase totalidade dos segmentos de sinal sem sinais de fala ou com fala defeituosa, procurou-se armazenar todo o corpus num meio mais acessível e seguro. Por forma a optimizar o espaço de armazenamento disponível, os sinais de fala foram sujeitos a um processo vulgar de compressão de dados que utiliza a codificação Lempel-Ziv (LZ77). Desta forma, foi possível condensar todo o corpus num único disco compacto (CD-ROM).

A recolha dos sinais de fala dos oradores portugueses foi efectuada no ano de 1993 (cerca de dois anos depois dos restantes sinais de fala) na câmara anecóica do CAPS (Centro de Análise e Processamento de Sinais[2]). Trata-se de uma câmara com um volume interno de $76,3m^3$ ($5,3m \times 3,8m \times 3,8m$) ou $175m^3$ ($7m \times 5m \times 5m$) de volume externo se se considerar o material de isolamento sonoro.

Um estudo detalhado das características desta câmara (Brázio et al., 1979) determinou: a variação do ruído de fundo com a frequência; o desvio da lei de variação da pressão correspondente à energia radiada, com o inverso da distância da fonte sonora; e a atenuação de sinais no exterior da câmara para diversas frequências. A título ilustrativo destas características refira-se que a atenuação obtida a 2 kHz é de 70 dB entre o interior da câmara e o espaço circundante da câmara e de 25 dB entre este espaço e o exterior. De realçar ainda que o exterior corresponde a espaços de gabinetes de estudo e como tal

[2]Complexo Interdisciplinar do Instituto Superior Técnico em Lisboa

relativamente silenciosos.

A câmara dispõe ainda de uma rede metálica que permite colocar o orador aproximadamente no seu centro geométrico. O microfone foi suspenso a partir do tecto.

3.3 Outros corpora multissotaque

Existem outros corpora além do SUNSTAR multissotaque que incluem fala de oradores estrangeiros, quase todos em língua inglesa.

NATO AC/243 "language data base" — A sigla "DRG/AC243/Panel III/RSG 10" refere um grupo de investigação da NATO interessado em aplicações militares do processamento de fala. Um dos aspectos que procurou esclarecer foi o efeito da língua no desempenho dos sistemas de reconhecimento automático. Para dar início a este estudo, os países membros do "Research Study Group (RSG)" recolheram e permutaram entre si um corpus com quatro línguas: alemão, francês, inglês, (europeu e americano) e holandês. Foram utilizados 19 oradores para gravar um vocabulário de dez dígitos num total de 21.400 palavras isoladas e mais de 25.000 sequências de dígitos. Este corpus foi utilizado para testar sistemas comercialmente disponíveis dos países membros do RSG 10. Embora de dimensões modestas, foi o primeiro corpus multilíngua de que se tem registo (Vonusa et al., 1982; Moore, 1986), para o estudo de aplicações de reconhecimento automático de fala. Simultaneamente, é também o primeiro corpus multissotaque conhecido. Efectivamente nove dos oradores utilizados (4 alemães, 2 franceses e 3 holandeses) fizeram simultaneamente gravações na sua língua nativa e em inglês.

SC1 (Accents) "Strange Corpus 1" — Inclui a história do "Nordwind und Sonne" lida em alemão por 16 oradores nativos e 72 oradores apresentando sotaque estrangeiro. Os dados foram compilados num disco compacto (do catálogo da ELRA[3]).

TED "Translanguage English Database" — É um corpus criado com gravações das apresentações orais em inglês feitas no decorrer da conferência internacional Eurospeech'93 em Berlim. O nome deste corpus (igualmente conhecida nos meios da especialidade por "The Terrible English Database") traduz a alta percentagem de

[3]ELRA é a sigla de "European Language Resources Association" para a organização sem fins lucrativos criada no Luxemburgo em 1995. Os seus objectivos são os de centralizar a organização da validação, gestão e distribuição de recursos e utensílios para texto e sinais de fala. Deve ainda promover o uso destes meios junto da comunidade de investigação e desenvolvimento tecnológico da União Europeia (http://www.icp.grenet.fr/ELRA/).

comunicações feitas por oradores não nativos do inglês. Consiste em 224 apresentações de cerca de quinze minutos sobre um tópico específico, mais cinco minutos de discussão com alguns dos outros participantes, num total de cerca de 75 horas de material de fala. Como material de texto associado, dispõe-se dos artigos publicados na acta da respectiva conferência e ainda de questionários feitos a estes oradores. Um subconjunto deste corpus, com 188 apresentações e excluindo os períodos de discussão, é disponibilizado em cinco discos compactos (TEDspeeches). Existe ainda um subconjunto do corpus TED que foi recolhido separadamente de acordo com regras semelhantes às dos corpora "Polyphone" (o corpus TEDphone).

COST232 "Multi-English Speech Database" — O consórcio COST232 ("Speech Recognition Over the Telephone Line") recolheu em Junho e Julho de 1993 um corpus, em inglês, via rede telefónica com oradores de doze países da Europa: Alemanha, Bélgica, Checoslováquia, Dinamarca, Eslovénia, Espanha, Inglaterra, Itália, Noruega, Portugal, Suécia, e Suíça. O vocabulário seleccionado inclui o nome do orador, as designadas *palavras TI*[4] e o nome do parceiro do consórcio correspondente ao país do orador (FTZ, Univ. Gent – F.P. Mons, Czech T. Univ., JT, Univ. Ljubljana, ETSIT, BT, FUB, NTH, Univ. Coimbra, Telia Res., Ascom – Swiss Telecom).

Cada país contribuiu com oito oradores, tendo cada um feito oito chamadas utilizando linhas e equipamentos terminais diferentes (total de 64 chamadas por país).

Fora do espaço europeu é também conhecida a existência de alguns corpora de fala com características semelhantes aos anteriores. Verifica-se, contudo, que os corpora seguintes foram recolhidos em países de língua oficial inglesa (Austrália e Estados Unidos da América). Os oradores estrangeiros utilizados são em geral residentes nestes países e com uma forte motivação para aumentarem o seu desempenho com a língua local. Por outro lado, a maioria dos oradores utilizados na recolha dos corpora europeus nunca viveram em países de língua oficial inglesa (pelo menos num período significativo da sua vida) e utilizam a língua inglesa esporadicamente. Estes aspectos determinam diferenças importantes na produção da fala, de acordo com o que é conhecido, nomeadamente, acerca da aprendizagem de L2 (Schulte-Pelkum, 1976; Best, 1995; Bohn, 1995; Boysson-Bardies e Halle, 1995; Flege, 1995; Kuhl, 1995; Llisterri, 1995; McAllister, 1995; Polka, 1995; Rochet, 1995; Strange, 1995a; Mixdorff, 1996).

[4] O corpus de fala TI46 foi recolhidos pela companhia americana "Texas Instruments, Inc." (TI) e contém um vocabulário de 46 palavras que consiste no alfabeto de 26 letras e um conjunto de vinte palavras conhecido na gíria da especialidade pelas *palavras TI*. Este subvocabulário inclui os dígitos em inglês de zero a nove e as palavras de comando "enter", "erase", "go", "help", "no", "rubout", "repeat", "stop", "start", e "yes".

Na realidade, a maioria dos estudos relacionados com língua estrangeira ocorrem em países de língua inglesa que possuem em geral uma grande percentagem de população imigrante, ficando assim facilitada a recolha de sinais de fala estrangeira. Pelo contrário, na Europa, a motivação prática para estes estudos dirige-se essencialmente para pessoas temporariamente deslocadas do seu país de origem, por motivos de trabalho ou lazer. A União Europeia tem incentivado novas formas de comunicação e intercâmbio económico e cultural entre os seus estados membros, o que, mesmo que indirectamente, constitui uma motivação para este tipo de estudos.

ANDSOL "Australian National Database of Spoken Language" — É uma base de dados que inclui um corpus de fala em inglês falado por estrangeiros a residir na Austrália. Robin King e os seus colegas da universidade de Sydney publicaram estudos (Blackburn et al., 1993; Kumpf e King, 1997a; Kumpf e King, 1997b) que visavam classificar automaticamente os sotaques de oradores originários de três países diferentes: Austrália, (oradores nativos) Líbano e Vietname do Sul.

Universidade de Duke — John Hansen e os seus colegas desta universidade em Durham, (EUA) recolheram um corpus de fala de inglês americano (Arslan, 1996; Arslan e Hansen, 1996). Com base na vasta literatura disponível acerca da educação do inglês como segunda língua, seleccionaram um vocabulário de teste com vinte palavras isoladas e quatro frases. Cada um destes elementos do vocabulário foi repetido cinco vezes por cerca de uma centena de oradores apresentando os sotaques seguintes: inglês americano neutro, alemão, chinês, turco, francês, persa, espanhol, italiano, hindu, romeno, japonês, grego, etc.

FAE "Foreign Accented English Corpus"[5] — Disponível desde Março de 1998, contém 4961 frases de inglês americano pronunciadas por oradores nativos de 22 línguas, incluindo o português europeu e do Brasil. As frases foram recolhidas por via telefónica, sendo cada orador solicitado a falar sobre de si próprio em inglês durante 20 segundos. Os sinais de fala foram amostrados a 8 kHz e armazenados no formato "μ-law" de 8 bits.

Entre outras vantagens este corpus inclui os resultados de testes perceptuais (secção 6.7). Todas as locuções foram classificadas por três oradores nativos do inglês americano, com base na variação fonética. Ou seja, não foram considerados erros gramaticais ou de escolha de palavras. A classificação consiste na detecção do sotaque de acordo com a escala seguinte:

[5] "URL: http://www.cse.ogi.edu/CSLU/corpora/fae/".

1. Sem sotaque: sem nenhum sotaque ou, se existente, difícil de detectar.
2. Sotaque moderado: o sotaque é notado durante a maior parte da locução, mas não impede a compreensão.
3. Sotaque acentuado: o sotaque é acentuado durante toda a locução e torna a compreensão difícil.
4. Sotaque muito acentuado: a compreensão é muito perturbada e são necessárias múltiplas audições da locução por forma a compreender o orador.

Com base nestas decisões gerou-se uma lista das locuções que foram classificadas com 1 por um dos ouvintes e com 4 por um outro ouvinte. Estas classificações foram posteriormente revistas. Além disto, o corpus inclui ainda matrizes de confusão (secção 2.7) entre as decisões de cada par de ouvintes.

3.4 Conclusões

Neste capítulo apresentou-se o corpus de fala utilizado na maioria das experiências realizadas no decorrer dos estudos apresentados nos capítulos seguintes. Descreveram-se, para além dos detalhes técnicos, as diversas motivações que estiveram na origem da criação deste corpus. Descreveram-se igualmente, embora de forma mais resumida, outros corpora multissotaque actualmente disponíveis.

Capítulo 4

Detecção de palavras-chave

4.1 Introdução

Como se referiu em capítulos anteriores o reconhecimento automático de fala é uma tarefa complexa que apenas tem conseguido alguns resultados práticos de interesse em domínios limitados. As limitações referem-se, nomeadamente aos ambientes acústicos em que a fala é produzida, ao conteúdo linguístico do sinal de fala e às características dos oradores. Os trabalhos científicos na área do reconhecimento têm procurado ultrapassar, tanto quanto possível, estas limitações. Nesta perspectiva, o assunto deste capítulo enquadra-se no âmbito do conteúdo linguístico do sinal de fala. Por sua vez, os assuntos dos capítulos 5 e 6 enquadram-se melhor no âmbito das características dos oradores. O presente capítulo debruça-se sobre uma restrição que limita, em geral de forma drástica, o conteúdo linguístico do sinal de fala e que é muito comum no reconhecimento automático: a utilização de vocabulários de pequena ou média dimensão.

As tarefas de reconhecimento com maior complexidade requerem um vocabulário de média ou de grande dimensão. Contudo, em particular se o reconhecedor disponível for de palavras isoladas, a aplicação é geralmente desenhada de modo a que a interacção com o orador seja efectuada por etapas. Este tipo de interacção é semelhante ao utilizado nos programas de computador em que se introduzem parâmetros através de sucessivas listas de opções. Em cada uma destas etapas o reconhecedor considera apenas uma pequena parte do vocabulário total, designada por *subvocabulário* ou de *vocabulário activo*. Ou seja, os vocabulários activos dos reconhecedores de palavras isoladas possuem em geral umas escassas dezenas de palavras. Estes vocabulários consistem, quase exclusivamente, em substantivos e verbos. Não incluem outras palavras comuns na fala espontânea, tais como pronomes, adjectivos, palavras de ligação ou de saudação. Por exemplo, numa

simples aplicação de resposta binária, do tipo *sim* ou *não*, é habitual o seu utilizador responder com um *sim se faz favor* ou *não muito obrigado* ou ainda *agora não*. A obtenção de modelos que contemplem a totalidade ou a maioria das frases possíveis, representa um custo acrescido que ultrapassa largamente o custo de desenvolvimento da aplicação inicial, quer na concepção, quer na recolha dos sinais de fala representativos.

O uso de palavras adicionais aumenta, em geral, a redundância e como tal a robustez da comunicação entre humanos. Contudo, para os referidos reconhecedores de fala, estas palavras causam habitualmente graves quebras no seu desempenho global. Este problema e as respectivas soluções formam uma área de trabalho no reconhecimento de fala conhecida pela designação de *rejeição de palavras* ou de *detecção de palavras-chave* ("keyword spotting"). A designação de *palavra-chave* aplica-se habitualmente aos elementos de um vocabulário de pequena ou média dimensão. Para maior simplicidade designam-se aqui os respectivos modelos por *modelos-chave*. As palavras que não pertencem ao vocabulário activo do reconhecedor são referidas no texto seguinte pela designação de *palavras estranhas* e os respectivos modelos por *modelos de escoamento* (Chigier, 1992; Teixeira e Trancoso, 1992).

O conceito de modelo de escoamento decorre, em termos práticos, da utilização dos designados modelos de silêncio e dos modelos de "background" (Wilpon et al., 1990). Os modelos de "background" são obtidos a partir dos ruídos de fundo do sinal de fala, tais como os existentes, por exemplo, no interior de uma viatura a velocidade elevada, na transmissão por linha telefónica, ou numa estação central de camionagem. Em qualquer destes casos, pretende-se modelar segmentos do sinal de fala que não deverão ser relevantes para o reconhecimento, de modo a ser mais robusto em termos do ruído. Na perspectiva aqui apresentada, os modelos de "background" funcionam como alternativa, num determinado instante de tempo, aos modelos de palavra. Ou seja, os modelos de "background" contribuem, no essencial, para uma determinação mais robusta da segmentação do sinal de fala. Na realidade, os ruídos modelados deverão estar sobrepostos ao longo da duração de toda a locução, perturbando ainda o reconhecimento. Existe, contudo, a possibilidade de utilizar estes modelos na descodificação de um segundo processo de Markov, em paralelo com o primeiro, dedicado exclusivamente à fala limpa de ruídos (Varga e Ponting, 1989; Varga e Moore, 1990).

Neste capítulo descrevem-se algumas experiências de detecção de palavras-chave em que se procurou encontrar soluções mais eficazes para diversas aplicações de reconhecimento destinadas a oradores nativos e não nativos. Seguidamente apresenta-se a estrutura deste capítulo.

Na secção que se segue, referem-se as soluções conhecidas para o problema da detecção de palavras-chave.

Na secção 4.3, apresentam-se alguns dos parâmetros mais comuns utilizados na avaliação do desempenho das técnicas de detecção de palavras-chave. Estes parâmetros são essenciais quer para a comparação entre técnicas diferentes, quer para a afinação de determinada técnica.

Seguidamente, na secção 4.4, descrevem-se as primeiras experiências (Teixeira e Lindberg, 1992) que indicaram a possibilidade de se obterem ganhos significativos no desempenho da detecção de palavras-chave através do uso de modelos de escoamento múltiplos.

Na secção 4.5, procurou-se determinar procedimentos para a distribuição do material de fala disponível para o treino de modelos de escoamento múltiplos.

Na secção 4.6, estudou-se a relação entre o número de modelos de escoamento utilizados e a dimensão do vocabulário da aplicação. Nomeadamente, verificou-se a necessidade de utilizar maior número de modelos de escoamento quando a dimensão do vocabulário era maior. De igual modo se verificou não ser necessário mais do que um único modelo de escoamento quando a dimensão do vocabulário da aplicação era muito pequena (≤ 5 palavras). Confirmou-se, deste modo, as conclusões anteriores de investigadores reconhecidos que realizaram um estudo semelhante, embora com o uso exclusivo de um vocabulário com estas dimensões (Wilpon et al., 1990).

Na secção 4.7, procuraram-se algumas características do sinal de fala adequadas para o treino de modelos de escoamento. Comparou-se o desempenho dos modelos HMM contínuos com o de um reconhecedor baseado em modelos semicontínuos em experiências de detecção de palavras-chave.

Na secção 4.8, estudou-se a aplicação da detecção de palavras-chave no contexto do reconhecimento da fala de oradores estrangeiros.

Na secção 4.9, descrevem-se algumas experiências com sinais de fala contínua. Determinou-se o modelo linguístico a utilizar e as condições em que é vantajoso o uso de modelos de escoamento múltiplos. Por último, apresentam-se as conclusões deste capítulo.

4.2 Metodologias conhecidas

Os estudos publicados sobre o problema da detecção de palavras-chave surgem com alguma insistência desde o final da década de 80 (Rohlicek et al., 1989; Wilpon et al.,

1989; Asadi et al., 1990; Rose e Paul, 1990) existindo contudo referências esporádicas a trabalhos anteriores das quais se destaca o de (Higgins e Wohlford, 1985). Referências a outros trabalhos encontram-se em (Wilpon et al., 1990; Rose e Paul, 1990).

Um detector de palavras-chave pode ser descrito como um classificador de padrões que deve distinguir padrões referentes a duas classes: as palavras-chave pertencentes ao vocabulário ou subvocabulário activo de uma aplicação específica e todas as restantes manifestações possíveis do sinal de fala. Esta última classe pode ser convenientemente representada pelas manifestações mais prováveis do sinal de fala no contexto específico da aplicação (Wilpon et al., 1989). Um reconhecedor de fala é também por inerência um classificador, o qual deve possuir uma descrição do conjunto dos padrões ou modelos das palavras-chave. Muito do conhecimento acumulado com os reconhecedores de fala tradicionais pode assim ser utilizado na detecção das palavras-chave. A obtenção de uma solução integrada no próprio reconhecedor deverá também ser vantajosa em termos de espaço de armazenamento de informação e de tempo de processamento.

Os detectores de palavras-chave surgem assim, na sua maioria, integrados em reconhecedores de fala. Os primeiros foram desenvolvidos com base na técnica DTW, enquanto que os mais recentes são baseados nos modelos de Markov não observáveis e em redes neuronais artificiais, existindo contudo algumas excepções (O'Kane e Kenne, 1993).

A estratégia mais adoptada para o problema da detecção de palavras-chave é exactamente a de detectar a ocorrência de alguma das palavras do vocabulário activo num determinado segmento de sinal, considerando todo o sinal restante como uma espécie de ruído, independentemente de este conter ou não outros sinais de fala. Este sinal de sobra pode, por sua vez, ser associado a diversos tipos de padrões ou modelos. Higgins e Wohlford (Higgins e Wohlford, 1985) utilizaram seis palavras-função ("the, of, for, at, to, from") e algumas unidades de fala semelhantes a sílabas para a obtenção dos designados *padrões de preenchimento* ("filler templates"). Nesta altura predominavam as técnicas de reconhecimento baseadas nos algoritmos DTW. Com a divulgação dos HMMs, o conceito de modelo substitui progressivamente o de padrão. Assim, Rose e Paul (Rose e Paul, 1990) utilizaram modelos de preenchimento acústicos de palavras e subpalavras baseadas em monofones e trifones. Actualmente, consideram-se também as sílabas para a obtenção destes modelos, tendo-se conseguido melhores resultados na detecção de palavras novas (El meliani e O'Shaughnessy, 1998). Rohlicek et al. (Rohlicek et al., 1989) criaram os designados *modelos alternativos*, que são treinados com segmentos das próprias palavras-chave. Wilpon et al. (Wilpon et al., 1989; Wilpon et al., 1990; Wilpon et al., 1991) utilizaram as palavras mais comuns, não pertencentes ao vocabulário, para treinar os designados *modelos de lixo* ("garbage models"). Posteriormente, alguns autores utilizaram

modelos semelhantes, embora com diferenças no tipo de material de treino, que designaram por *modelos de escoamento* ("sink models") (Chigier, 1992; Teixeira e Trancoso, 1992). Esta designação, utilizada como alternativa à de modelos de lixo, foi a adoptada no presente trabalho.

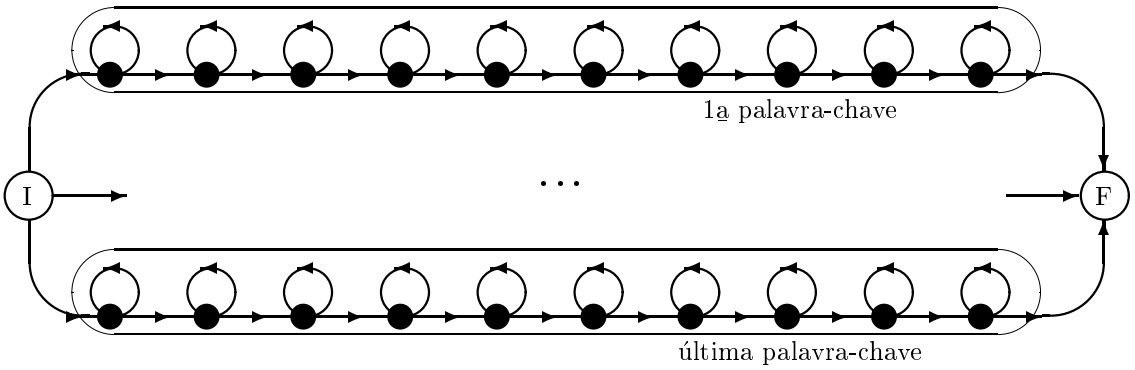
Os modelos de escoamento de Wilpon et al. são utilizados em simultâneo com um modelo de "background", exclusivamente destinado ao silêncio e aos ruídos da transmissão na linha telefónica (Wilpon et al., 1990). Algumas das experiências descritas avaliaram o desempenho do reconhecedor com a utilização de vários modelos de escoamento concluindo que não era justificável a utilização de mais do que um destes modelos. A tarefa de reconhecimento estudada é baseada num pequeno vocabulário de apenas cinco palavras o que permite sustentar a hipótese, discutida neste capítulo, sobre a utilidade do uso simultâneo de vários modelos deste tipo.

As soluções propostas para o problema da detecção de palavras-chave são, em geral, muito semelhantes variando principalmente com o tipo de reconhecedor utilizado, com as exigências de determinadas aplicações e com os sinais de fala disponíveis.

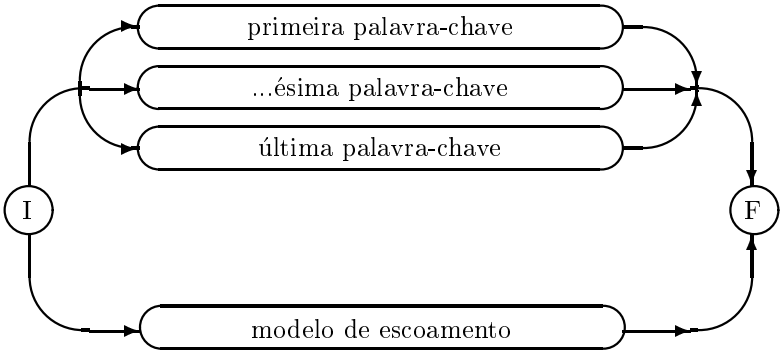
Para a realização das experiências seguidamente apresentadas dispunha-se de um corpus de fala construído com o vocabulário de diversas aplicações de reconhecimento de fala (subsecção 3.2). Houve necessidade de providenciar cada uma destas aplicações com um mecanismo de detecção de palavras-chave. Simultaneamente procurava-se evitar a recolha ou utilização de outros corpora. A estratégia adoptada para modelar as palavras estranhas ao vocabulário foi a utilização de modelos de escoamento específicos para cada aplicação. Estes modelos foram treinados com todo o corpus disponível à excepção da parte que incluía o vocabulário da aplicação em que seriam utilizados.

Também nos reconhecedores modernos, com vocabulários na ordem das dezenas de milhar de palavras, se pretende criar estratégias que permitam atenuar os efeitos causados pelo surgimento de palavras novas. Nos reconhecedores mais elaborados, após a detecção de cada uma destas palavras novas, são accionados procedimentos para a inserção da nova palavra de modo a fazer parte do vocabulário activo (Asadi et al., 1990; Asadi et al., 1991; Hetherington, 1995; El meliani e O'Shaughnessy, 1998). A *detecção de palavras novas* opõe-se ao conceito de rejeição de palavras, uma vez que as palavras não pertencentes ao vocabulário são agora consideradas elementos não desprezáveis para o reconhecimento. Os procedimentos envolvidos nesta detecção afectam necessariamente não só os modelos acústicos e linguísticos como também os modelos semânticos da aplicação (no caso de estes existirem).

No caso dos reconhecedores baseados em modelos HMM, os modelos de escoamento



(a)



(b)

Figura 4.1: Representação esquemática de um reconhecedor de palavras isoladas convencional (a) e com capacidade de rejeição de palavras (b).

podem ser da mesma natureza do que os modelos-chave e competir com estes em paralelo, no reconhecedor, por um valor máximo de verosimilhança. Deste modo, a topologia convencional de um reconhecedor de fala baseado em HMMs (figura 4.1a) é pouco alterada, conforme se pode verificar na figura 4.1b. Os modelos de escoamento deverão ser capazes de representar a generalidade dos sinais de fala, nomeadamente os que não contêm as palavras-chave, por forma a apresentarem verosimilhanças superiores aos obtidos com os modelos-chave. Devem assim permitir "escoar" as palavras que apresentam menores valores de verosimilhança nos modelos do vocabulário. Um modelo-chave representa uma única palavra, devendo por isso ser treinado apenas com repetições dessa mesma palavra. A diferença essencial entre estes modelos e os de escoamento reside no facto de estes últimos representarem todo o sinal de fala que não inclui palavras-chave. Devem, portanto, ser treinados com locuções do maior número de palavras possíveis, em princípio, diferentes das palavras-chave. Na secção 4.7, verificar-se-á que esta última condição não é tão

relevante como inicialmente se poderia supor.

Após a descodificação de uma palavra no reconhecedor representado na figura 4.1, os valores de verosimilhança fornecidos pelo algoritmo de Viterbi podem ser objecto de um processo de decisão mais elaborado do que a simples maximização (Feng e Mazor, 1992). Este processo pode mesmo socorrer-se de estimativas de características adicionais do sinal de fala, por forma a obterem-se melhores desempenhos na detecção de palavras-chave (Gish et al., 1992). Um tipo de problema para o qual têm surgido novas soluções é o da determinação de medidas de confiança que permitam avaliar convenientemente os resultados do reconhecimento automático (Gillik et al., 1997; Vaver, 1998). Estas medidas visam, nomeadamente, determinar qual a probabilidade de uma palavra reconhecida automaticamente estar efectivamente correcta. A definição mais óbvia de uma medida de confiança baseia-se na própria probabilidade a posteriori

$$-\log Pr(W|O).$$

Estas medidas podem ser utilizadas na detecção de palavras-chave por forma a ser dispensável o uso de modelos de escoamento (Junkawitsch et al., 1997).

4.3 Avaliação do desempenho da detecção de palavras

Nesta secção, aborda-se o problema da avaliação dos resultados obtidos com métodos de detecção de palavras-chave. Na secção 2.7, apresentaram-se diversas taxas que permitem avaliar com rigor o desempenho de um reconhecedor, no pressuposto de que todas as locuções a reconhecer correspondem a palavras-chave. No problema da detecção de palavras-chave, são igualmente apresentadas ao reconhecedor palavras estranhas ao respectivo vocabulário. Pretende-se agora avaliar, com parâmetros semelhantes, a capacidade adicional do reconhecedor em rejeitar estas palavras.

Os investigadores da área da detecção de palavras têm utilizado diversos parâmetros de avaliação de desempenho, sendo muitos deles equivalentes. Em termos do reconhecimento de palavras isoladas podem ser calculadas taxas de detecção, de rejeição e de alarme falso de uma forma simples tal como foi feito com a taxa de reconhecimento, a de inserção, a de supressão, a de exactidão e a do respectivo erro.

Considerem-se as seguintes quantidades:

N_p = número de palavras-chave apresentadas ao reconhecedor;

N_{p_p} = número de palavras-chave correctamente detectadas;

N_{p_e} = número de palavras-chave rejeitadas;

N_e = número de palavras estranhas apresentadas ao reconhecedor;

N_{e_e} = número de palavras estranhas correctamente rejeitadas;

N_{e_p} = número de palavras estranhas detectadas como sendo palavras-chave ou número de alarmes falsos;

$N_t = N_p + N_e = N_{p_p} + N_{p_e} + N_{e_e} + N_{e_p}$ número total de palavras apresentadas ao reconhecedor.

Quando N_p e N_e são suficientemente elevados, é possível obter as seguintes taxas:

$P_D = 100\% \; N_{p_p}/N_p$ = taxa de detecção (Rohlicek et al., 1989) é uma estimativa da probabilidade de detecção de (Higgins e Wohlford, 1985);

$P_F = 100\% \; N_{p_p}/N_p/T$ = taxa de alarmes falsos (Rohlicek et al., 1989). É uma estimativa da probabilidade de alarme falso de (Higgins e Wohlford, 1985). T é o tempo total de um sinal de fala durante o qual se contabiliza N_{p_p} e N_p. A unidade habitual é o número de alarmes falsos por palavra-chave e por hora (fa/kw/hr — "false-alarms per keyword per hour"). Esta medida de insucesso é particularmente adequada à fala contínua, quando se torna difícil de contabilizar N_e e consequentemente N_{e_e}. O mesmo não se passa no reconhecimento de palavras isoladas onde cada locução pode ser facilmente classificada;

$R_j = 100\% \; N_{e_e}/N_e$ = taxa de rejeição (Wilpon et al., 1989).

Com um reconhecedor típico, desprovido de qualquer protecção em termos de palavras estranhas ao seu vocabulário, obtém-se $R_j = 0\%$ e pode-se assumir que a taxa de reconhecimento se mantém inalterável, continuando a depender exclusivamente dos resultados obtidos com as palavras pertencentes ao vocabulário. O mesmo não se deverá passar com um reconhecedor capaz de rejeitar palavras estranhas ao respectivo vocabulário. O mecanismo de detecção de palavras-chave será tanto melhor quanto mais alta for a taxa R_j, mas deverá igualmente não rejeitar palavras-chave. Estes dois objectivos revelam-se na prática antagónicos, sendo necessário encontrar soluções de compromisso. Embora não tivessem sido utilizados neste trabalho, é possível impor limiares à verosimilhança obtida com os modelos de palavras-chave e com os de escoamento, por forma a controlar o referido antagonismo. Para se estudar este tipo de soluções utiliza-se a designada curva *característica*

de operação do receptor ou ROC ("receiver operating characteristic") representando a taxa de detecção P_D em função da taxa de alarmes falsos P_F (Higgins e Wohlford, 1985; Rohlicek et al., 1989). Grande parte dos trabalhos publicados sobre a detecção de palavras utilizam medidas de avaliação de resultados baseadas nos parâmetros aqui sistematizados, nomeadamente os referentes à obtenção da referida curva característica de operação do receptor (Clary e Hansen, 1992; Okawa et al., 1993; O'Kane e Kenne, 1993; Lleida et al., 1993; El meliani e O'Shaughnessy, 1995; Lleida e Rose, 1996). Alguns destes trabalhos utilizam a média de P_D para P_F entre 0 e 10 fa/kw/hr (Gish et al., 1992; Lau e Seneff, 1997) inicialmente proposta por (Rohlicek et al., 1989) e actualmente designada por *figura de mérito* ou FOM ("figure of merit"). Existem ainda outras propostas baseadas no integral da ROC (Marcus, 1992).

Por vezes, o reconhecedor pode identificar uma palavra estranha com uma das palavras-chave. Este tipo de erro é contabilizado na taxa de inserção definida na secção 2.7. Esta taxa varia de forma inversa da taxa de rejeição, constituindo assim uma medida equivalente. A rejeição incorrecta de uma palavra-chave é por sua vez contabilizada na taxa de supressão. A taxa de exactidão, ao considerar simultaneamente a taxa de inserção e de supressão, fornece assim uma medida global do desempenho do sistema de reconhecimento. Contudo, para efeitos de desenvolvimento e afinação do sistema de detecção de palavra-chave, é importante a análise em separado da taxa de reconhecimento e de rejeição. No presente trabalho, optou-se por este tipo de análise, à semelhança do efectuado por outros autores (Wilpon et al., 1989).

4.4 Modelos de escoamento múltiplos

Desde os primeiros trabalhos em detecção de palavras-chave que ocorreu a ideia de se dispor em vez de um único, vários padrões (Higgins e Wohlford, 1985) ou modelos (Wilpon et al., 1990) que deveriam permitir modelar, de forma mais eficaz, uma grande variedade de sinais de fala não incluindo palavras-chave (figura 4.2). A verificação desta ideia, revelou, contudo, vantagens irrisórias (Wilpon et al., 1990), face a algumas desvantagens mais evidentes, tais como um custo acrescido no processamento e armazenamento dos modelos. O facto de nestas experiências se terem utilizado, em geral, um número elevado de componentes gaussianas, permitiu por certo uma modelação eficaz do restante sinal de fala. Assim, Wilpon et al. utilizaram 9 componentes para cada um dos 10 estados dos seus modelos (Wilpon et al., 1990) e Komori e Rainton usaram 64 componentes para um único estado (Komori e Rainton, 1992). Deste modo, durante algum tempo e com raras excepções (HMMs com distribuições discretas (Feng e Mazor, 1992)) muitos dos

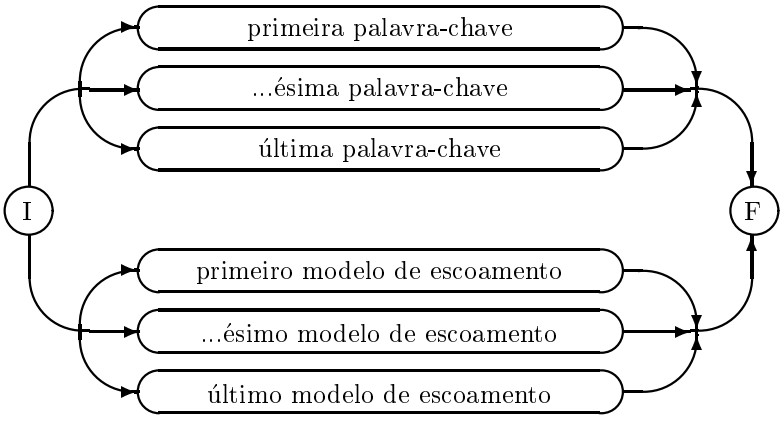


Figura 4.2: Representação esquemática de um reconhecedor de palavras isoladas com modelos de escoamento múltiplos.

trabalhos desenvolvidos com modelos de escoamento, não utilizaram mais do que um destes modelos (Komori e Rainton, 1992). Por vezes, surgem alguns modelos adicionais, procurando contudo modelar aspectos claramente distintos e bem identificados como é o caso de modelos de silêncio ou sem fala (Chigier, 1992).

Para o trabalho descrito na presente secção, dispunha-se unicamente de um reconhecedor baseado em HMMs com a função densidade de probabilidade de observação correspondente a cada estado descrita por uma única gaussiana. Esta limitação do reconhecedor foi posteriormente ultrapassada, pelo menos parcialmente, com o desenvolvimento de um reconhecedor baseado em observações semicontínuas, (secção 4.7) mas assinala uma diferença considerada significativa em relação às experiências originais de Wilpon et al. (Wilpon et al., 1989; Wilpon et al., 1990). Outra diferença significativa advém de se pretender aplicar a detecção de palavras-chave em vocabulários com um tamanho médio de algumas dezenas de palavras, em contraste com o pequeno vocabulário de 5 palavras então utilizado. Por último, enquanto que Wilpon et al. dispunham dos sinais que ocorriam com maior frequência juntamente com as palavras-chave, no caso presente apenas se dispunha de gravações de outras palavras correspondentes a outras aplicações. Para tornar os testes mais realistas testaram-se palavras estranhas diferentes das utilizadas no treino dos modelos de escoamento. Deste modo, a presente tarefa configura-se substancialmente mais complexa do que a utilizada no trabalho original dos referidos autores.

A necessidade de rejeitar palavras estranhas menos esperadas, com modelos estatisticamente menos poderosos alimentou a hipótese de se conseguirem melhores resultados com a utilização de vários modelos de escoamento. Cada um destes modelos deveria ser capaz de rejeitar determinado tipo de palavras estranhas. De facto, a existência de um

único modelo de escoamento sugere a possibilidade de uma palavra estranha, de alguma forma semelhante a alguma das palavras-chave, poder obter uma probabilidade maior junto do modelo correspondente a essa palavra-chave do que junto de um modelo de escoamento demasiado incaracterístico. Este problema deverá ser tanto mais visível quanto maior o vocabulário do reconhecedor, uma vez que aumenta a possibilidade de existir uma palavra-chave semelhante a cada uma das eventuais palavras estranhas.

Por outro lado, é previsível o surgimento de ligeiras quebras nas taxas de reconhecimento, uma vez que existe sempre a possibilidade da locução de alguma palavra-chave obter um valor de verosimilhança mais alto junto de um modelos de escoamento do que entre os próprios modelos-chave. Esta possibilidade aumenta com o número de modelos de escoamento, pelo que os aumentos que se consigam obter na taxa de rejeição deverão ser justificadamente compensadores face às eventuais quebras das taxas de reconhecimento. Naturalmente que a ponderação destes dois valores só pode ser feita com a realização de testes de campo. Estes deverão permitir a aferição da proporção média entre as palavras-chave e as palavras estranhas proferidas. De facto, verifica-se que o aumento do tempo de processamento é o maior inconveniente do uso de um número elevado de modelos de escoamento, uma vez que a descida associada da taxa de reconhecimento é habitualmente pouco significativa.

A forma de dividir o material de fala disponível para o treino de cada um dos modelos de escoamento múltiplo é o assunto da secção 4.5. Na presente secção utilizou-se um procedimento muito simples para este fim, que se designou por *método alfabético*. Ordenou-se alfabeticamente numa lista, o vocabulário de palavras estranhas disponível para o treino. A partir desta lista criaram-se n_s sublistas (n_s= número de modelos de escoamento pretendidos). Cada sublista retira sucessivamente uma palavra de cada vez da lista anterior até se esgotarem todas as palavras estranhas disponíveis para o treino. Deste modo, a m-ésima palavra da lista inicial é incluída na (m mod n_s)-ésima sublista. Por fim, todas as locuções das palavras incluídas em determinada sublista são utilizadas no treino de um dos modelos de escoamento múltiplos. Este procedimento procura distribuir de forma ordenada as palavras disponíveis na expectativa de que ocorram poucas correlações, em termos fonéticos, entre as palavras escolhidas para o treino do mesmo modelo de escoamento.

4.4.1 Condições experimentais

O reconhecedor de palavras isoladas utilizado (SIRtrain, 1991; Jacobsen, 1992) possui as características gerais descritas no capítulo 2. Sublinham-se algumas características

que são particularmente relevantes para a análise das experiências seguintes. Todos os modelos descritos nesta secção, quer de palavras-chave, quer de escoamento, possuem uma topologia linear com oito estados, sem transições que permitam saltar estados intermédios (figura 2.1). Nas experiências descritas a partir da secção 4.6 em diante, adoptaram-se modelos com uma topologia semelhante, mas com dez estados. O aumento do número se estados foi devido às palavras compostas existentes no vocabulário, as quais apresentam uma duração média superior às palavras simples. A cada estado corresponde uma função densidade de probabilidade de observação gaussiana com uma matriz de covariância diagonal.

Os sinais utilizados nesta secção fazem parte do corpus SUNSTAR multissotaque descrito no capítulo 3. Procurou-se testar um vocabulário com um número de palavras-chave representativo das aplicações representadas neste corpus. Seleccionou-se a aplicação de despertar automático "Alarm Call", desenvolvida pela companhia de telefones dinamarquesa Jydsk Telefon, cuja dimensão do vocabulário é de 40 palavras e que é um valor próximo da média da dimensão dos vocabulários das restantes aplicações. Nestas experiências preliminares procurou-se simultaneamente obter uma aproximação razoável do que aconteceria numa situação real, em que o sotaque do orador fosse diferente dos sotaques utilizados no treino. Assim, utilizaram-se as duas repetições de cada palavras-chave de 80 oradores. Para o teste utilizaram-se os oradores dinamarqueses, (25%) enquanto que para o treino dos modelos foram utilizados os restantes oradores de três outras nacionalidades: alemães, (25%) espanhóis (25%) e ingleses (25%). Cada um destes grupos tem um número aproximadamente igual de oradores do sexo feminino e masculino. Esta proporção foi também utilizada nas restantes experiências deste capítulo, salvo menção expressa em contrário.

As restantes gravações disponíveis do corpus SUNSTAR multissotaque, correspondentes a outras duas aplicações, foram utilizadas para o desenvolvimento e teste das capacidades de detecção das palavras-chave. Deste subconjunto de sinais foram eliminadas não só as palavras-chave, como também todas as palavras compostas que incluíam palavras-chave. Por exemplo, um dos elementos destes vocabulários era a palavra composta "stop voice" que foi eliminada pelo facto da palavra "stop" ser uma das palavras-chave da aplicação de despertar automático. O material restante foi ainda dividido em duas partes: uma parte, com uma única repetição de 69 palavras estranhas, foi utilizada para o treino dos modelos de escoamento; a outra parte tem duas repetições de 57 palavras estranhas, diferentes das anteriores, e foi utilizada no teste.

A escolha de palavras estranhas de teste intencionalmente diferentes das utilizadas no treino dos modelos de escoamento, simula um caso extremo em relação às situações reais

de aplicação. Efectivamente, nestes casos, surge fala quase espontânea na qual quaisquer dessas palavras têm, em geral, uma probabilidade muito baixa de ocorrer. Deste modo, os modelos de escoamento não devem ficar excessivamente especializados em algumas palavras estranhas. Por isso, excluiu-se do treino dos modelos de escoamento, a outra repetição disponível de cada orador para cada palavra estranha.

4.4.2 Resultados

Na tabela 4.1 apresentam-se os resultados obtidos com diversos reconhecedores construídos de acordo com as descrições da subsecção anterior. Cada linha desta tabela refere-se a um reconhecedor que difere dos restantes apenas nos modelos de escoamento utilizados ($E_1, E_3, \ldots$). Isto é, os modelos das palavras-chave são sempre os mesmos.

A primeira coluna da tabela 4.1 (E_0) corresponde a uma experiência de aferição do reconhecedor sem modelos de escoamento e portanto sem a capacidade de rejeitar palavras estranhas ao vocabulário. A taxa de reconhecimento obtida é aceitável tendo em conta a perplexidade da tarefa, as limitações do reconhecedor e principalmente o facto de se testar um sotaque diferente dos disponíveis no treino dos modelos.

O modelo de escoamento designado por E_1, na coluna seguinte, foi treinado com todo o material de fala descrito na subsecção anterior. Posteriormente, utilizando o *método alfabético*, dividiu-se este material em três partes iguais por forma a incluir 23 palavras estranhas diferentes em cada parte. Cada uma destas partes foi utilizada para o treino de um modelo de escoamento: e_1, e_2 e e_3. Deste modo, pretende-se averiguar da utilidade do uso de modelos de escoamento múltiplos em circunstâncias idênticas ao da utilização de um único destes modelos, nomeadamente com o mesmo material de treino.

As três colunas seguintes apresentam os resultados obtidos com cada um dos modelos de escoamento e_1, e_2 e e_3 treinados com uma quantidade equivalente de material de fala. Os resultados obtidos são bastante semelhantes entre si, com uma média de taxa de reconhecimento de 86,8% e de rejeição de 47,1%. Confirma-se uma ligeira quebra da taxa de reconhecimento, conforme foi referido no início desta secção. Também como seria de esperar, a utilização de um modelo de escoamento treinado com maior número de palavras estranhas (E_1) revelou-se vantajoso, com um aumento significativo (cerca de 14%) da taxa de rejeição.

A primeira experiência com modelos de escoamento múltiplos utiliza simultaneamente os três modelos e_1, e_2 e e_3 de acordo com a topologia representada na figura 4.2. Para simplificar, esta associação de modelos de escoamento é representada por E_3. Tal como se

previu, a taxa de reconhecimento diminuiu, embora de novo, de forma à qual não se pode atribuir grande significado (cerca de 0,1%). A taxa de rejeição, por sua vez, aumentou 12,3% em relação à experiência anterior com o modelo E_1 (cerca de 28% em relação ao uso de cada um dos modelos de escoamento e_1, e_2 e e_3).

modelos de escoamento	E_0	E_1	e_1	e_2	e_3	E_3	E_{10}	E_{20}
reconhecimento (%)	87,0	86,7	86,7	86,8	86,8	86,6	86,3	85,7
rejeição (%)	00,0	53,6	48,2	47,7	45,4	60,2	67,8	71,5

Tabela 4.1: Taxas (%) de reconhecimento e de rejeição obtidas com o uso combinado de vários modelos de escoamento (Teixeira e Lindberg, 1992).

Perante os resultados descritos, considerou-se de interesse ensaiar a utilização de um maior número de modelos de escoamento. Para tal, subdividiu-se uma vez mais o material de treino disponível por forma a se poderem comparar os novos resultados com os anteriormente obtidos. Utilizando um processo igual ao utilizado para o conjunto de modelos de escoamento E_3, consideram-se subconjuntos de palavras com 7 e de 3 palavras diferentes. Obtiveram-se assim novos conjuntos de 10 e 20 modelos de escoamento, aqui representados por E_{10} e E_{20}, respectivamente. Os resultados obtidos verificam a tendência para uma ligeira descida da taxa de reconhecimento que começa a ser significativa com o conjunto E_{20}. Em contrapartida, a taxa de rejeição aumenta de forma significativa com a adopção do conjunto E_{10} (26,5% em relação a E_1 e 12,6% em relação a E_3) mas não tanto com o uso do conjunto E_{20} (33,4% em relação a E_1 mas apenas 5,5% em relação a E_{10}). Além disso, com 20 modelos de escoamento, aumenta-se o tempo de processamento de forma indesejável (cerca de 50%). Verifica-se assim, que a escolha do número de modelos de escoamento se deve basear num compromisso entre a taxa de reconhecimento e a de rejeição. Com o aumento deste número, a taxa de reconhecimento aparenta descer cada vez mais depressa, enquanto que a de rejeição obtém ganhos cada vez menos significativos.

O número de modelos de escoamento aqui utilizados, não deve ser considerado em termos absolutos em outras experiências de reconhecimento. Será de esperar que, se a quantidade de material de treino for maior, se consigam ganhos superiores na taxa de rejeição com o uso de um maior número de modelos de escoamento. Atenda-se assim que, na experiência realizada com o conjunto E_{20}, cada modelo de escoamento dispõe de locuções de apenas três palavras diferentes para o seu treino.

Em aplicações em que seja previsível um número elevado de palavras estranhas por cada palavra-chave proferida, as taxas de rejeição obtidas não podem ser consideradas satisfatórias. Contudo, na falta de outro mecanismo mais promissor, estes resultados

revelaram-se encorajadores, nomeadamente para aplicações em que se preveja um surgimento esporádico de palavras estranhas.

4.5 Treino de modelos de escoamento

Na secção 4.4, o material disponível para o treino de cada um dos modelos de escoamento múltiplos foi seleccionado pelo referido *método alfabético*. Com o intuito de se poder aumentar a taxa de rejeição, pretende-se agora desenvolver procedimentos com o mesmo fim, mas que determinem a divisão do material de fala com base nas características desse mesmo material ("data-driven methods"). Considera-se que o número de modelos de escoamento (n_s) é pré-definido e adoptou-se, tal como na secção 4.4, a restrição de se manter o mesmo número de locuções para o treino de cada um destes modelos. Em geral, manteve-se a restrição que considera todas as locuções da mesma palavra no conjunto de treino do mesmo modelo de escoamento. Para verificar a utilidade desta restrição considera-se, num dos métodos que a seguir se descrevem, uma variante em que esta restrição é levantada.

4.5.1 Método iterativo

Desenvolveu-se um método simples, aqui designado por *iterativo* baseado nos valores de verosimilhança obtidos na última iteração de reestimação do modelo de escoamento singular $V_0 = E_1$ (Teixeira e Lindberg, 1992). Neste modelo, utilizou-se todo o material disponível para o treino dos modelos de escoamento, obtendo-se para cada locução o respectivo valor de verosimilhança. Este valor é uma estimativa da probabilidade de observação da respectiva locução dado todo o material de treino disponível para os modelos de escoamento. A primeira variante deste método, considera a soma dos valores correspondentes às locuções da mesma palavra.

Considere-se n_w o número de palavras estranhas diferentes disponíveis para o treino. As locuções das n_w/n_s palavras estranhas, cujas somas apresentem os valores mais altos, são utilizadas no treino do primeiro dos modelos de escoamento finais X_1. As restantes locuções são utilizadas no treino de um outro modelo de escoamento auxiliar V_1. Repetindo o procedimento adoptado com V_0 obtêm-se, sucessivamente, os modelos finais $X_2, X_3, \ldots, X_{n_s}$ e os respectivos modelos auxiliares $V_2, V_3, \ldots, V_{n_s}$, em que $X_{n_s} = V_{n_s-1}$ e V_{n_s} é o conjunto vazio.

Desenvolveu-se outra variante deste método em que as locuções foram consideradas

individualmente, independentemente da palavra que continham. Considere-se agora n_u o número de locuções disponíveis para o treino de modelos de escoamento. Assim, em cada iteração são seleccionadas as n_u/n_s locuções que apresentam os valores de verosimilhança mais altos para serem utilizadas no treino de um modelo de escoamento final.

4.5.2 Método k-médias

O *método k-médias*, também por vezes designado por *IsoData*, é um dos métodos de agrupamento mais conhecidos (Duda e Hart, 1973). Este tipo de métodos pressupõe a existência de distâncias ou de medidas de similaridade, que relacionem entre si cada um dos objectos a agrupar. No caso presente, estes objectos são as locuções de palavras estranhas. Assim, treinaram-se modelos HMM individuais para cada uma das palavras estranhas e obtiveram-se valores de verosimilhança com o algoritmo de Viterbi para todas as locuções disponíveis para o treino de modelos de escoamento. Os valores correspondentes a locuções da mesma palavra x obtidos com um dado modelo de palavra estranha y constituem parcelas da soma $s_{x,y}$. O valor $S(x,y) = S(y,x) = (s_{x,y} + s_{y,x})/2$ é utilizado no método descrito nesta subsecção e na seguinte, como uma medida de similaridade entre as palavras estranhas x e y.

O método k-médias assume a existência de um conjunto de objectos auxiliares A com representação no espaço de características dos objectos (palavras) a agrupar. Cada objecto auxiliar encontra-se associado a um grupo que, neste caso, será o grupo de palavras cujas locuções treinam um único modelo de escoamento. Deverão existir tantos objectos auxiliares quantos os grupos a formar (n_s). Em geral, designam-se estes objectos auxiliares de *centróides* e são determinados com base na média das características dos objectos associados ao respectivo grupo. No caso presente, seria necessário testar esse centróide com todos os modelos correspondentes às palavras estranhas. Para o cálculo da medida de similaridade, tal como foi efectuado com os restantes objectos, seria ainda necessário treinar um novo modelo HMM com apenas esse mesmo centróide, para ser testado com todas as locuções de palavras estranhas. Por forma a eliminar os inconvenientes deste procedimento, optou-se por escolher para objectos auxiliares um dos objectos já existentes. Para representar cada grupo G escolheu-se o objecto z que apresentava um valor máximo para a soma das medidas de similaridade entre si e os restantes objectos do grupo ($\max_{z \in G} \sum_{x \in G} S(x,z)$).

Este método requer uma escolha inicial dos referidos objectos auxiliares e estes por sua vez são determinados a partir da definição dos grupos que se pretendem obter. Além disto, a eficácia do método depende da qualidade desta escolha, pelo que deve ser feita com algum

critério. Escolheram-se primeiramente os dois objectos apresentando a menor medida de similaridade entre si. Os restantes objectos auxiliares são sucessivamente determinados de modo a que apresentem um valor mínimo na soma das medidas de similaridade em relação aos objectos auxiliares já determinados (Teixeira, 1989).

O método pode ser descrito pelos seguintes passos elementares:

1. Efectua-se o agrupamento. Associa-se cada objecto x ao grupo referente ao objecto auxiliar z que apresenta um valor máximo para a medida de similaridade $\max_{z \in A} S(x, z)$.

2. Para cada grupo estabelecido no passo anterior, determina-se o correspondente objecto auxiliar. Obtém-se, assim, um novo conjunto de objectos auxiliares A. O processo termina se este conjunto for igual em duas iterações sucessivas deste passo. De outro modo, regressa-se ao passo 1.

Considere-se o somatório das medidas de similaridade entre cada objecto e o objecto auxiliar do grupo a que se encontra associado. O mecanismo utilizado para a escolha dos objectos iniciais determinou para este somatório um valor[1] (-163747) superior ao obtido com alguns ensaios realizados com escolhas aleatórias dos objectos auxiliares iniciais (-170466 em média). Este resultado é indicativo do interesse desta escolha inicial e do mecanismo de selecção adoptado.

4.5.3 Método do grafo

Considere-se um grafo no qual cada nó representa um objecto a ser agrupado. Neste grafo, uniu-se por um ramo, todos os nós que apresentem entre si um valor de similaridade acima de determinado limiar. Considera-se que pertencem ao mesmo grupo, todos os objectos que partilhem o mesmo caminho ligado por ramos. Utilizando um processo iterativo é, em geral, possível encontrar um limiar que determine o número de caminhos ligados desejados (n_s). Este método foi inicialmente desenvolvido para o agrupamento de segmentos quase-estacionários de tamanho variável de sinal electroencefalográfico (Teixeira, 1989) tendo sido posteriormente adaptado de acordo com os requisitos do presente problema (Teixeira e Lindberg, 1992).

[1] Os valores para a medida de similaridade foram obtidos a partir de logaritmos por forma a facilitar os cálculos.

4.5.4 Resultados

Os métodos anteriormente descritos foram utilizados na determinação do material de treino a utilizar em cada um dos modelos de escoamento. Os resultados encontram-se representados na tabela 4.2. As primeiras duas linhas são resultados já apresentados na tabela 4.1, obtidos com o *método alfabético*. As taxas de reconhecimento e de rejeição indicadas, podem ser comparadas com as obtidas nas experiências anteriores, uma vez que utilizaram o mesmo material de treino e de teste.

método	selecção de	# modelos de escoamento	taxa (%) de reconheci/	taxa (%) de rejeição
alfabético	palavras	3	86,6	60,2
		10	86,3	67,8
iterativo	palavras	3	86,5	60,5
		10	86,0	67,8
	elocuções	3	86,6	60,0
		10	86,3	65,8
k-médias	palavras	10	85,9	62,3
grafo	palavras	10	86,0	56,7

Tabela 4.2: Taxas (%) de reconhecimento e de rejeição obtidas com modelos de escoamento treinados com diferentes combinações de material de fala. Determina-se o material de fala, utilizado no treino de cada modelo de escoamento, de acordo com vários métodos de agrupamento (Teixeira e Lindberg, 1992).

As experiências efectuadas com o *método iterativo* para $n_s = 3$ e $n_s = 10$, determinaram taxas sem diferenças significativas entre as duas variantes consideradas e o mesmo se passa em relação às obtidas com o *método alfabético*. Tal sugere três tipos de considerações. Em primeiro lugar, não parece ser relevante o facto de se dividirem ou não as locuções da mesma palavra em diferentes modelos de escoamento. Em segundo lugar, o tipo de seriação de dados obtido com o *método iterativo* não permitiu gerar qualquer alteração significativa em relação às taxas anteriormente obtidas com o *método alfabético*. Por último, as vantagens obtidas em termos da taxa de rejeição, com dez modelos de escoamento, parecem ser consistentemente superiores às obtidas com apenas três destes modelos. De acordo com esta última consideração, os restantes métodos foram testados

unicamente para a obtenção de conjuntos com dez modelos de escoamento.

Os resultados apresentados do método k-médias e do método do grafo referem-se ambos a taxas de rejeição inferiores às anteriormente obtidas. No caso do método do grafo, registou-se uma quebra significativa (cerca de 16%) da taxa de rejeição. Neste caso, verificou-se ainda a formação de um grupo com a maioria das palavras enquanto que os restantes ficaram apenas com uma única palavra cada. Esta tendência foi também verificada no caso do método k-médias embora de forma menos declarada: determinaram-se dois grupos dividindo entre si a maioria das palavras e a maioria dos grupos restantes ficaram com apenas uma única palavra cada.

Estes resultados parecem indicar que se deverá manter os subconjuntos de treino de cada modelo de escoamento o mais heterogéneo possível em termos das palavras que são utilizadas. Principalmente, deve-se distribuir o material de fala uniformemente, em termos da quantidade de locuções, pelos vários modelos de escoamento.

4.6 O número de modelos de escoamento e a dimensão do vocabulário

Como se referiu na secção 4.4, à data da primeira publicação dos resultados aqui apresentados, a utilização de modelos de escoamento múltiplos era, no mínimo, polémica. Tal ficou a dever-se a experiências anteriores (Wilpon et al., 1989) realizadas em condições muito particulares, nomeadamente com vocabulários muito pequenos (cinco palavras). Assim, é importante determinar qual a relação entre a dimensão do vocabulário e as possíveis vantagens do uso de modelos de escoamento múltiplos no desenvolvimento de novos reconhecedores. As conclusões a surgir deste estudo poderiam de alguma forma colidir com conclusões anteriores de investigadores prestigiados. Por isso, tendo em consideração alguns dos factores de variabilidade do sinal de fala referidos na secção 1.2, procurou-se avaliar o significado estatístico dos resultados aqui apresentados, determinando-se os respectivos intervalos de confiança (secção 2.7).

O novo conjunto de experiências utilizou exclusivamente o material de fala recolhido dos oradores dinamarqueses. Os modelos das 40 palavras-chave foram treinados com duas repetições de 16 oradores enquanto que os modelos de escoamento foram treinados com uma única repetição de 70 palavras diferentes, proferidas por seis oradores. Para teste, utilizaram-se quatro oradores que repetiram duas vezes todas as palavras-chave mais 53 palavras diferentes das usadas no treino. Para a divisão do material de treino dos modelos de escoamento adoptou-se o *método alfabético*, (secção 4.5) uma vez que nenhum

dos outros procedimentos revelou vantagens acrescidas, sendo contudo mais complexos.

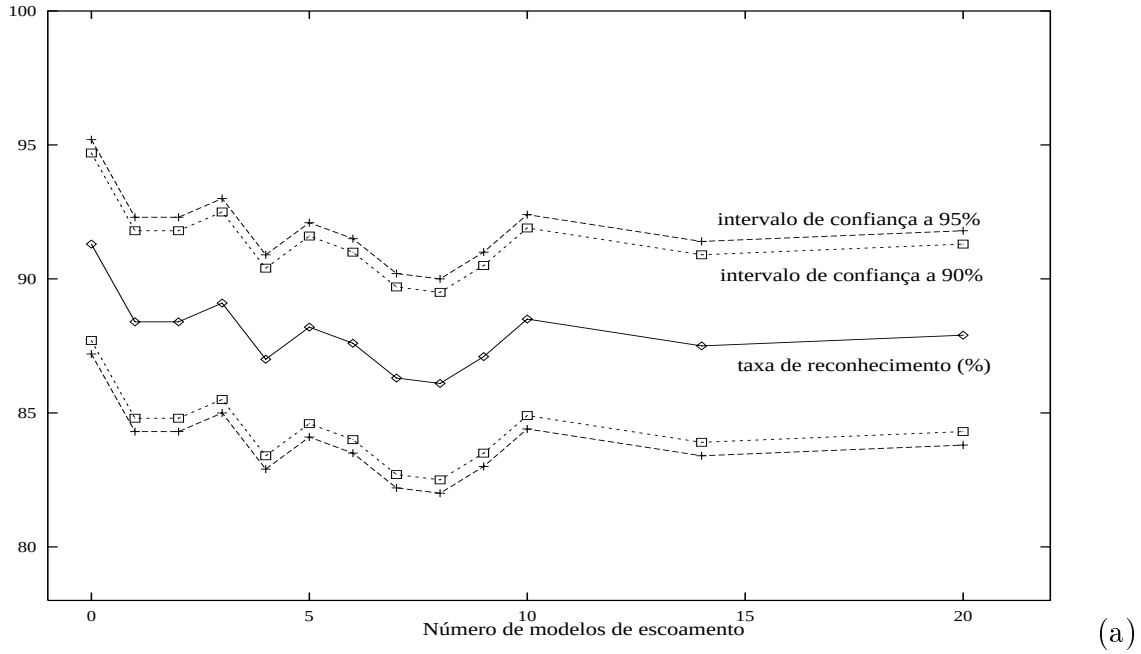
4.6.1 Experiências de aferição

Os resultados obtidos com o vocabulário de 40 palavras encontram-se representados na figura 4.3. Verificou-se, mais uma vez, o decréscimo da taxa de reconhecimento com o aumento do número de modelos de escoamento. Contudo, este decréscimo continua a ser estatisticamente irrelevante, uma vez que não é passível de comprovação em intervalos de confiança inferiores a 90%. Analisando a figura 4.3b, verifica-se um crescimento quase monótono da taxa de rejeição com o número de modelos de escoamento utilizados. Este crescimento deixa de ser significativo com os reconhecedores com mais de, aproximadamente, cinco modelos de escoamento. A vantagem de se utilizarem cinco modelos de escoamento em relação a um único é verificada em intervalos de confiança superiores a 95%.

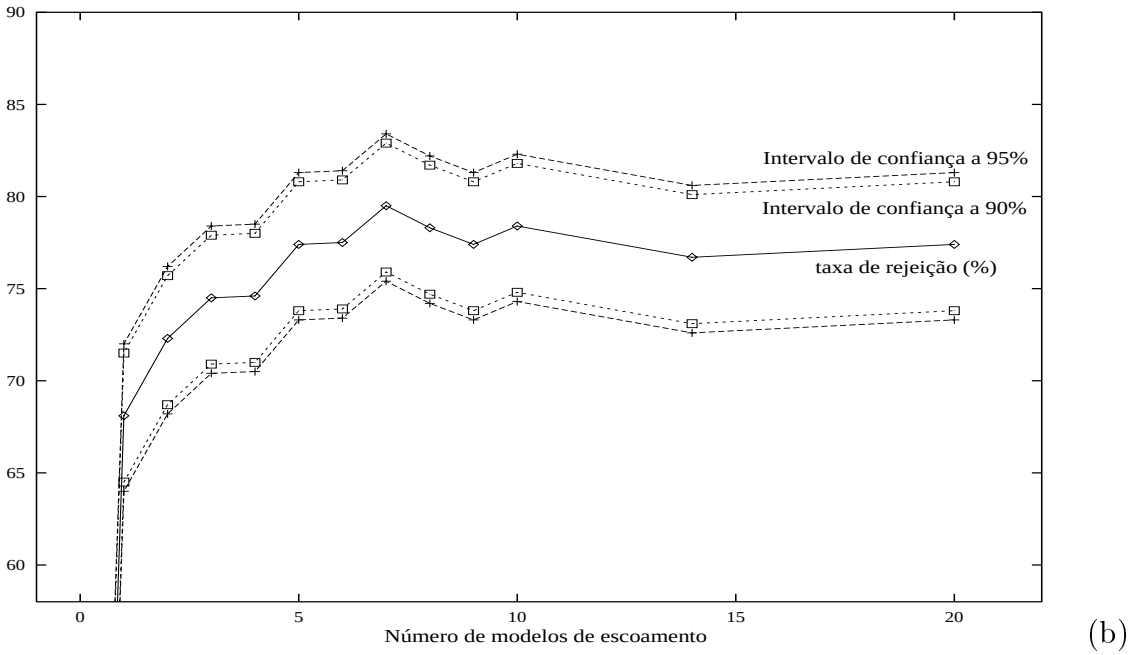
Comparando estes resultados com os representados na tabela 4.2 verifica-se que tanto as taxas de reconhecimento como as de rejeição são agora maiores. Sublinha-se que o conteúdo dos respectivos vocabulários utilizados é diferente e que nas experiências das secções 4.4 e 4.5 se utilizou uma quantidade superior de oradores. Os resultados melhores agora obtidos são, contudo, justificáveis, uma vez que os oradores utilizados no treino apresentam o mesmo sotaque dos que são utilizados nos testes.

4.6.2 Experiências com diversas dimensões de vocabulário

Procura-se agora obter alguma relação entre a dimensão do vocabulário e o número de modelos de escoamento a utilizar. Para tal, utilizou-se o vocabulário anterior (40 palavras) dividindo-o em subvocabulários de 5, 10, 20 e 30 palavras. A obtenção destes subvocabulários foi sistematizada da seguinte forma: o vocabulário inicial foi ordenado numa lista por ordem alfabética; a q-ésima palavra do novo subvocabulário de p palavras ($1 \leq q \leq p$) é a m-ésima palavra desta lista, em que m é a parte inteira do quociente $40q/p$. Com cada um destes novos subvocabulários, construíram-se reconhecedores com diferentes números de modelos de escoamento. De acordo com os resultados obtidos com o vocabulário de 40 palavras, não se justificaria utilizar mais de cinco modelos de escoamento, pelo que não se experimentaram reconhecedores com um número superior destes modelos. Tendo em consideração que o subvocabulário mais pequeno a ser utilizado é de precisamente cinco palavras, também não é razoável, em termos computacionais, que se utilize um número de modelos de escoamento superior ao dos próprios modelos-chave



(a)



(b)

Figura 4.3: Taxa (%) de reconhecimento (a) e de rejeição (b) obtidas com reconhecedores com diferentes números de modelos de escoamento. Nos gráficos representam-se igualmente os intervalos de confiança a 90% e a 95% (Teixeira et al., 1992).

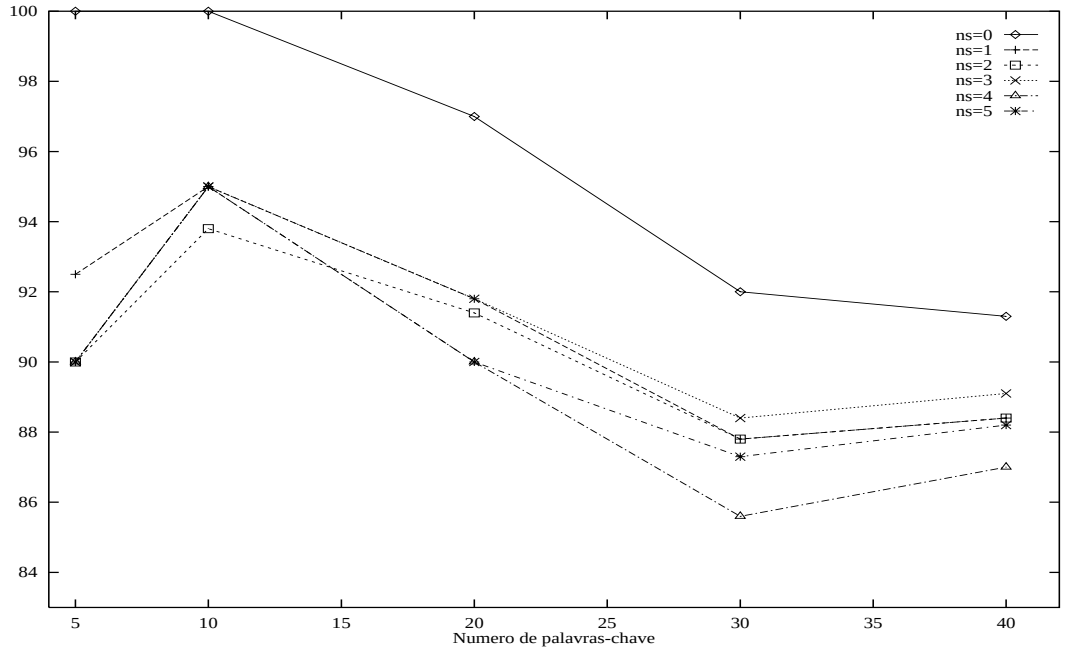
(conforme já foi referido).

Na figura 4.4, apresentam-se os resultados das experiências anteriormente descritas. Cada ponto representa o resultado de uma experiência de reconhecimento com uma dimensão de vocabulário e um número de modelos de escoamento diferentes. Os segmentos de recta unem os pontos das experiências em que se utilizou o mesmo número de modelos de escoamento. Os valores obtidos para as taxas (%) de reconhecimento e de rejeição foram representados em duas figuras diferentes (a) e (b), respectivamente. Em relação à primeira verifica-se, uma vez mais, a tendência de descida da taxa de reconhecimento com o aumento do número de modelos de escoamento. Além disso, os trajectos representados evidenciam uma redução desta quebra com o aumento da dimensão dos vocabulários. Por exemplo, de um reconhecedor com um vocabulário de 5 palavras e sem capacidade de rejeição para um outro com cinco modelos de escoamento, obtém-se uma quebra de 10% na taxa de reconhecimento. Nos reconhecedores com um vocabulário de 40 palavras, a referida quebra é reduzida para cerca de 5%. Esta tendência só foi possível de verificar num intervalo de confiança de 90%, quando se substitui um reconhecedor de 5 palavras por outro de 40 palavras.

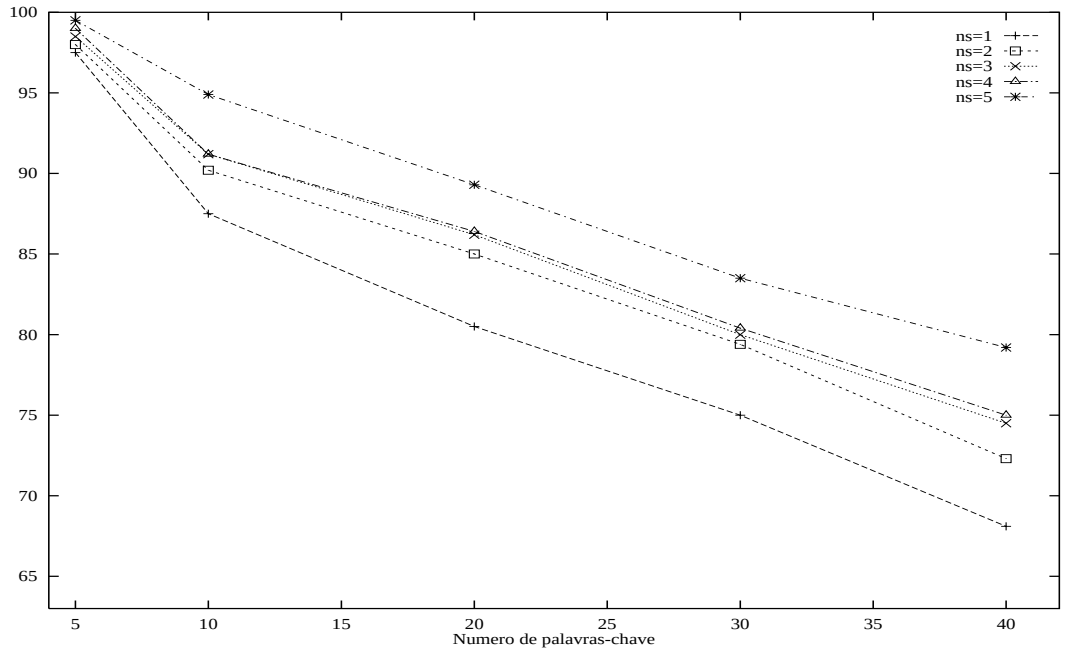
Os resultados obtidos mostram um claro declínio da taxa de rejeição com o aumento da dimensão dos vocabulários, bem como uma tendência (menos evidente) para o afastamento entre os trajectos assinalados na figura 4.4b. Assim, a vantagem do uso de modelos de escoamento múltiplos parece ser maior quanto maior for o número de palavras-chave. Esta tendência é mais evidente nos vocabulários mais pequenos. Por exemplo, de um reconhecedor de 20 palavras com um único modelo de escoamento para um outro com cinco, obtém-se um acréscimo de 10% na taxa de rejeição. Com os reconhecedores de 5 palavras, o referido acréscimo é reduzido para cerca de 2%. Assim, a conclusão essencial das experiências descritas nesta secção, é a confirmação da utilidade do uso de modelos de escoamento múltiplos para o caso dos vocabulários a reconhecer serem de dimensão superior a cerca de uma dezena de palavras.

4.7 Escolha de material de treino e uso de modelos semicontínuos

Na secção 4.5 procurou-se separar o material de fala para o treino dos modelos de escoamento múltiplos, de forma a obterem-se melhores taxas de rejeição de palavras estranhas. Na presente secção, descrevem-se experiências que procuram determinar algumas características genéricas deste material (por exemplo, a sua quantidade) que permitam



(a)



(b)

Figura 4.4: Representação de taxas (%) de reconhecimento (a) e de rejeição (b). Cada ponto representa o resultado de uma experiência de reconhecimento com uma dimensão de vocabulário e um número de modelos de escoamento diferentes. Os segmentos de recta unem os pontos das experiências em que se utilizou o mesmo número de modelos de escoamento (Teixeira et al., 1992).

obter melhorias significativas dessas mesmas taxas. Testou-se também um reconhecedor diferente do utilizado nas secções anteriores, baseado em observações semicontínuas (secção 2.4). Outros autores já haviam considerado o uso deste tipo de modelos na detecção de palavras-chave, contudo, tirando partido de um sistema de extracção de características baseado em redes neuronais artificiais (Clary e Hansen, 1992). Nesta dissertação procura-se determinar eventuais vantagens no uso de modelos de escoamento deste tipo, tendo em consideração as limitações da quantidade de material de fala disponível para o respectivo treino (Teixeira et al., 1993a).

O reconhecedor baseado em modelos semicontínuos permite modelar cada função densidade de probabilidade de observação a partir de uma mistura de gaussianas cujas descrições se associam numa espécie de dicionário de quantificação (secção 2.4). Neste caso, treinou-se um dicionário com 128 gaussianas das quais se escolheram as 12 mais representativas (com pesos c_m mais altos) em cada estado dos HMMs, para construir a respectiva função densidade de probabilidade de observação (subsecção 2.4.2). O referido dicionário resulta do treino conjunto dos modelos-chave, sendo depois utilizado, sem ser de novo reestimado, durante o treino dos modelos de escoamento. No treino dos modelos de escoamento efectua-se, exclusivamente, a reestimação das probabilidades de transição e dos pesos das componentes gaussianas por cada estado.

tipo de HMMs	CHMM			SCHMM		
HMMs de esc.	0	1	5	0	1	5
c40×2	93,7/0,0	91,4/56,1	91,0/73,3	90,8/0,0	90,0/46,4	89,1/60,2
c40×1	93,7/0,0	91,2/55,0	89,6/69,5	90,8/0,0	90,0/47,8	87,9/57,9
c̃40×1	93,7/0,0	92,5/57,4	90,8/70,9	90,8/0,0	90,6/52,8	90,0/62,4
c̃70×1	93,7/0,0	92,5/61,0	91,9/71,2	90,8/0,0	90,8/51,6	90,4/62,3

Tabela 4.3: Taxas (%) de reconhecimento/rejeição obtidas com diferentes selecções de material de fala para o treino de modelos de escoamento (Teixeira et al., 1993a).

Nas experiências descritas nesta secção consideram-se todos os oradores de um único sotaque não nativo (dinamarqueses). O vocabulário a reconhecer é igual ao empregue nas secções 4.4 e 4.5. Utilizaram-se 30% dos oradores disponíveis para o teste dos diferentes reconhecedores, repetindo cada um duas vezes o mesmo conjunto de locuções: 40 palavras-chave e 70 palavras estranhas. Para o treino dos modelos-chave utilizaram-se duas repetições das palavras-chave proferidas por cada um dos restantes oradores (70%). Para maior simplicidade, designou-se este conjunto de treino de *c40×2*. Exclusivamente

para o treino dos modelos de escoamento, seleccionaram-se ainda mais três conjuntos de locuções proferidas uma única vez por dez oradores e designados, por analogia, por:

c40×1 — subconjunto com cerca de um terço das locuções de *c40×2*;

$\tilde{c}$***70×1*** — 70 palavras estranhas, diferentes das utilizadas no teste;

$\tilde{c}$***40×1*** — subconjunto de $\tilde{c}$*70×1* com apenas 40 palavras estranhas.

Estes conjuntos de locuções de treino foram, por sua vez, divididos pelo *método alfabético* (secção 4.4) em 5 subconjuntos de palavras. Este procedimento permitiu treinar, respectivamente, 5 modelos de escoamento com palavras diferentes. Pretende-se desta forma, mais uma vez, verificar as vantagens dos modelos de escoamento múltiplos para cada uma das referidas selecções. A escolha do número de modelos de escoamento foi feita de acordo com os resultados na secção 4.6.

Na tabela 4.3, representam-se os resultados referentes a estas experiências. Os resultados obtidos com os reconhecedores sem capacidade de rejeição (repetidos ao longo das colunas encabeçadas por "0") correspondem apenas a duas experiências: uma para os modelos contínuos (CHMM); outra para os semicontínuos (SCHMM).

O reconhecedor de modelos contínuos apresenta uma taxa de reconhecimento ligeiramente superior ($\approx 3\%$) ao dos modelos semicontínuos. Embora não se verificando, neste caso, condições de escassez de material de treino, nas quais os modelos semicontínuos seriam presumivelmente vantajosos, não é justificável qualquer quebra significativa no desempenho em relação aos modelos contínuos. Contudo, deve ter-se em consideração que se tratam dos primeiros resultados obtidos com uma primeira versão de reconhecedor de modelos semicontínuos. Neste caso não se introduziram quaisquer alterações ou ajustes equivalentes aos anteriormente efectuados no reconhecedor de modelos contínuos. Pelo contrário, existiam várias versões anteriores do reconhecedor de modelos contínuos, a maioria delas já utilizadas com alguma frequência por diversos grupos de trabalho europeus (do consórcio SUNSTAR — secção 3.2). A apreciação dos resultados obtidos com o uso de modelos de escoamento, revela quebras nas taxas de rejeição ainda mais significativas ($\approx 10\%$). Como se referiu, estes modelos não foram treinados em conjunto com o dicionários de gaussianas, o que pode ter contribuído para este facto. Contudo, no caso da primeira linha de resultados, o material utilizado no treino do dicionário é exactamente o mesmo do utilizado no treino dos modelos de escoamento. Por outro lado, com o uso de modelos de escoamento, atenuaram-se as diferenças entre as taxas de reconhecimento dos reconhecedores CHMM e SCHMM. Verificou-se que algumas das palavras que eram

incorrectamente reconhecidas pelos modelos semicontínuos e correctamente reconhecidas pelos modelos contínuos, são agora rejeitadas pelos modelos de escoamento contínuos.

De seguida detalha-se a análise dos resultados da tabela 4.3 em termos do material de treino dos modelos de escoamento. As primeiras duas linhas de resultados (*c40×2* e *c40×1*) referem-se a um tipo de solução particularmente útil para o desenvolvimento de reconhecedores com capacidade de rejeição: utilizou-se o mesmo material de fala seleccionado para o treino dos modelos das palavras-chave. Esta solução facilita a recolha de material de fala para o treino de um reconhecedor, uma vez que bastará considerar o próprio vocabulário da aplicação. Contudo, com a proliferação actual de corpora de fala, esta vantagem tem cada vez menos significado. Pretende-se averiguar da necessidade de treinar os modelos de escoamento com maior variedade lexical. Assim, comparam-se os resultados obtidos com este tipo de solução com os obtidos com uma selecção com palavras não pertencentes ao vocabulário da aplicação (*$\tilde{c}$40×1*) e outra ainda com maior variedade lexical (*$\tilde{c}$70×1*).

O material da selecção *c40×1* contém aproximadamente um terço das locuções da selecção *c40×2*. A redução da quantidade de material de fala determina um decréscimo na taxa de rejeição, quando este material é subdividido para o treino de modelos de escoamento múltiplos. Este decréscimo é aproximadamente igual para os modelos contínuos e semicontínuos.

O número de locuções da selecção *$\tilde{c}$40×1* é aproximadamente igual ao da selecção *c40×1*. Pretende-se com as experiências associadas a esta selecção verificar o impacto do uso de palavras diferentes das palavras-chave no treino de modelos de escoamento. Espera-se obter, no essencial, uma recuperação das taxas de reconhecimento, uma vez que os modelos de escoamento deverão apresentar características mais afastadas das dos modelos-chave, evitando-se deste modo que algumas palavras-chave sejam rejeitadas. De facto, não só as taxas de reconhecimento como também as de rejeição aumentam ligeiramente ($\approx$ 1%). No caso dos modelos semicontínuos verificam-se aumentos de cerca de 5% nas taxas de rejeição.

Por último, com a selecção *$\tilde{c}$70×1*, quase se duplica o número de locuções utilizando exclusivamente palavras diferentes de todas as restantes. O acréscimo de quantidade e variedade lexical do material de treino dos modelos de escoamento, não alterou de forma significativa o desempenho deste reconhecedor em relação ao que utilizou a selecção *$\tilde{c}$40×1*.

As conclusões possíveis de estabelecer a partir deste conjunto de experiências confirmam a superioridade do uso de modelos de escoamento múltiplos, desta vez no caso dos modelos semicontínuos. Além disso, prevê-se a possibilidade de se utilizarem as mes-

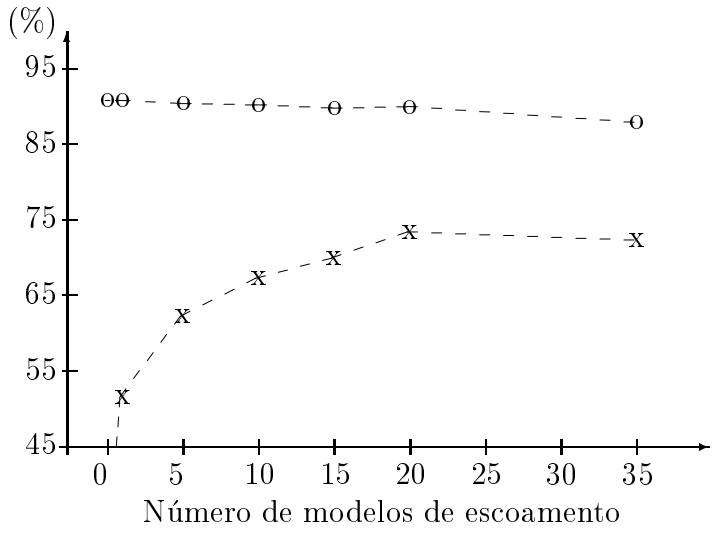


Figura 4.5: Taxas de reconhecimento (o) e de rejeição (x) obtidas de experiências com modelos HMM de observações semicontínuas (Teixeira et al., 1993a).

ma locuções de treino dos modelos-chave para o treino dos modelos de escoamento, sem grande prejuízo das taxas de rejeição. De facto, os melhores resultados de rejeição foram obtidos nestas circunstâncias com modelos de escoamento múltiplos (CHMM). O aumento da variedade lexical deste material de treino não pareceu contribuir para um acréscimo significativo do desempenho destes reconhecedores.

Na sequência destas experiências, pretendeu-se determinar, no caso dos modelos semicontínuos, vantagens eventuais da divisão do material de treino por um número superior de modelos de escoamento. Para tal, adoptou-se a selecção de locuções com mais palavras diferentes ($\tilde{c}70\times1$) e continuou-se a sua divisão, pelo processo anteriormente adoptado, de modo a obterem-se 10, 15, 20 e 35 modelos de escoamento. Os resultados obtidos encontram-se representados no gráfico da figura 4.5. Da observação desta figura conclui-se que, no caso dos modelos semicontínuos, é possível melhorar substancialmente as taxas de rejeição com o uso de um número superior de modelos de escoamento. De facto, com 20 destes modelos foi possível obter uma taxa de rejeição de 73,4% (90,0% de reconhecimento). Com 35 destes modelos obteve-se para a mesma taxa 72,3% (87,9% de reconhecimento) valores que parecem indicar uma zona de saturação ou declínio de desempenho, tal como foi detectado para os modelos contínuos na secção 4.6.

4.8 Influência do sotaque estrangeiro

O problema do reconhecimento automático de fala de oradores estrangeiros é o assunto do capítulo 5. Nesta secção, este problema é abordado de uma forma preliminar, no seguimento das experiências iniciais das secções 4.4, 4.5 e na perspectiva da detecção das palavras-chave. No essencial, procura-se perceber o efeito da inclusão selectiva de material de fala de oradores nativos e não nativos no treino, quer dos modelos-chave, quer dos modelos de escoamento singulares ou múltiplos.

4.8.1 Modelos separados para oradores nativos e não nativos

Os oradores não nativos utilizados nas experiências apresentadas nesta secção são de nacionalidade dinamarquesa e os oradores nativos são ingleses. As palavras-chave e as palavras estranhas utilizadas são as mesmas que as da secção 4.7. Contudo, o número de oradores utilizados no treino de modelos é agora 80% do total em vez dos 70% anteriormente utilizados. Tal facto poderá justificar o aumento agora obtido nas taxas de reconhecimento com os modelos treinados com oradores dinamarqueses ($\approx 3\%$).

Obteve-se para cada sotaque, nativo e não nativo, modelos-chave, modelos de escoamento singulares e múltiplos. De acordo com os resultados da secção 4.6, consideram-se exclusivamente associações de cinco modelos de escoamento.

Na tabela 4.4 apresenta-se os resultados de reconhecimento obtidos com oradores nativos e não nativos, respectivamente. Cada linha de resultados refere-se a um reconhecedor diferente, indicando as taxas (%) obtidas de reconhecimento e de rejeição.

A análise dos resultados obtidos indica uma significativa quebra nas taxas de reconhecimento (em média mais de 15%) quando o sotaque do material de teste difere do de treino. Esta quebra é, contudo, mais acentuada, quando se testam os oradores não nativos com modelos treinados com sotaque nativo do que na situação inversa. A justificação pode estar no facto de os oradores estrangeiros apresentarem maior variabilidade de pronúncia para uma mesma palavra. Quando se treina um modelo para cada palavra-chave essas variações são "filtradas" no cálculo de parâmetros estatísticos. Por outro lado, quando estas locuções são comparadas com modelos-chave treinados com oradores nativos, as referidas variações determinam valores de verosimilhança mais baixos.

Verifica-se também um decréscimo das taxas de rejeição com os oradores estrangeiros quando se utilizam os reconhecedores treinados com oradores nativos. Contudo, comparando as taxas de rejeição obtidas com oradores nativos, verifica-se que os reconhecedores

teste / modelos	nativos (n) reconheci/	rejeição	não nativos (ñ) reconheci/	rejeição
n0	96,9	00,0	79,7	00,0
n1n	96,6	62,3	77,5	59,0
n5n	96,3	68,4	76,6	69,8
ñ0	86,9	00,0	96,9	00,0
ñ1ñ	79,1	71,5	95,9	64,9
ñ5ñ	79,1	76,4	95,3	71,0
n1ñ	96,6	37,0	66,3	83,0
ñ1n	74,7	84,7	95,0	32,8

Tabela 4.4: Resultados (%) de testes de reconhecimento. A notação xNy refere-se a um reconhecedor cujos modelos das palavras-chave foram treinados com locuções de oradores x e utilizou um número N de modelos de escoamento treinados com locuções de oradores y. A letra n refere-se às locuções dos oradores nativos. A letra $\tilde{n}$ refere-se às locuções dos oradores não nativos (Teixeira e Trancoso, 1992).

treinados exclusivamente com material de oradores não nativos apresentam valores significativamente superiores aos obtidos com os modelos nativos. Nos modelos de escoamento pretende-se, em geral, diluir a estrutura fonotáctica das diversas palavras com que foram treinados. A existência de maior variabilidade da pronúncia não nativa pode ser, uma vez mais, a justificação.

De acordo com os resultados das secções anteriores, verificou-se em todas as experiências realizadas um acréscimo nas taxas de rejeição, em geral superior a 5%, com o uso de modelos de escoamento múltiplos, quando comparado com o uso de um único modelo de escoamento.

Pretendeu-se avaliar até que ponto a diferença de sotaque se impõe em relação às diferenças lexicais entre palavras-chave e palavras estranhas. Para tal, realizaram-se também algumas experiências em que os modelos de escoamento foram treinados com um sotaque diferente do utilizado no treino dos modelos-chave. Os resultados destas experiências encontram-se representados nas duas últimas linhas da tabela 4.4. De acordo com o esperado, os reconhecedores com modelos-chave treinados com o mesmo sotaque do de teste, as taxas de reconhecimento mantiveram-se inalteráveis enquanto que as taxas de rejeição sofreram quebras da ordem dos 50%. Ou seja, as palavras estranhas determinaram valo-

res de verosimilhança mais elevados junto dos modelos-chave que foram treinados com o mesmo sotaque. Nos reconhecedores com modelos de escoamento treinados com o mesmo sotaque do de teste, as taxas de reconhecimento descem e as de rejeição atingem valores acima dos 80%. Verificou-se que muitas das locuções de palavras-chave obtêm valores de verosimilhança mais elevados junto dos modelos de escoamento treinados com o mesmo sotaque. Por outro lado, atenuou-se a tendência de algumas palavras estranhas obterem valores de verosimilhança mais elevados junto de modelos-chave, uma vez que neste caso, estes modelos foram treinados com um sotaque diferente.

4.8.2 Modelos treinados com dois sotaques

Para as experiências seguintes treinou-se um único conjunto de modelos de palavra-chave utilizando as locuções de oradores nativos e não nativos. Ensaiaram-se várias combinações deste modelos com os modelos de escoamento anteriormente utilizados e com modelos de escoamento treinados com as locuções de palavras estranhas com ambos os sotaques, nativo e não nativo.

teste / modelos	nativos (n) reconheci/	nativos (n) rejeição	não nativos (ñ) reconheci/	não nativos (ñ) rejeição
m0	96,9	00,0	95,9	00,0
m1n	95,9	52,8	95,0	23,4
m1ñ	96,6	29,3	95,9	57,1
m1n1ñ	95,9	54,5	95,0	60,6
m5n5ñ	95,6	62,7	94,4	70,6
m1m	96,3	48,6	95,3	47,9
m2m	95,9	54,3	95,6	55,2
m5m	95,6	59,7	95,0	62,5
m10m	95,6	64,6	94,4	67,9

Tabela 4.5: Resultados (%) dos testes de reconhecimento. A notação xNy refere-se a um reconhecedor cujos modelos das palavras-chave foram treinados com locuções de oradores x e utilizou um número N de modelos de escoamento treinados com locuções de oradores y. A letra m refere-se à utilização simultânea das locuções dos oradores nativos (n) e dos não nativos ($\tilde{n}$) (Teixeira e Trancoso, 1992).

Os resultados destas experiências encontram-se representados na tabela 4.5. Apresentando uma taxa de reconhecimento ligeiramente inferior (1%) à obtida pelos oradores não nativos com os modelos específicos desse sotaque, (tabela 4.4) este tipo de modelos-chave é uma boa solução para o reconhecimento da fala de oradores com diferentes sotaques. Naturalmente, há que ter em consideração que este reconhecedor utilizou no treino dos respectivos modelos, o dobro do material de fala utilizado no treino dos modelos anteriores. Esta duplicação não é meramente em termos quantitativos, mas também qualitativos, uma vez que existem dois sotaques diferentes. Este facto representa também um custo acrescido em termos de recolha de material de fala.

Ensaiaram-se os modelos de escoamento singulares anteriormente determinados para cada sotaque, individualmente e em conjunto (segunda, terceira e quarta linha de resultados). A associação destes modelos em paralelo produziu os melhores resultados de rejeição na média dos dois sotaques testados. Contudo, verifica-se uma quebra apreciável quando comparados com as taxas obtidas com os modelos específicos de cada sotaque (tabela 4.4). Na sequência destes resultados testou-se um reconhecedor associando os modelos de escoamento múltiplos anteriormente utilizados. Registou-se um aumento de quase 10% na taxa de rejeição, a qual continua, ainda assim, inferior aos valores referidos na tabela 4.4. Neste contexto é necessário ter em consideração, não só a duplicação do material de treino, já referida a propósito dos modelos-chave, como também a duplicação do número total de modelos de escoamento utilizados. Conforme se viu nas secções anteriores, o aumento deste número, dentro de certos limites, corresponde a um aumento da taxa de rejeição. O facto de se dispor de mais material de treino permite alargar estes limites, uma vez que se conseguem treinar adequadamente um número maior de modelos de escoamento.

Apresentaram-se até agora duas estratégias diferentes para utilizar todo o material de treino correspondente aos dois sotaques disponíveis: treino de modelos com todo o material de treino (utilizada para os modelos-chave); associação dos modelos treinados com o material de cada sotaque (utilizada para os modelos de escoamento). A análise desta segunda estratégia, aplicada aos modelos de palavra-chave, é efectuada no capítulo 6. De seguida, comentam-se os resultados obtidos com a aplicação da primeira estratégia aos modelos de escoamento.

Considerem-se as locuções de palavras estranhas pronunciadas por oradores nativos e não nativos, destinadas ao treino de modelos de escoamento. Este material é utilizado no treino de um, dois, cinco e dez modelos de escoamento, que correspondem às quatro últimas linhas de resultados da tabela 4.5. Para a determinação do material a utilizar no treino de cada modelo de escoamento múltiplo utilizou-se o *método alfabético* (secção 4.4).

Com o modelo de escoamento singular, obtém-se uma taxa de rejeição inferior às obtidas com a estratégia de associação de modelos. Contudo, deve-se comparar reconhecedores com o mesmo número de modelos de escoamento. Nestas circunstâncias, obteve-se para os reconhecedores com dois modelos de escoamento, uma taxa de rejeição média ligeiramente inferior. Nos reconhecedores com dez modelos de escoamento, esta diferença é ainda menor.

As duas estratégias adoptadas para o treino de modelos de escoamento não revelaram diferenças significativas nos resultados obtidos. Contudo, a associação de modelos é neste caso mais vantajosa, uma vez que permite a adaptação incremental de um reconhecedor a novos sotaques.

4.9 Reconhecimento de fala ligada

Nesta secção, descrevem-se algumas experiências com fala contínua e reconhecedores de fala ligada que integram capacidades equivalentes à da detecção de palavras-chave. Pretende-se melhorar a taxa de reconhecimento de frases completas que incluem palavras estranhas. Isto é, existe um conjunto pré-determinado de frases a reconhecer, definido no universo de uma aplicação. Os utentes desta aplicação pronunciam algumas destas frases inserindo inadvertidamente palavras adicionais, algures, na estrutura sintáctica. As experiências realizadas utilizaram, tal como anteriormente, modelos de palavras isoladas e modelos de escoamento. Estes modelos foram obtidos com os procedimentos anteriormente utilizados, de acordo com um plano para o desenvolvimento de diversas aplicações (Teixeira, 1992a) de reconhecimento de fala. Para a realização dos testes, foi necessário o desenvolvimento de um reconhecedor diferente, que permitisse o encadeamento de vários modelos elementares de modo a modelar a estrutura frásica. Este reconhecedor, designado por SUNCAR (Andersen et al., 1992), permite em alternativa, o uso de modelos subpalavra e pode portanto efectuar reconhecimento de fala contínua.

Os resultados produzidos pelo reconhecimento de fala ligada exigem parâmetros de avaliação de resultados diferentes dos anteriormente utilizados. Com um reconhecedor de palavras ligadas existe a possibilidade de uma palavra ter sido proferida e de não lhe ser atribuído, no processo de reconhecimento, qualquer classificação. É importante contabilizar separadamente estas ocorrências, que são indicadas nas tabelas seguintes como taxas de supressão medidas em percentagem do número total de palavras ou frases, consoante o caso. Recorde-se que a taxa de supressão de palavras definida na secção 2.7 tem o mesmo significado, embora as supressões num reconhecedor de palavras isoladas

sejam em geral exclusivamente devidos à ineficácia do detector de início e fim de palavra. Como é habitual nos casos em que são determinadas as supressões de palavras, determina-se simultaneamente a taxa de exactidão.

4.9.1 Corpus de fala com frases

Para realizar as experiências propostas no início desta secção é possível utilizar um corpus de palavras isoladas através da concatenação destas palavras. Esta solução artificial pode, contudo, conduzir a resultados com pouco significado para futuras aplicações práticas. Por forma a obter-se um corpus de fala adequado, desenvolveu-se uma aplicação que permitisse a demonstração de soluções para este tipo de problema.

A aplicação utilizada nesta secção foi desenvolvida no laboratório do "Center for PersonKommunication" da Universidade de Aalborg, com a designação de SAMOGO ("Speech Activated Manipulation Of Graphical Objects") (Christensen e Lindberg, 1992). A língua adoptada foi o dinamarquês embora não tenham sido considerados procedimentos específicos para esta língua na concepção da aplicação ou do respectivo corpus. A versão de 1991 inclui um vocabulário de 84 palavras de acordo com um modelo linguístico baseado numa gramática com número finito de estados (secção 2.6). Este modelo apresenta um factor de ramificação médio de 3,8 com um desvio padrão de 3,3 e uma perplexidade total de 8,50. O subcorpus de teste apresenta um valor de perplexidade de 6,64.[2]

O corpus de sinais de fala utilizado inclui locuções de:

palavras-chave — vocabulário recolhido em palavras isoladas;

palavras estranhas — palavras isoladas à escolha do locutor, em geral diferentes das palavras-chave e sem repetições;

frases correctas — frases de acordo com o modelo linguístico pré-estabelecido;

frases incorrectas — frases obtidas das anteriores mas com a inserção por parte do locutor de palavras adicionais, eventualmente não pertencendo ao vocabulário, que tornem a frase mais natural;

frases estranhas — locuções de frases à escolha do locutor, diferentes das frases correctas ou incorrectas e sem repetições.

[2]Dados gentilmente fornecidos por Tom Brøendsted, investigador de linguística computacional do CPK.

O subcorpus de frases incorrectas, denominado "SAMOGO Pseudo Natural Language" (PNL) foi recolhido em condições relativamente controladas. Apresentava-se uma frase correcta ao orador, ao qual era pedido que a pronunciasse da forma mais natural possível e inserindo-lhe palavras estranhas. Desde que se disponha de oradores treinados, torna-se fácil a obtenção da respectiva etiquetagem ortográfica. Tal não acontece com os sinais de fala recolhidos sem restrições, num ambiente real. Por estes motivos, a recolha do corpus foi restringida a investigadores e técnicos do CPK (exclusivamente do sexo masculino).

4.9.2 Métodos adoptados

Os métodos de reconhecimento adoptados nas experiências desta secção, foram exclusivamente baseados em modelos de palavras isoladas e em modelos de escoamento semelhantes aos utilizados no restante capítulo. A diferença mais evidente foi a possibilidade de controlar o uso destes modelos através da gramática com número finito de estados. No que se refere ao treino dos modelos de escoamento, ensaiou-se o treino com locuções de frases inteiras em vez de palavras.

Treino de modelos

Os modelos das palavras-chave foram treinados com as respectivas locuções em palavras isoladas. Utilizaram-se três repetições de cada palavra obtidas de cada um dos 14 oradores seleccionados para o efeito. Estes modelos apresentam as mesmas características de outros modelos HMM descritos em secções anteriores, nomeadamente, uma topologia linear de dez estados. Posteriormente revelou-se vantajoso modelar doze das palavras-chave, que são monossilábicas, com HMMs de apenas cinco estados.

Cada um dos 14 oradores pronunciou ainda 3 palavras e 3 frases quaisquer, sem restrições, pelo que não existe uma descrição ortográfica desta parte do corpus. Com este material treinaram-se três tipos de modelos de escoamento, todos eles com uma topologia linear de dez estados:

p.e. — com palavras estranhas;

f.e. — com frases estranhas;

$p.e. \wedge f.e.$ — com palavras e frases estranhas.

As frases recolhidas foram segmentadas com o mesmo algoritmo utilizado para palavras

isoladas (capítulo 2). Os silêncios ou segmentos de sinal sem fala, obtidos entre palavras, foram seleccionados para o treino de um modelo de silêncio HMM linear com cinco estados.

Para os testes, seleccionaram-se 8 oradores masculinos distintos dos utilizados para o treino. Cada um proferiu 149 frases correctas e 20 frases incorrectas.

Modelo linguístico

Com o objectivo de poder rejeitar de forma explícita palavras ou pequenas frases estranhas (fala estranha) inseridas em frases correctas, desenvolveram-se três extensões do modelo linguístico desenvolvido no decorrer da especificação da aplicação e do corpus SAMOGO (este modelo é aqui designado por $gram0$):

$gram1$ — Prevê uma das situações mais prováveis que é a da ocorrência de fala estranha antes ou depois da produção de uma frase correcta. Ou seja, prevê o surgimento de fala estranha onde o modelo $gram0$ prevê a possibilidade de ocorrência de silêncios de pequena duração. Portanto, associa-se em paralelo com o modelo de silêncio, um ou mais modelos de escoamento;

$gram2$ — Prevê a ocorrência de fala estranha entre todos os submodelos linguísticos, mas não no interior destas. Por exemplo, não considera fala estranha entre uma preposição ou artigo e o objecto, nem entre o objecto e o respectivo adjectivo. Uma vez que também prevê a ocorrência de fala estranha no início e no fim das frases, pode ser considerada uma extensão de $gram1$;

$gram3$ — É uma extensão do modelo $gram2$ (e portanto, da $gram1$ e da $gram0$) que prevê a ocorrência de fala estranha entre quaisquer duas palavras de uma frase correcta. Ou seja, antes e depois de qualquer modelo de palavra-chave existe um caminho alternativo com um ou mais modelos de escoamento.

4.9.3 Testes com frases correctas

Nas primeiras experiências com fala ligada utilizaram-se todos os modelos HMM com dez estados, à excepção do já referido modelo para o silêncio, que dispõe de apenas cinco estados. Realizaram-se testes com o subcorpus com frases correctas cujos resultados são apresentados na tabela 4.6. O modelo de escoamento utilizado com os modelos $gram1$, $gram2$ e $gram3$ foi treinado simultaneamente com palavras e frases estranhas ($p.e. \wedge f.e.$).

modelo	rec. de frases		rec. de palavras		
linguístico	corr.	apag.	corr.	apag.	exact.
gram0	70,0	8,4	82,9	4,2	82,4
gram1	66,0	17,0	74,4	10,9	73,9
gram2	58,8	25,2	67,6	16,7	67,1
gram3	58,6	25,3	67,5	16,8	67,1

Tabela 4.6: Resultados (%) obtidos com frases correctas (rec.= reconhecimento, corr.= correctas, supr.= supressões, exact.= exactidão) com modelos de palavras-chave HMM com dez estados

A utilização de modelos de escoamento introduziu quebras apreciáveis no desempenho do reconhecedor. As quebras mais notáveis ocorreram nos modelos $gram2$ e $gram3$ em que se prevê o surgimento de fala estranha no meio da frase correcta. Tal como se verificou com o reconhecimento de palavras isoladas, a utilização dos modelos de escoamento representa uma solução de compromisso que depende da quantidade de fala estranha existente.

Os resultados de reconhecimento obtidos na melhor situação, isto é, sem modelos de escoamento e com frases correctas, são já de si, mais baixos do que seria de esperar. Verificou-se que os resultados obtidos foram no geral penalizados pela existência de doze palavras-chave monossilábicas que são pronunciadas pelos oradores dinamarqueses num intervalo de tempo muito reduzido. Assim, o número de tramas existentes em cada locução não será suficiente para treinar adequadamente dez estados de um HMM. Numa tentativa de melhorar o desempenho global dos reconhecedores reduziu-se este número de estados para metade (cinco).

modelo	rec. de frases		rec. de palavras		
linguístico	corr.	apag.	corr.	apag.	exact.
gram0	75,9	4,1	87,8	2,0	87,3
gram1	73,7	8,6	84,3	5,4	83,9
gram2	69,4	14,3	79,9	9,0	79,6
gram3	69,1	14,6	79,9	9,1	79,6

Tabela 4.7: Resultados (%) obtidos com frases correctas (rec.= reconhecimento, corr.= correctas, supr.= supressões, exact.= exactidão) com modelos de palavras-chave com diferente número de estados. (Teixeira et al., 1992).

As taxas obtidas com os novos modelos de palavra-chave encontram-se representadas na tabela 4.7. Comparando com os resultados anteriores, (tabela 4.6) verifica-se um acréscimo significativo das taxas de reconhecimento, em particular nos reconhecedores com modelos de escoamento, os quais registaram aumentos destas taxas, superiores a 10%. Em relação aos reconhecedores sem capacidade de rejeição de fala estranha, o decréscimo das taxas de reconhecimento foi reduzido a menos de metade. Estes resultados parecem indicar a possibilidade de se obterem melhorias significativas nas taxas de reconhecimento através de um dimensionamento criterioso do número de estados dos modelos HMM utilizados. Além disso, nas condições em que os oradores são treinados para produzir apenas frases correctas, a introdução da capacidade de rejeitar fala estranha poderá ocorrer de forma a manter o desempenho do reconhecedor ao melhor nível possível.

4.9.4 Testes com frases incorrectas

As experiências seguintes testam os reconhecedores apresentados na secção anterior com as frases incorrectas do subcorpus SAMOGO PNL. Em alternativa aos modelos de escoamento $p.e. \wedge f.e.$, testaram-se reconhecedores com os modelos $p.e.$, $f.e.$ e com a associação destes dois modelos. Os respectivos resultados encontram-se representados na tabela 4.8.

As frases incorrectas obtêm junto do reconhecedor sem modelos de escoamento ($gram0$) uma taxa de reconhecimento de cerca de 50% da taxa obtida com frases correctas (tabela 4.7). Esta quebra acentuada deve-se essencialmente a um aumento de substituições das frases e das palavras, embora também se verifique um aumento do número de supressões.

Analisando os resultados obtidos com o modelo $gram1$, verifica-se que apresentam as taxas de reconhecimento mais baixas, muito inferiores às obtidas com a $gram0$, inclusive. Em contrapartida, os modelos $gram2$ e $gram3$, que na tabela 4.7 apresentavam os valores mais baixos nas taxas de reconhecimento, surgem agora com os valores mais altos. De facto, a maioria dos oradores utilizados nas gravações da SAMOGO PNL, seguiram textualmente as instruções iniciais de inserir fala estranha "no meio das frases correctas".

A alteração do modelo de escoamento utilizado em cada um dos reconhecedores não evidenciou alterações apreciáveis nos resultados, embora o modelo de escoamento treinado simultaneamente com palavras e frases estranhas ($p.e. \wedge f.e.$) tenha apresentado o melhor desempenho com frases incorrectas. Em relação ao uso isolado dos modelos $p.e.$ e $f.e.$, apresenta a vantagem de acumular o material de treino desses dois modelos, pelo que deverá ser mais representativo da fala estranha.

modelo de escoamento	modelo linguístico	rec. de frases		rec. de palavras		
		corr.	apag.	corr.	apag.	exact.
–	*gram0*	37,5	11,2	65,1	3,6	61,3
p.e.	*gram1*	26,9	33,1	56,5	14,9	55,6
	gram2	51,9	25,0	69,8	10,1	69,2
	gram3	54,4	18,1	74,8	7,5	74,0
f.e.	*gram1*	30,0	26,9	58,9	11,7	57,7
	gram2	51,2	22,5	69,8	9,1	69,0
	gram3	55,6	14,4	76,6	6,0	75,8
p.e. $\wedge$ *f.e.*	*gram1*	28,8	31,9	57,9	14,3	56,5
	gram2	54,4	23,1	70,8	9,5	70,2
	gram3	56,2	15,6	75,8	6,5	75,2
p.e. + *f.e.*	*gram1*	28,1	33,8	55,8	15,5	55,8
	gram2	51,9	23,8	69,6	10,1	69,0
	gram3	55,0	16,2	76,0	6,9	75,2

Tabela 4.8: Resultados (%) das frases com palavras estranhas (rec.= reconhecimento, corr.= correctas, supr.= supressões, exact.= exactidão). *p.e.* = palavras estranhas, *f.e.* = frases estranhas (Teixeira et al., 1992).

O uso conjunto dos modelos *p.e.* e *f.e.* não revelou qualquer vantagem em relação às situações anteriores (três últimas linhas da tabela 4.8). O uso de dois modelos de escoamento em simultâneo, de acordo com os resultados obtidos ao longo deste capítulo com os modelos de escoamento múltiplos, deveria permitir obter algum ganho no desempenho destes reconhecedores. O facto pode ser explicado se se tiver em consideração o factor de ramificação estimado para o modelo linguístico da presente aplicação e os resultados da secção 4.6. Assim, uma vez que na maior parte dos casos o presente modelo reduz o vocabulário activo em cada instante a menos de seis palavras, não se deveria esperar qualquer vantagem com o uso de modelos de escoamento múltiplos (figura 4.4). De acordo com este raciocínio e tomando a solução representada por $gram3$ para o problema das frases incorrectas, a activação de determinado subvocabulário poderia ser associada à escolha de um número adequado de modelos de escoamento. Na realidade, os modelos de escoamento não funcionam neste contexto como alternativa aos modelos-chave, pelo que não são de esperar vantagens evidentes com os modelos de escoamento múltiplos mesmo em aplicações com perplexidades superiores.

4.9.5 Taxas de reconhecimento globais

Considere-se a taxa de reconhecimento global de frases dada por

$$R_g(f_i, r_c, r_i) = r_c.(1 - f_i) + r_i.f_i,$$

em que r_c e r_i representam, respectivamente, as taxas de reconhecimento de frases correctas e incorrectas. f_i é a fracção de frases incorrectas.

Os resultados obtidos nesta secção indicam que a utilização de modelos de escoamento pode ser útil no reconhecimento de palavras ligadas, se a fracção de frases incorrectas for superior a um determinado valor α. Uma vez que no caso do reconhecedor convencional $gram0$ se tem r_c=75,9% e r_i=37,5%, este valor pode ser calculado para cada um dos reconhecedores a partir da igualdade

$$R_g(\alpha, r_c, r_i) = 75,9.(1 - \alpha) + 37,5.\alpha,$$

ou, de forma equivalente

$$\alpha = \frac{75,9 - r_c}{38,4 + r_i - r_c}.$$

Os resultados obtidos para os reconhecedores que utilizaram o modelo de escoamento $p.e. \wedge f.e.$ encontram-se na tabela 4.9.

taxa (%) de reconheci/	$gram0$	$gram1$	$gram2$	$gram3$
de frases correctas	75,9	73,7	69,4	69,1
de frases incorrectas	37,5	28,8	54,4	56,2
α (%)	0,0	$< 0,0$	27,8	26,7

Tabela 4.9: Percentagem de frases incorrectas (α) necessárias para se obter um desempenho equivalente ao do reconhecedor convencional ($gram0$).

Deste modo, pode concluir-se que o uso de modelos de escoamento é justificado com a obtenção de taxas de reconhecimento mais elevadas, sempre que mais de cerca de 30% das frases apresentadas para reconhecimento sejam frases incorrectas. A classificação de frases correctas ou incorrectas é, contudo, uma forma demasiado grosseira de definir a quantidade de fala estranha. É preferível dispor-se de uma contagem de palavras estranhas em cada frase, ou mesmo da duração da fala estranha, por forma a obterem-se medidas semelhantes às mencionadas na secção 4.3.

4.10 Conclusões

Neste capítulo apresentaram-se várias experiências de utilização de modelos de escoamento para a detecção de palavras-chave. Estas experiências permitiram verificar a vantagem da utilização de modelos de escoamento múltiplos no reconhecimento de vocabulários de média dimensão. Na sequência desta verificação, determinaram-se algumas condições para o treino destes modelos de modo a se obterem taxas de rejeição mais elevadas.

Descreveram-se experiências no contexto do reconhecimento de fala não nativa. Assim, verificou-se a possibilidade de se adaptarem de forma incremental para novos sotaques, os reconhecedores com capacidade de detecção de palavras-chave.

Estudou-se o reconhecimento de fala ligada quando existem palavras estranhas, utilizando-se gramáticas com um número finito de estados. O uso de modelos de escoamento só é justificável em algumas situações, as quais são determinadas de forma aproximada. Neste tipo de reconhecimento, concluiu-se que o uso dos modelos de escoamento múltiplos dificilmente poderá apresentar vantagens evidentes.

Capítulo 5

Reconhecimento automático da fala de oradores estrangeiros

5.1 Introdução

Um dos objectivos principais deste trabalho é o de reduzir as quebras de desempenho verificadas no reconhecimento automático de fala de oradores estrangeiros. Os reconhecedores de fala são habitualmente treinados para um vocabulário numa determinada língua, utilizando-se exclusivamente locuções de oradores nativos. As taxas de reconhecimento obtidas no teste destes reconhecedores com oradores não nativos são significativamente inferiores às que se obtêm com oradores nativos.

Um estudo de 1989 (Barry et al., 1989) aborda um problema de sotaques regionais no âmbito do reconhecimento de fala analisando, nomeadamente, os diferentes sistemas de vogais entre oradores do norte e do sul do Reino Unido, da Escócia e da América do Norte. O estudo dos sotaques regionais distingue-se em alguns aspectos essenciais do estudo dos sotaques estrangeiros, sendo este último geralmente mais complexo. Os oradores estrangeiros apresentam diferentes níveis de competência de leitura e de pronúncia (Mengel, 1993), isto é, o conhecimento das conversões grafema-fonema da língua estrangeira apresenta grandes variações de orador para orador, assim como a capacidade de pronunciar sons que não fazem parte do inventário de sons nativo.

O problema da variabilidade interorador representa o quadro geral onde se pode classificar o problema dos sotaques estrangeiros. Contudo, este problema revela-se mais complexo, uma vez que um reconhecedor treinado com oradores com o mesmo sotaque estrangeiro dos oradores de teste, apresenta igualmente um decréscimo significativo na taxa de reconhecimento. As soluções típicas para o problema da variabilidade interorador não

dispensam um corpus de fala representativo dessa mesma variabilidade. A recolha de um corpus específico para cada sotaque estrangeiro não é, em geral, uma tarefa exequível, pelo que se pretende, também neste trabalho, procurar formas de a evitar ou de a reduzir ao mínimo indispensável.

As experiências descritas neste capítulo utilizaram todo o corpus SUNSTAR multissotaque (descrito na secção 3.2). Recorde-se que este corpus resulta de gravações de um vocabulário constituído exclusivamente por palavras inglesas. Para além de oradores nativos britânicos (*en*), encontram-se representados cinco sotaques estrangeiros do inglês: dinamarquês (*da*), alemão (*de*), espanhol (*es*), italiano (*it*) e português (*pt*). Em representação de cada sotaque dispõem-se de aproximadamente 20 oradores (10 homens e 10 mulheres). Cada orador repete duas vezes um vocabulário com cerca de 200 palavras diferentes. Utilizaram-se corpus separados de treino e de teste para cada um dos sexos. Os oradores de cada um destes corpus foram ainda divididos por forma a serem obtidos testes independentes do orador. Assim, 60% dos oradores foram seleccionados para o conjunto de treino dos modelos, enquanto que os restantes 40% foram utilizados no teste. A selecção do vocabulário e dos oradores para o treino e para o teste dos diversos reconhecedores utilizados, é diferente da selecção utilizada no capítulo 4. Estas diferenças ficam a dever-se à necessidade de se utilizarem condições equivalentes nas experiências com unidades subpalavra, para efeitos comparativos. Contudo, a dimensão do vocabulário e o tipo de observações estimadas é o mesmo. Assim, apenas 25% do vocabulário disponível foi utilizado no treino e no teste dos modelos de palavras inteiras. Este subvocabulário foi igualmente utilizado nos testes de reconhecimento com modelos subpalavra. Uma vez que não se utilizou qualquer modelo linguístico, a selecção usada para os testes determina uma perplexidade relativamente alta. O restante vocabulário (75%) foi utilizado no treino dos modelos subpalavra, por forma a garantir a independência do vocabulário de teste (este foi o principal argumento para a utilização destes modelos).

Sublinha-se o facto de todos os sinais de fala terem sido filtrados numa simulação de um canal telefónico convencional e segmentados em palavras isoladas. Os sinais resultantes foram pré-processados, por forma a serem obtidos os respectivos coeficientes cepstrais (Teixeira, 1992b). Durante o processamento dos algoritmos de reconhecimento foram determinados coeficientes de delta-cepstrum (secção 2.2).

O reconhecedor empregue nas experiências seguidamente descritas é consideravelmente mais avançado do que o utilizado no capítulo 4. Este reconhecedor dispõe de um processo de inicialização mais aperfeiçoado e permite o processamento de uma mistura de múltiplas componentes gaussianas no modelamento da função densidade da probabilidade de observação. Além disso dispõe do algoritmo de *token passing* (subsecção 2.5.6) e de

outras facilidades para o processamento de fala contínua, tais como o uso de léxicos e de modelos linguísticos (secções 2.5 e 2.6, respectivamente). Os modelos de palavra aqui utilizados obedecem a uma topologia linear de 10 estados.

De seguida descreve-se resumidamente o conteúdo do presente capítulo.

Na secção 5.2 são apresentadas as experiências efectuadas com modelos de palavra. Os resultados obtidos servem de referência para as experiências com modelos subpalavra descritas nas secções posteriores.

Na secção seguinte (5.3) apresentam-se as experiências efectuadas com modelos subpalavra dependentes e independentes do sotaque e discute-se a possibilidade de utilização dos designados *polifones*.

Na secção 5.4 descrevem-se alguns métodos, propostos por outros autores, para a determinação automática de transcrições fonéticas largas. Pretende-se utilizar métodos deste tipo para determinar elementos distintivos dos sotaques estudados ao nível da transcrição fonética.

Na secção 5.5 é proposto um método que permite obter uma rede probabilística de transcrições fonotípicas para cada palavra. Apresentam-se e analisam-se resultados de reconhecimento obtidos com estas redes de transcrições.

Por último, na secção 5.6 conclui-se quanto à utilidade e aplicabilidade dos métodos testados neste capítulo e do seu uso na resolução do problema dos oradores estrangeiros.

5.2 Reconhecimento com modelos de palavra

As primeiras experiências realizadas com oradores estrangeiros testaram exclusivamente modelos de palavra (Teixeira e Trancoso, 1992). Os resultados apresentados na secção 4.8 evidenciaram uma quebra de cerca de 15% na taxa de reconhecimento quando os oradores testados eram não nativos (dinamarqueses) em relação a experiências equivalentes com oradores nativos.

Na mesma altura Brousseau e Fox (Brousseau e Fox, 1992) apresentaram uma comunicação que indicava resultados semelhantes obtidos em experiências com dialectos. Estudaram os dialectos europeu e americano (canadiano) do francês e as semelhanças com os dialectos homólogos do inglês. O reconhecedor utilizado foi o do laboratório destes investigadores, o DragonDictate^{TM} da Dragon Systems Inc. (Baker, 1975). Os testes foram realizados com apenas três oradores por dialecto, mas com vocabulários de duas

mil palavras num caso e de dezoito mil palavras no outro. Foram utilizados modelos linguísticos baseados em *N-gramas* (secção 2.6).

Por forma a verificar os resultados obtidos na secção 4.8 e de modo a estabelecer uma referência para as experiências posteriores com modelos subpalavra, efectuaram-se testes com os restantes sotaques disponíveis no corpus SUNSTAR multissotaque.

5.2.1 Modelos de palavra para cada sotaque

O primeiro grupo de experiências usa modelos de palavras isoladas treinados exclusivamente com oradores nativos, (britânicos) separadamente para o sexo feminino e masculino. Estes modelos foram testados com os vários grupos de oradores nativos e estrangeiros do inglês. Os resultados encontram-se representados nas figuras 5.1a e 5.2a, respectivamente.

Para o segundo grupo de experiências treinou-se um conjunto de modelos para cada grupo de oradores estrangeiros correspondente a determinado sotaque. Cada conjunto destes modelos foi testado com um grupo de oradores diferentes, mas com o mesmo sexo e sotaque. Os resultados encontram-se representados nas figuras 5.1b e 5.2b, respectivamente.

As figuras 5.1 e 5.2 representam o resultado de cada experiência de reconhecimento num ponto de um plano bidimensional cujas coordenadas são: a taxa de reconhecimento (%) e o número de componentes gaussianas utilizadas para o modelamento da função densidade de probabilidade de observação. Os pontos unidos pelo mesmo trajecto resultam do teste de um conjunto de modelos, obtidos em sucessivas etapas de treino. Em cada trajecto, os sinais de fala utilizados no treino e no teste são sempre os mesmos[1]. A linha etiquetada de *média* corresponde à média aritmética obtida sobre todos os sotaques.

Seguidamente, descreve-se em termos gerais o processo de treino dos modelos. Os modelos iniciais foram construídos a partir de uma segmentação fixa, seguida de um alinhamento de Viterbi (subsecção 2.3.5). Posteriormente, reestimam-se os respectivos parâmetros até se verificar um critério de convergência num ciclo com um número máximo de 10 iterações (quando não se determina um máximo local, assume-se que ao fim deste número de iterações se dispõe de uma solução aceitável). A estas reestimações soma-se um ciclo de mais quatro reestimações *embutidas*. Os modelos obtidos descrevem a função densidade de probabilidade das observações para cada estado com apenas uma única componente gaussiana. Os resultados obtidos com os testes destes modelos encontram-

[1] Diferindo os do treino dos de teste, uma vez que provêm de oradores diferentes e por forma a garantir testes independentes do orador

se representados no valor mais baixo (1) nos eixos das abcissas das figuras 5.1 e 5.2. Utiliza-se posteriormente o processo de separação iterativa descrito na subsecção 2.3.5, de modo a adicionar outras componentes gaussianas, uma a uma, a cada estado dos modelos. Após a determinação de cada nova componente repete-se o ciclo de quatro reestimações embutidas por forma a estimar convenientemente todos os parâmetros. Estes modelos são então testados de novo, sempre com os mesmos sinais de fala. Nas curvas de resultados apresentados nas figura 5.1 verificou-se que apenas uma delas (*de* na figura (b)) regista um valor máximo nas ordenadas, com mais do que cinco componentes gaussianas por mistura. Mesmo neste caso, o valor obtido é aproximadamente igual ao máximo conseguido com um número inferior destas componentes. Nos resultados referentes aos oradores masculinos verificou-se que os valores máximos da taxa de reconhecimento foram obtidos com um número inferior de componentes gaussianas. Por este motivo, não se representaram na figura 5.2 os pontos experimentais com mais componentes por mistura.

Os oradores britânicos registaram de forma clara os melhores resultados. Tal foi verificado, mesmo quando cada grupo de oradores dispunha de modelos específicos para o reconhecimento do respectivo sotaque. A fala dos oradores nativos apresenta naturalmente menor variabilidade, de acordo com o que foi verificado nas experiências descritas na secção 4.8. Os oradores dinamarqueses obtiveram os melhores resultados entre os restantes grupos de oradores estrangeiros. A única excepção coube aos oradores masculinos ibéricos, que conseguiram melhores resultados com os modelos específicos do respectivo sotaque.

Os piores resultados foram obtidos com os oradores italianos e alemães e devem-se essencialmente aos problemas referidos na subsecção 3.2.3, a propósito da detecção de início e fim de palavra.

Tal como se verificou no capítulo 4, o facto de se dispor de modelos específicos para cada grupo de oradores com a mesma língua materna, aumentou significativamente o desempenho dos respectivos reconhecedores. Os resultados obtidos com oradores nativos foram repetidos nas figuras 5.1b e 5.2b, para efeitos de comparação, verificando-se que nenhum dos restantes reconhecedores se aproxima destes resultados. Quanto ao uso de mais de uma componente gaussiana, este foi vantajoso para os modelos referentes aos sotaques estrangeiros, mas não para os modelos treinados com oradores nativos. A justificação destes factos poderá também ficar a dever-se a uma maior variabilidade na fala da segunda língua, mesmo quando os oradores em análise partilham a mesma língua materna. Em concordância com o referido a propósito deste assunto na subsecção 3.1.1, a segunda língua é em geral adquirida e exercitada em circunstâncias muito variadas quando comparadas com as da língua materna, o que determina a correspondente variabilidade no sinal de fala.

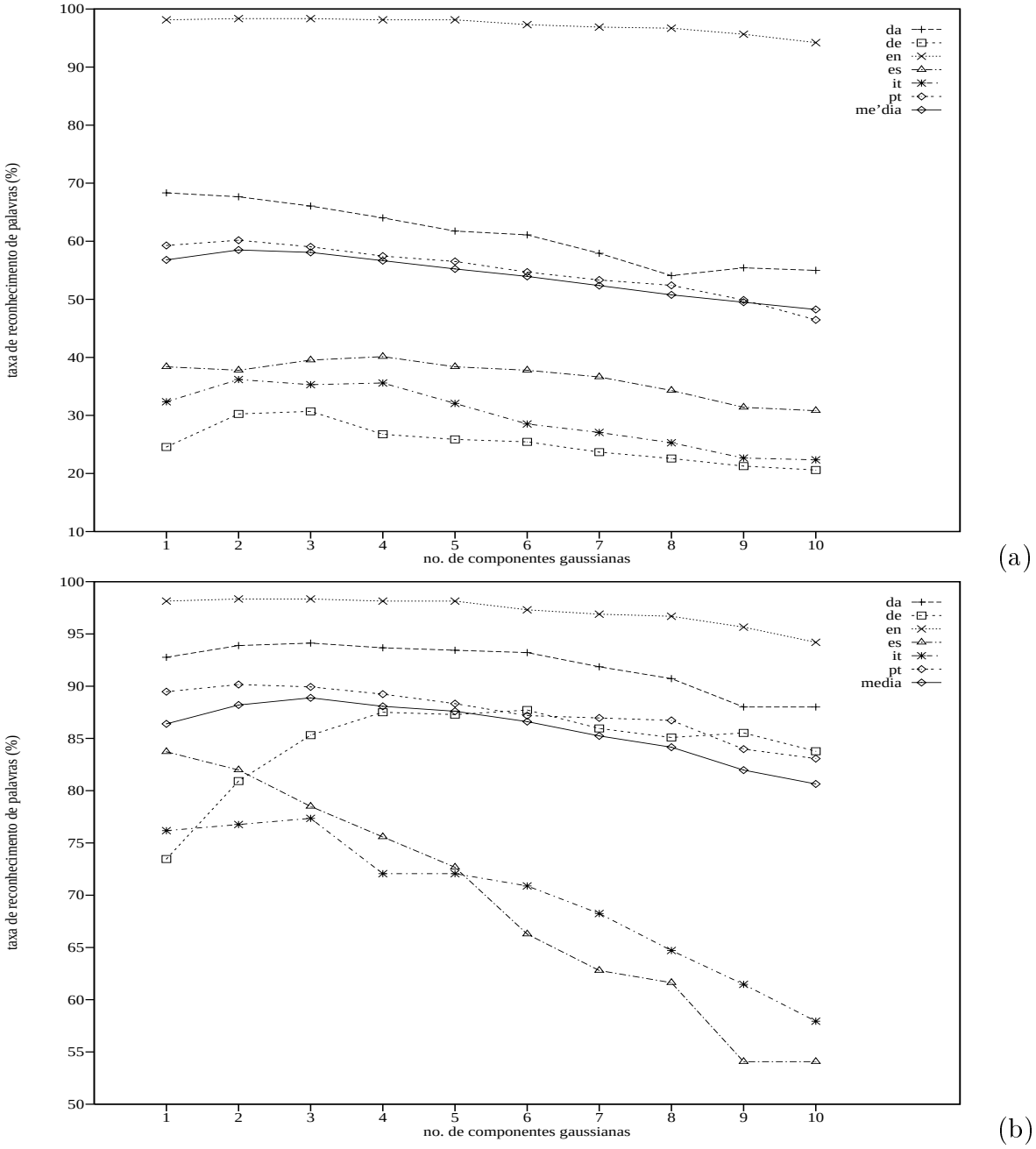


Figura 5.1: Taxa de reconhecimento (%) obtida com reconhecedores de modelos de palavra com diversas componentes gaussianas, treinados com: (a) oradores britânicos; (b) oradores de cada sotaque específico. Os corpora de treino e de teste incluem exclusivamente oradores do sexo feminino.

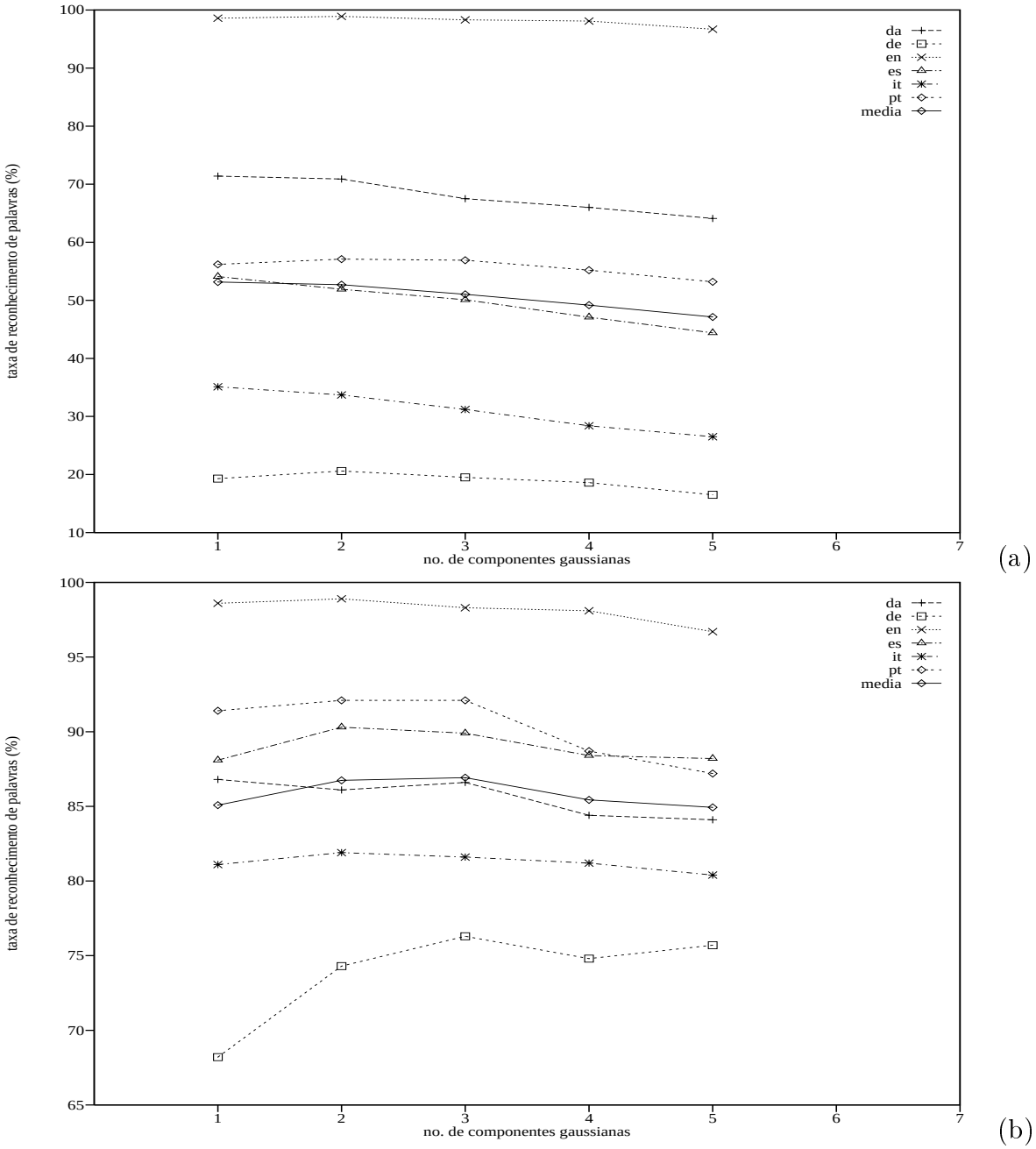


Figura 5.2: Taxa de reconhecimento (%) obtida com reconhecedores de modelos de palavra com diversas componentes gaussianas, treinados com: (a) oradores britânicos; (b) oradores de cada sotaque específico. Os corpora de treino e de teste incluem exclusivamente oradores do sexo masculino (Teixeira et al., 1997).

As experiências realizadas com os reconhecedores específicos de cada sotaque (figuras 5.1b e 5.2b) apresentaram acréscimos globais de mais de 20% na taxa de reconhecimento (linhas com a etiqueta de *média*) em relação aos resultados obtidos com o reconhecedor treinado exclusivamente com oradores nativos (figuras 5.1a e 5.2a). Em termos gerais podem ser identificados três motivos para esta melhoria no desempenho:

- Utilizou-se uma quantidade maior de material de fala para o treino dos modelos. Este facto é sobejamente conhecido no reconhecimento de fala. Contudo, neste caso, o material de fala em acréscimo foi utilizado para treinar outros reconhecedores.

- O material de treino acrescentado foi proferido por oradores não nativos, ou seja, incorporando características mais próximas do material de teste. Como consequência, os modelos obtidos deverão ser mais capazes de modelarem a fala de outros oradores não nativos, nomeadamente se tiverem as mesmas nacionalidades dos de treino, tal como aqui acontece.

- O material disponível para o treino e para os testes foi utilizado de forma selectiva. Os oradores foram seleccionados a priori de acordo com o respectivo sexo e língua materna: os de treino treinaram um conjunto específico de modelos; os de teste utilizaram o reconhecedor com o conjunto de modelos específicos para o respectivo presumível sotaque.

De acordo com o último motivo apresentado, a utilidade prática do segundo grupo de experiências é condicionada ao conhecimento prévio da língua materna de cada orador. Tal não aconteceu com o primeiro grupo de experiências, em que todos os oradores disponíveis para o teste utilizaram o mesmo reconhecedor. Este aspecto é mais grave na fase de teste, uma vez que se pretende eliminar qualquer decisão não contemplada pelos métodos automáticos disponíveis. A decisão implícita nestas experiências responde à pergunta: "qual o reconhecedor a utilizar com cada grupo de oradores?" De facto, a resposta foi dada a priori de acordo com os dados disponíveis para cada orador. Da qualidade desta decisão dependerá obviamente o desempenho global do sistema. A discussão de um sistema deste tipo será feita no capítulo 6. As experiências seguintes procuram utilizar o material de fala dos oradores estrangeiros para o treino de modelos, sem se efectuar uma decisão prévia de qual o sotaque associado a cada locução.

5.2.2 Modelos de palavra independentes do sotaque

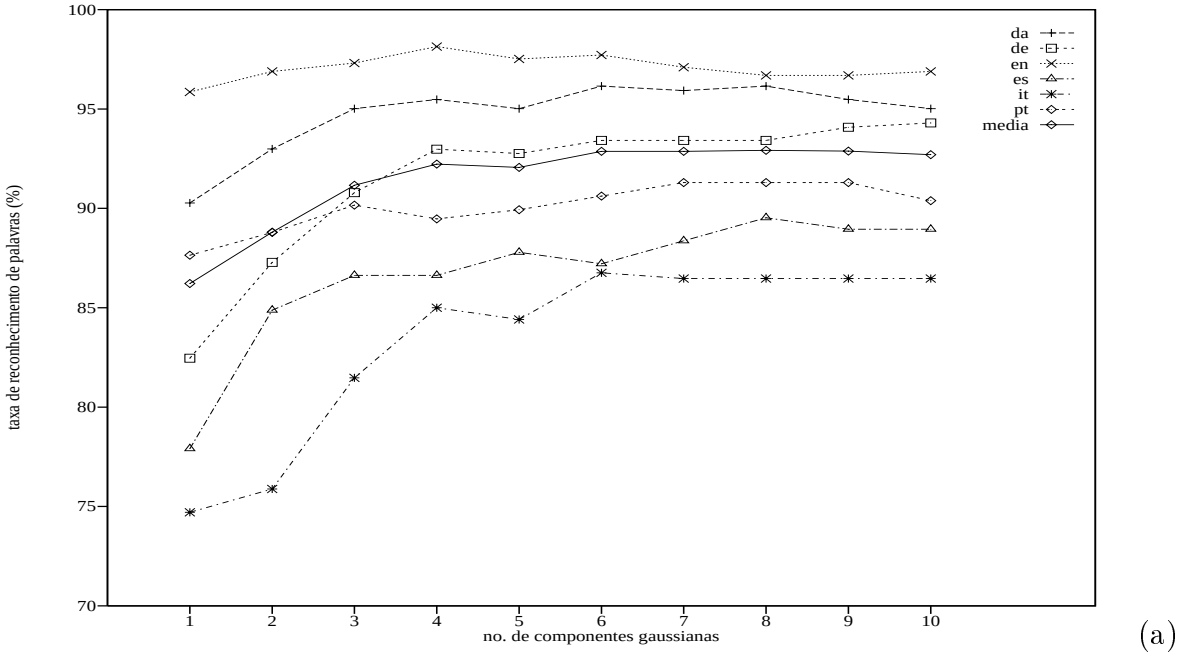
Sublinham-se dois factos conhecidos que estão na base da estratégia adoptada nesta secção para o reconhecimento automático dos sotaques estrangeiros: o problema pode ser considerado no quadro mais geral do reconhecimento independente do orador; os reconhecedores dependentes do orador ou de um grupo restrito de oradores, apresentam taxas de reconhecimento mais elevadas. Alguns dos compromissos impostos por estes dois tipos de reconhecedores foram já discutidos no capítulo de introdução deste trabalho.

A solução mais utilizada para o problema do reconhecimento independente do orador é a de considerar o maior número possível de oradores na recolha do corpus de treino e utilizar todo este material como se apenas de um orador se tratasse, ou seja, atendendo apenas às respectivas transcrições (por exemplo, no caso presente, a etiquetagem ortográfica). A solução adoptada nesta secção pode ser considerada uma generalização da anterior para o caso dos sotaques estrangeiros. Na prática, é em tudo igual à anterior, ou seja, embora existindo diferenças significativas entre os diversos grupos de oradores e consequentemente nos respectivos sinais de fala, estes são utilizados independentemente destas diferenças. Estas deverão ficar modeladas de forma conveniente nos modelos probabilísticos HMM independentes do orador.

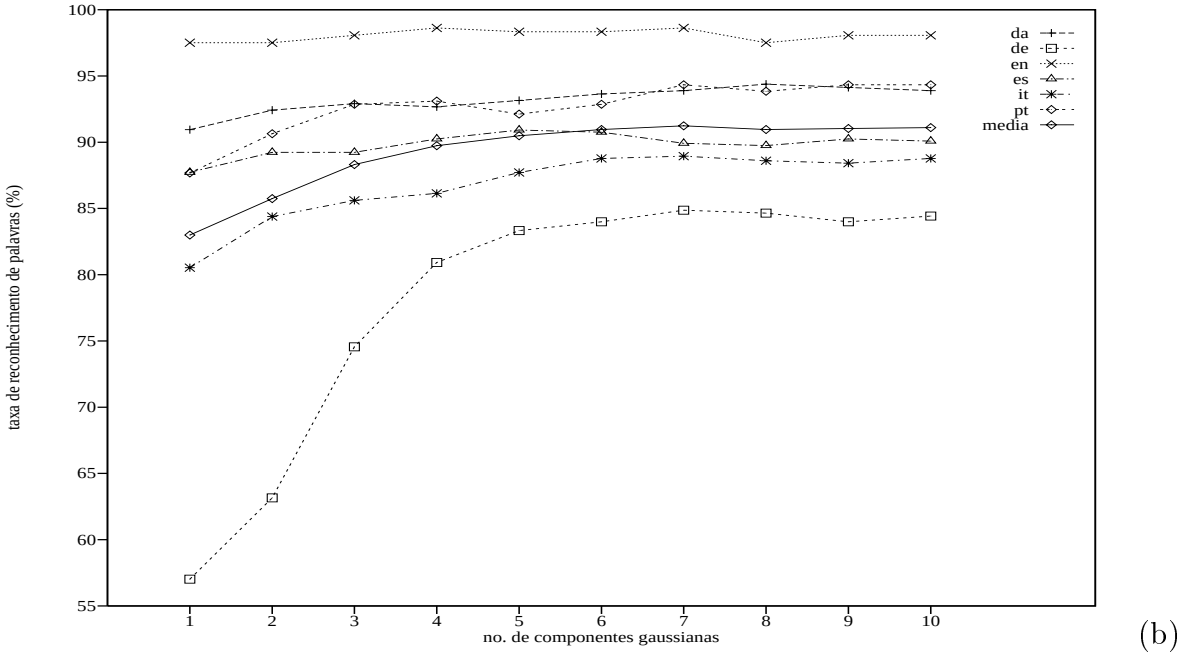
Na figura 5.3 apresentam-se os resultados obtidos com esta solução. Como seria de esperar, os oradores britânicos obtêm de novo os melhores resultados. A primeira vantagem deste método, quando confrontado com o uso dos reconhecedores específicos de cada sotaque é a de não necessitar de nenhum mecanismo de pré-determinação do sotaque do orador. Contudo, mais do que isso, apresenta um acréscimo significativo de desempenho com a generalidade dos oradores estrangeiros.

Outro aspecto a assinalar é o de, ao contrário das restantes experiências realizadas com modelos de palavra, se terem verificado também acréscimos de desempenho com o uso de múltiplas componentes de observação por estado. Na média dos sotaques o melhor resultado foi obtido com 7 componentes. Tal facto pode ser associado ao acréscimo de material de fala disponível para treinar cada um dos modelos de palavra (seis vezes mais material, ou seja, tantas quantos os sotaques disponíveis).

A concluir esta secção, deve ainda referir-se que existem outras alternativas, em geral mais elaboradas, em relação ao processo de obtenção destes modelos. Por exemplo, os investigadores Kubala e Schartz treinaram inicialmente modelos dependentes do orador para posteriormente serem combinados em modelos únicos, independentes do orador (Kubala e Schwartz, 1991). Desta forma conseguiram reduzir o número de oradores de treino



(a)



(b)

Figura 5.3: Taxa de reconhecimento (%) obtida com os reconhecedores de modelos de palavra com diversas componentes gaussianas, treinados com todos os sotaques. Os corpora de treino e de teste incluem oradores do sexo feminino (a) e masculino (b).

de 100 para 12 sem perdas significativas no desempenho do reconhecedor. No mesmo trabalho também foram detectadas vantagens no treino separado de modelos para cada um dos dois sexos.

5.3 Reconhecimento com transcrição fixa

As experiências realizadas com modelos de palavras isoladas, descritas na secção anterior, foram repetidas com modelos subpalavra. Em termos práticos, a característica mais marcante dos reconhecedores subpalavra é a de permitirem o reconhecimento de palavras, mesmo quando estas não se encontram representadas no respectivo corpus de treino. Esta característica é habitualmente designada por *independência do vocabulário*, por oposição aos reconhecedores de palavras isoladas, que são dependentes do vocabulário de treino.

De acordo com o referido na subsecção 2.5.5 utilizaram-se 46 modelos de fones e um léxico de pronúncia apropriado, com uma transcrição fixa por cada palavra do vocabulário de teste (apêndice A). Este conjunto de fones e de transcrições fonémicas foram utilizados independentemente dos diversos grupos de oradores empregues no treino e no teste. A utilização do conhecimento da fonética da língua materna de cada orador permitiria eventualmente a obtenção de modelos formalmente mais coerentes e com melhores resultados práticos. Contudo, a partir de um conjunto de modelos de fones nativos, é possível desenvolver estratégias com vista ao adequamento do reconhecedor a um novo sotaque (subsecção 5.3.2 e secção 5.4).

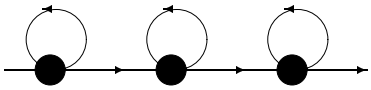


Figura 5.4: Topologia linear um modelo HMM de três estados emissores, utilizado na modelação de um fone.

Em termos gerais, houve a intenção de se utilizarem procedimentos e parâmetros tanto quanto possível semelhantes aos utilizados nas experiências anteriores, por forma a se poderem fazer comparações de resultados que não dependam de aspectos acessórios. Contudo, não foi possível evitar algumas diferenças que caracterizam os reconhecedores subpalavra utilizados. Nomeadamente, o número de estados utilizados na topologia dos modelos de palavras isoladas é manifestamente exagerado para o modelamento de elementos subpalavra, uma vez que se mantém a duração da trama de análise. Assim, adoptou-se uma topologia também linear mas de apenas três estados para os modelos

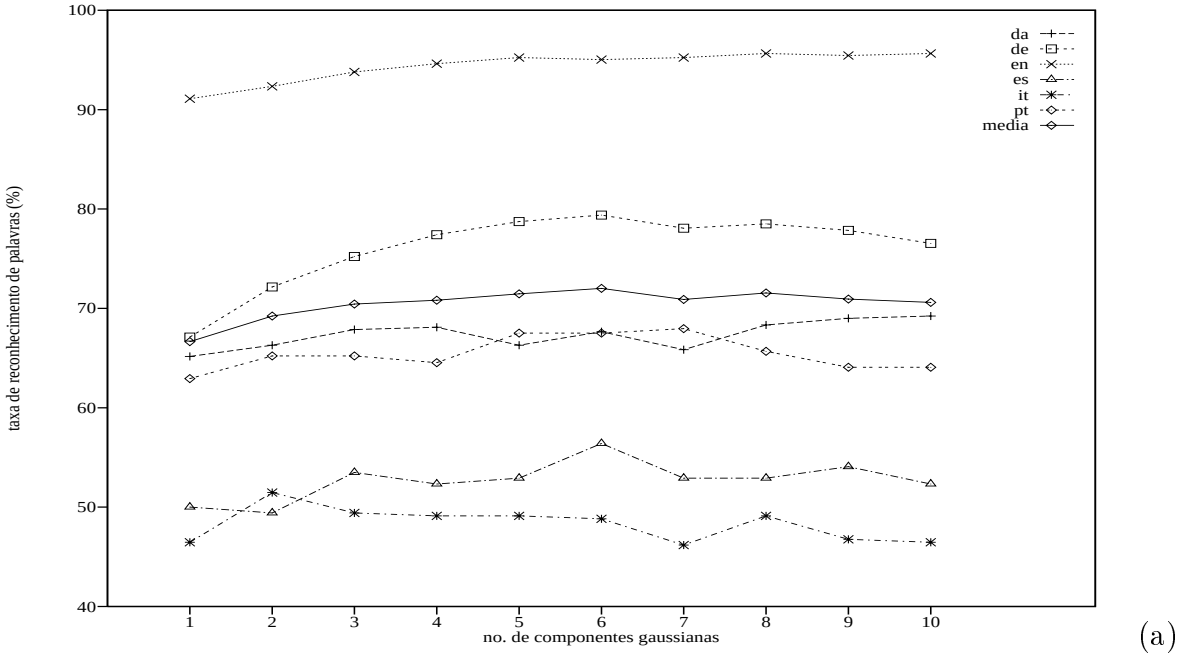
subpalavra (figura 5.4).

5.3.1 Modelos subpalavra para cada sotaque

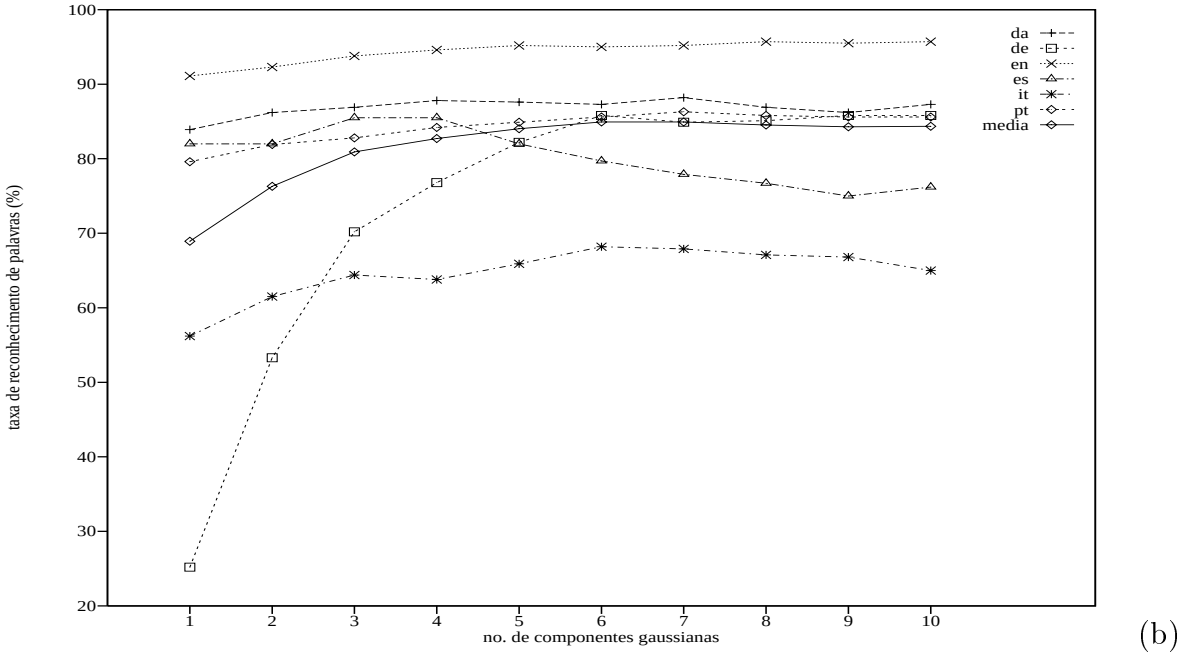
Os resultados apresentados nas figuras 5.5a e 5.6a referem-se às experiências realizadas com modelos subpalavra treinados exclusivamente com oradores ingleses nativos.

A análise qualitativa dos resultados obtidos com modelos de palavra mantém-se, (figuras 5.1a e 5.2a) nomeadamente no que se refere ao posicionamento relativo das taxas de desempenho de cada um dos grupos de oradores. As alterações mais evidentes são de dois tipos. A **primeira alteração** refere-se à diminuição da diferença de desempenho entre o reconhecedor nativo e da média de todos os reconhecedores. Tal facto deve-se, nomeadamente, a uma esperada quebra do desempenho deste tipo de reconhecedor com os oradores ingleses, mas também a uma evolução em sentido inverso da média dos reconhecedores estrangeiros. Esta evolução não encontra uma explicação nas experiências de reconhecimento tradicionais. O facto poder-se-á explicar pela falta de modelação dos aspectos de coarticulação associada aos modelos de fones utilizados. Os oradores estrangeiros têm em geral dificuldade em lidar com estes aspectos mais subtis da língua estrangeira, adoptando coarticulações menos dependentes do contexto ou, pelo menos, diferentes das proferidas pelos oradores nativos. Os fones que não existam ou sejam pouco comuns na língua materna do orador podem ser substituídos, respectivamente, pelos que lhes estão mais próximos ou são mais comuns no inventário fonético dessa língua. O mesmo pode acontecer com as coarticulações, embora tal facto seja mais difícil de verificar apenas por audição. A **segunda alteração** a realçar é o facto de os melhores resultados terem sido obtidos com cerca de três componentes gaussianas de observação, enquanto que nos modelos de palavra a utilização de uma única gaussiana era igualmente eficaz em termos de desempenho. Tal pode ser justificado pela disponibilidade de mais repetições de algumas das unidades subpalavra para o treino de cada modelo em comparação com o número aproximadamente fixo de palavras disponíveis para o treino dos modelos de palavras.

Nas experiências seguintes, foram treinados conjuntos de modelos de fones para cada sotaque utilizando material de fala obtido exclusivamente a partir de oradores da correspondente nacionalidade. Nas figuras 5.5b e 5.6b apresentam-se os resultados de reconhecimento obtidos com o teste de cada conjunto destes modelos com o correspondente grupo de oradores com o mesmo sotaque. Testaram-se de novo modelos com diversos números de componentes gaussianas de observação. Tal como nos resultados das figuras 5.5a e 5.6a, verifica-se novamente a vantagem do uso de componentes múltiplas. Os



(a)



(b)

Figura 5.5: Taxa de reconhecimento (%) obtida com reconhecedores de modelos subpalavra com diversas componentes gaussianas, treinados com: (a) oradores britânicos; (b) oradores de cada sotaque específico. Os corpora de treino e de teste incluem exclusivamente oradores do sexo feminino.

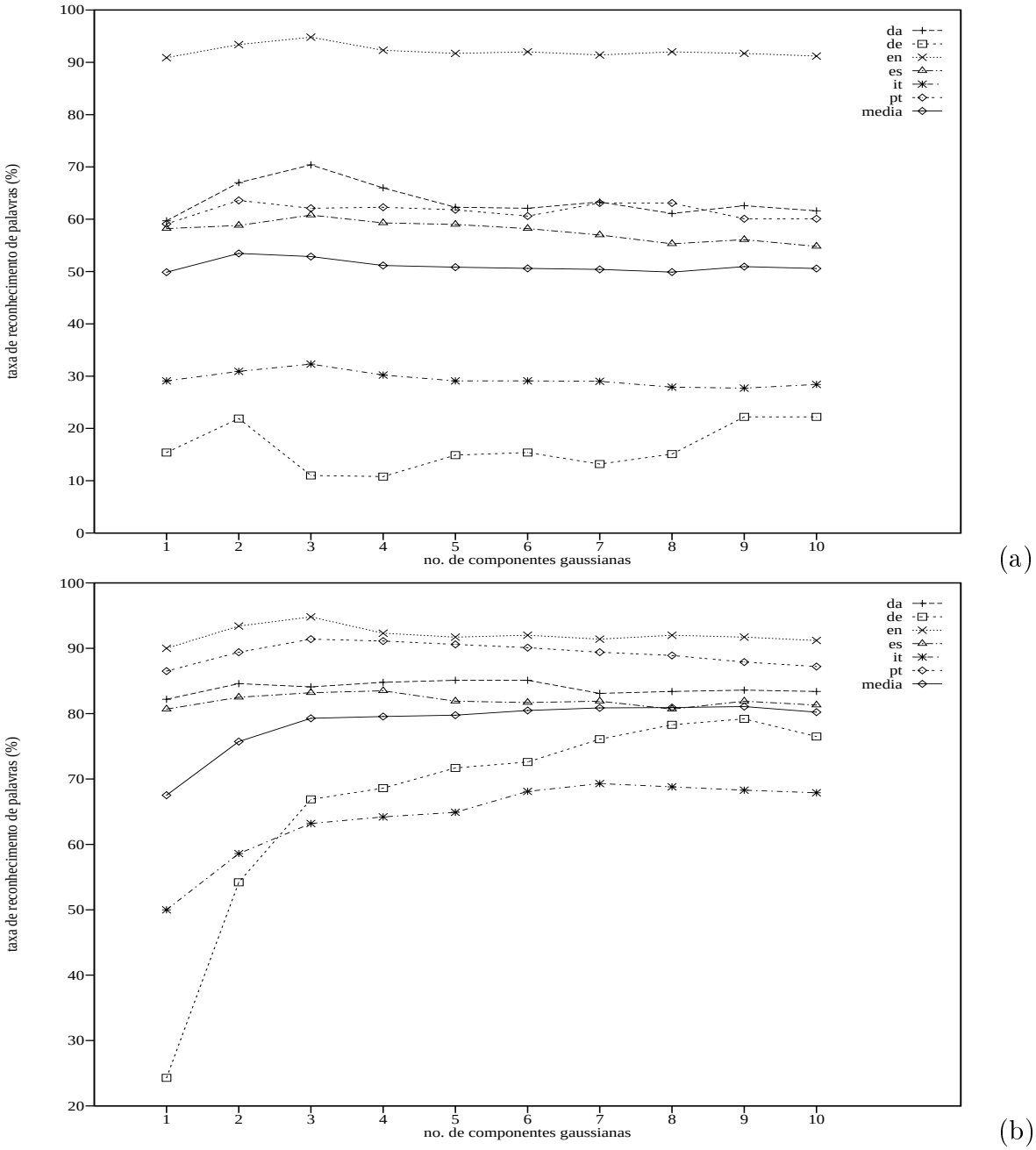
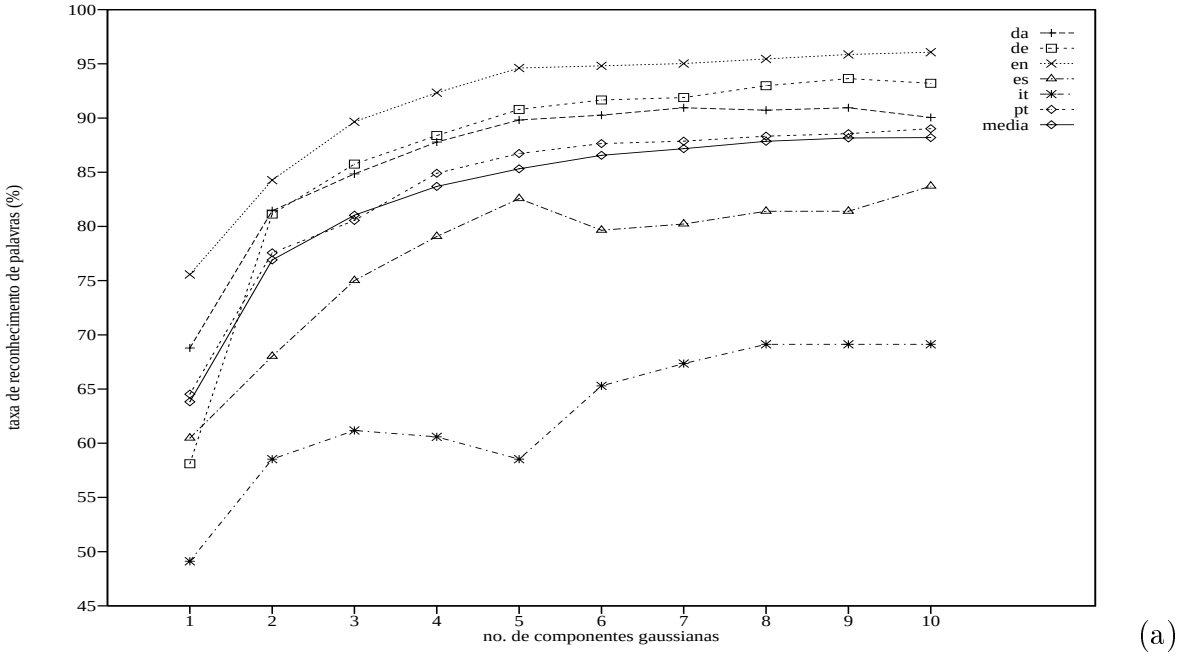


Figura 5.6: Taxa de reconhecimento (%) obtida com reconhecedores de modelos subpalavra com diversas componentes gaussianas, treinados com: (a) oradores britânicos; (b) oradores de cada sotaque específico. Os corpora de treino e de teste incluem exclusivamente oradores do sexo masculino (Teixeira et al., 1997).

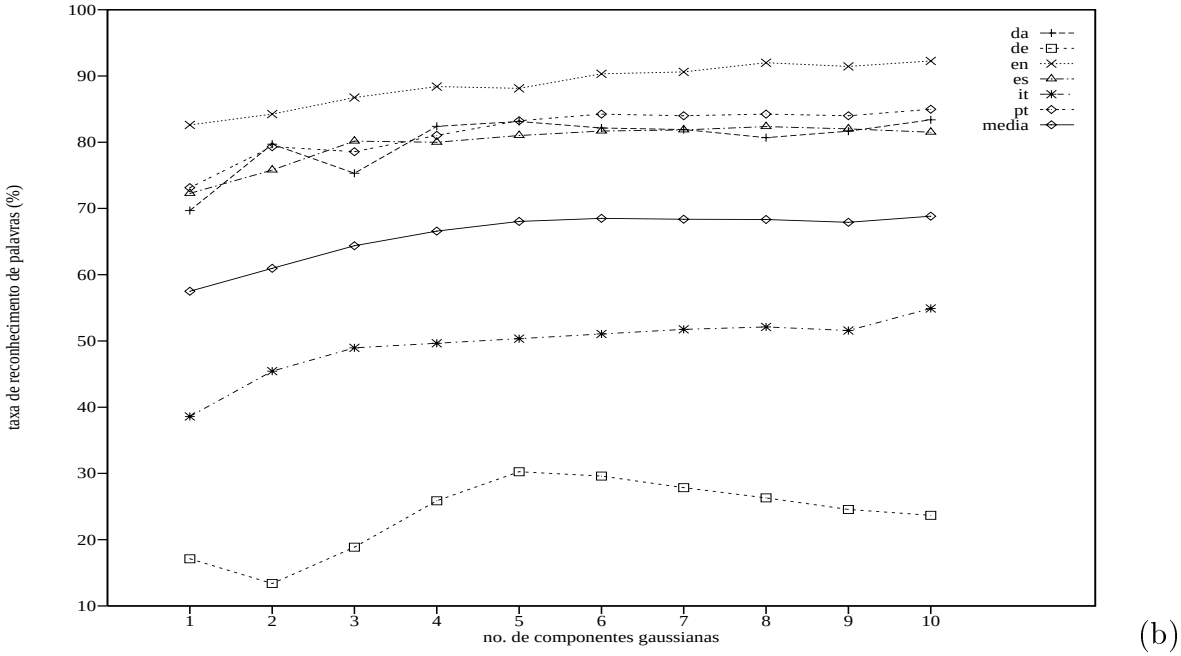
reconhecedores específicos de cada sotaque estrangeiro assim obtidos, apresentaram taxas de reconhecimento muito superiores às obtidas com os modelos dos oradores nativos (tal como se verificou com os modelos de palavra, nas figuras 5.1b e 5.2b). Os motivos que podem justificar esta diferença não deverão ser diferentes dos anteriormente apontados para o caso dos modelos de palavras. Também com os modelos de fones se verifica que nenhum dos restantes reconhecedores se consegue aproximar dos resultados obtidos com os oradores nativos. Contudo esta diferença diminuiu, sendo apenas de 3,4% para os oradores portugueses. A quebra da taxa de reconhecimento verificada entre os reconhecedores de modelos de palavra e os de fones dos oradores britânicos não é acompanhada pelos modelos dos oradores estrangeiros. No caso particular dos oradores alemães, obtiveram-se agora taxas de reconhecimento muito próximas ou mesmo superiores à média dos oradores testados (figura 5.5b). A ausência de modelamento da coarticulação em conjugação com a pouca consistência desta entre os diversos oradores de cada grupo pode ser justificativa destes factos. No caso dos oradores alemães, a justificação óbvia deveria ser atribuída ao facto de ser agora possível ultrapassar os problemas referentes a uma detecção de início e fim de palavra com menos qualidade e que resultaram em segmentos de sinal com maior duração antes e depois da ocorrência da própria palavra (subsecção 3.2.3). Assim, a adopção de um modelo explícito de três estados emissores para o silêncio deverá permitir o modelamento de durações superiores deste fone, comparativamente ao obtido com os modelos de palavras isoladas. Este facto será tanto mais relevante quanto maior for a duração da palavra ou o seu número de fones, uma vez que se adoptou um número fixo de estados para os modelos de palavras isoladas.

5.3.2 Modelos subpalavra independentes do sotaque

De acordo com as considerações feitas na subsecção 5.2.2, repetiram-se as experiências aí descritas utilizando modelos subpalavra. Assim, cada um dos modelos subpalavra foi treinado com todas as repetições de todos os sotaques disponíveis. Os resultados podem ser avaliados por inspecção da figura 5.7. Como se poderia esperar, de acordo com as experiências anteriores com modelos subpalavras e com os resultados apresentados na figura 5.3 com modelos de palavras, acentuam-se ainda mais os acréscimos de desempenho com a utilização de múltiplas componentes de observação. Também conforme o esperado, melhoram-se os resultados em relação aos descritos na subsecção 5.3.1. Contudo, os acréscimos de desempenho não foram tão significativos como os conseguidos com os modelos de palavras isoladas, sendo mesmo globalmente inferiores aos obtidos nas experiências com os reconhecedores específicos para cada sotaque. Tal facto confirma a ideia de que cada grupo de oradores terá tendências específicas deste mesmo grupo, manifes-



(a)



(b)

Figura 5.7: Taxa de reconhecimento (%) obtida com os reconhecedores de modelos subpalavra com diversas componentes gaussianas, treinados com todos os sotaques. Os corpora de treino e de teste incluem oradores do sexo feminino (a) e masculino (b).

tadas com particular intensidade em alguns dos fones. Enquanto que com os modelos de palavra a vantagem de se dispor de mais material de fala não podia discriminar partes específicas da palavra, nos modelos de fones essa vantagem não é distribuída de igual modo pelos diversos fones. O exemplo mais evidente vem de novo do caso dos oradores alemães, apontando-se mais uma vez a causa dos maus resultados obtidos, às diferenças na modelação dos silêncios.

5.3.3 Uso de polifones

De acordo com Dalsgaard e Andersen, a investigação na área do reconhecimento multilingue necessitava de aprofundar considerações teóricas de modo a que as metodologias desenvolvidas para cada língua pudessem ser extendidas para uma análise conjunta de várias línguas (Dalsgaard e Andersen, 1992). Para tal, procuraram identificar conjuntos de fonemas de diversas línguas com propriedades de realização suficientemente semelhantes entre si e que designaram por polifonemas. Por sua vez os restantes fonemas, que deveriam continuar a ser considerados isoladamente para a língua a que pertencem, foram designados por monofonemas. De forma análoga ao desenvolvimento do conceito de fone a partir do conceito de fonema (secção 2.5.1) define-se agora o conceito de *polifone* como extensão do conceito de *polifonema*.

A utilização de polifones foi ensaiada no decorrer deste trabalho, não revelando contudo qualquer vantagem significativa para os métodos de reconhecimento de fala desenvolvidos. Tal pode ser explicado pelo facto de o polifonema neste contexto ser em geral um fonema que ocorre com grande frequência na segunda língua. A vantagem que poderia advir deste conceito para o reconhecimento de fala, seria o de poder treinar os modelos de polifones sem ter que proceder a uma recolha com todos os oradores estrangeiros. Esta forma de economizar meios implicaria uma definição de vocabulários ricos em monofonemas para cada sotaque estrangeiro. Contudo, não só esta definição representa uma tarefa complexa, como o corpus de fala resultante pode ficar sujeito a uma utilização muito restrita. De qualquer modo, esta definição de vocabulários específicos não foi feita, nem poderia ser feita a priori. Assim, compreende-se que o excesso de repetições de fones de treino dos polifonemas não tenha produzido melhorias assinaláveis no desempenho dos reconhecedores.

5.4 Determinação automática de transcrições

A determinação manual de transcrições fonéticas estreitas, isto é, com a intervenção directa de um operador humano, é uma tarefa a evitar, tanto quanto possível. A dimensão dos corpora de fala exigidos para as novas aplicações do reconhecimento de fala tornam esta tarefa impraticável. Por outro lado, os resultados obtidos com as transcrições fonéticas manuais não são necessariamente os melhores.

Qualquer tipo de transcrição obtida por processos que envolvam um operador humano são, em geral, baseados em unidades subpalavra tipificadas (subsecção 2.5.1): o fonema, a semi-sílaba, a sílaba, etc. Os processos automáticos eliminam esta dependência, existindo mesmo estudos que indicam bons resultados com o uso de unidades subpalavra muito dificilmente caracterizáveis ou identificáveis por operadores humanos (Lee et al., 1988).

A utilidade de se obterem transcrições precisas por via automática é da maior utilidade em todas as áreas do processamento do sinal de fala, nomeadamente, na codificação de fala, com o advento possível da nova geração de codificadores ditos fonéticos. Este tipo de codificadores separa junto da fonte emissora, tanto quanto possível, as características específicas do orador do restante conteúdo informativo da mensagem. As características do orador não necessitam de ser enviadas continuamente e a restante informação pode ser representada numa espécie de transcrição fonética estreita. Desta forma obtêm-se taxas de ocupação do canal extremamente baixas. Contudo, estas transcrições têm de ser obtidas por via automática e em tempo real por forma a não comprometer o diálogo entre os utilizadores do canal.

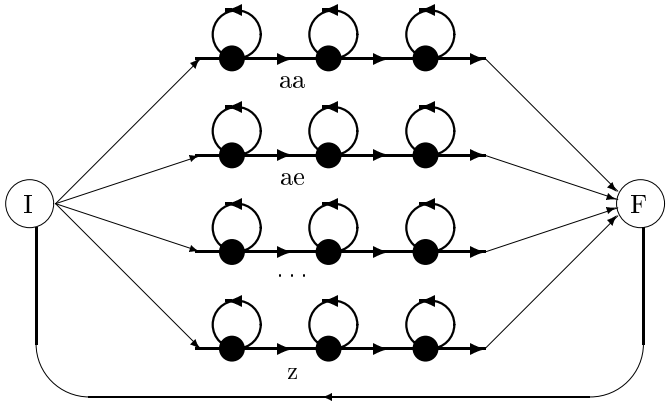


Figura 5.8: Representação do descodificador fonético convencional.

Tirando partido das técnicas de reconhecimento de fala tradicionais, podem ser for-

muladas algumas soluções para o problema da determinação de transcrições por via automática. A solução mais conhecida pode ser descrita com base na topologia de HMMs descrita na figura 5.8. No essencial, existe um estado inicial a partir do qual é possível atingir qualquer modelo subpalavra e um estado final para qual convergem todas as transições de saída destes modelos. Existe ainda uma transição que permite o regresso do estado final ao estado inicial até se esgotarem as observações disponíveis no sinal a transcrever. A transcrição final pode assim ser obtida com o uso de um algoritmo de Viterbi convencional. Este dispositivo é designado por *descodificador fonético*. Através das probabilidades de saída do estado inicial é possível introduzir a informação correspondente a um modelo linguístico do tipo unigrama (secção 2.6).

Uma variante da topologia descrita na figura 5.8 prescinde do estado inicial e final, admitindo transições directas do último estado de cada modelo subpalavra para o estado inicial de todos os modelos subpalavra. Considerando cada modelo subpalavra como um macroestado, configura-se uma nova topologia em que existem transições entre todos os macroestados. Os modelos que apresentam esta topologia são designados por ergódicos. Os modelos ergódicos admitem a implementação de modelos linguísticos baseados em bigramas, uma vez que se pode dispor das probabilidades de transição entre duas unidades subpalavra quaisquer.

As sequências de unidades subpalavra obtidas com um descodificador fonético variam muito entre repetições da mesma palavra. Tal facto verifica-se mesmo entre repetições do mesmo orador o que não permite determinar o que geralmente se convencionou designar por transcrição. Estes inconvenientes foram identificados desde muito cedo entre os investigadores que começaram a desenvolver reconhecedores para fala contínua: Jelinek et al. propuseram um método para ultrapassar este problema (Jelinek et al., 1975) que estima as probabilidades de uma cadeia de Markov num determinado sinal de fala; Lucassen e Mercer usaram uma abordagem da teoria da informação de modo a determinarem a transcrição fonémica a partir da representação ortográfica (Lucassen e Mercer, 1984). Este método exige dicionários de treino de grande dimensão. A sua utilização pode ser extendida a unidades acústicas determinadas automaticamente (Bacchiani et al., 1996).

Em alguns sistemas de reconhecimento mais avançados surgiram outras motivações para a utilização de transcrições fonéticas obtidas por via automática. Asadi et al. introduziram no reconhecedor de fala contínua BYBLOS, da BBN[2] um detector de *palavras novas* (Asadi et al., 1991). Consideram como *palavras novas* aquelas que não constam no léxico de pronúncia nem nos modelos linguísticos do sistema BYBLOS. Em vez de simplesmente rejeitarem estas palavras (como é feito no capítulo 4) decidiram incorporar a nova

[2] "URL:http://www.bbn.com".

palavra no próprio sistema. Esta atitude, a seu tempo inédita, só faz sentido em aplicações com vastas implicações semânticas tais como as decorrentes das áreas da inteligência artificial ou no caso dos conversores de fala para editores de texto, também designados por sistemas de ditado. Os aspectos da incorporação de uma nova palavra no modelo linguístico e eventualmente no modelo semântico da aplicação, estão claramente fora do âmbito do presente trabalho. O que interessa aqui realçar é forma como foi determinada a transcrição fonémica da palavra nova. Para este efeito propuseram um novo método que toma como dados de entrada a transcrição fonética obtida a partir do sintetizador de fala DECtalk[3], uma locução da nova palavra do orador que a produziu pela primeira vez e uma transformação probabilística. Esta transformação opera sobre a transcrição do DECtalk de modo a obter uma rede de pronúncias possíveis que irá impor restrições no processo de reconhecimento fonético. Este processo será efectuado sobre a única locução disponível e permitirá identificar qual a transcrição que apresenta o menor custo na rede de pronúncias. A transformação referida é construída a partir de uma matriz de confusão de fonemas, obtida de dois conjuntos de transcrições de um grande número de palavras. Um dos conjuntos é o das transcrições correctas, o outro, o das transcrições correspondentes obtidas a partir do DECtalk. Esta matriz é normalizada de forma a dispor-se da probabilidade do DECtalk confundir um dado fonema com outros. Deste modo, dada a transcrição fonética da nova palavra obtida com o DECtalk pode-se construir uma rede com um número finito de estados representando as transcrições possíveis.

Tal como os respectivos autores referem, talvez o maior inconveniente deste método seja o facto da escolha da transcrição final se basear numa única locução da palavra nova. Contudo, tal resulta essencialmente da imposição prática de se actualizar o sistema de reconhecimento logo a partir do momento em que surge uma palavra nova.

Michael Riley publicou um trabalho no mesmo ano do de Asadi et al. (1991) sobre um modelo estatístico para a geração de redes de pronúncia (Riley, 1991). Definindo fones como realizações acústicas distintas de fonemas, pretende com este trabalho predizer variações alofónicas, isto é, obter um mapa de conversão de fonemas para fones. O tipo de contexto fonémico que se pretendeu modelar não depende de fonemas individualmente considerados mas de classes de fonemas em determinadas posições, por exemplo, se o fonema anterior é ou não uma vogal. Riley considerou que a utilização das estatísticas do tipo *N-gramas* não seriam a opção ideal para este modelamento e preferiu utilizar uma árvore de decisão gerada estatisticamente a partir das transcrições fonéticas de 3024 frases da TIMIT. Cada ramo da árvore é associado a um subconjunto de fonemas em determinada posição. Esta árvore é depois usada para determinar as realizações fonéticas

[3]DECtalk é uma marca registada da Digital Equipment Corporation

de outras 336 frases da TIMIT. Riley determinou a entropia condicional de um fone dado o fonema que se lhe ajusta, em cerca de 1,5 bits. O método que propõe permite-lhe reduzir essa incerteza para cerca de 0,8 bits. Contudo, não existe conhecimento de aplicações práticas, nomeadamente no reconhecimento de fala.

Bahl et al., apresentaram um trabalho em que utilizavam unidades acústicas que designaram por *fenones* (Bahl et al., 1993) (subsecção 2.5.1). No desenvolvimento desse trabalho surgiu a necessidade de determinar as correspondentes transcrições *fenónicas* com base em várias (N_p) locuções de determinada palavra $y^{(1)}, y^{(2)}, \ldots, y^{(N_p)}$. Para tal, desenvolveram um método que começa por determinar as sequências de fenones $\theta^{(i)} \in \mathcal{S}; i = 1, \ldots, N_p$ mais prováveis para cada repetição ($\mathcal{S}$ é o conjunto de todas as sequências de fenones possíveis):

$$\theta^{(i)} = \arg\max_{s \in \mathcal{S}} P(y^{(i)}|s)$$

A transcrição pretendida $\hat{\theta}$ é escolhida no conjunto de transcrições $\Theta = \{\theta^{(i)}; i = 1, \ldots, N_p\}$ determinadas com um descodificador fonético convencional. A transcrição escolhida $\hat{\theta}$ é aquela que poderia ser produzida com maior probabilidade por todas as N_p repetições, isto é, aquela para qual o produto das verosimilhanças de todas as repetições dada a mesma transcrição, apresente o valor máximo:

$$\hat{\theta} = \arg\max_{\theta^{(i)} \in \Theta} \prod_{j=1}^{N_p} P(y^{(j)}|\theta^{(i)})$$

Os testes de reconhecimento de fala realizados com palavras isoladas e frases naturais revelaram melhores desempenhos, atribuídos essencialmente ao uso dos *fenones*.

Haeb-Umbach et al. apresentaram um estudo comparativo com este método que designaram por *método dos candidatos múltiplos* e outros métodos de transcrição automática equivalentes (Haeb-Umbach et al., 1995). Os desempenhos de cada método foram avaliados para uma mesma tarefa de reconhecimento de fala.

Um segundo método descrito por estes autores foi designado por *método da transcrição média*. Neste método foi criado um modelo de palavra com apenas uma única densidade de probabilidade de emissão treinado com todas as N_p repetições. Este modelo pode ser interpretado como uma "média" $\bar{y}$ dessas repetições. A transcrição da palavra desconhecida é então dada pela sequência de unidades subpalavra que produziu com maior probabilidade esta "locução média":

$$\bar{\theta} = \arg\max_{s \in \mathcal{S}} P(\bar{y}|s)$$

Ressalta o facto de, ao contrário do método dos candidatos múltiplos, esta transcrição não se encontrar em geral representada no conjunto Θ. Pode por isso mesmo ser incluída como candidata adicional $\theta^{(n+1)}$ no método dos candidatos múltiplos, por forma a constituir um terceiro método, também testado. O quarto método testado resulta do anterior, adicionando à lista de candidatos a transcrição fonémica $\theta^{(n+2)}$ obtida de um léxico de pronúncia.

Os dois primeiros métodos descritos não apresentam grandes diferenças de desempenho em reconhecimento de fala. Já a terceira alternativa, a que associa o método da transcrição média com a dos candidatos múltiplos, apresenta melhorias significativas, enquanto que a inclusão da transcrição fonémica não altera estes resultados significativamente. Haeb-Umbach et al. compararam ainda o uso destes métodos de transcrição automática com o método convencional baseado nas transcrições de um léxico de pronúncia. Obtiveram assim desempenhos ligeiramente superiores aos obtidos com um reconhecedor de modelos de palavra. Contudo, para este acréscimo de desempenho pode ter contribuído o facto do reconhecedor de palavras isoladas dispor apenas de uma única densidade gaussiana na probabilidade de observação.

Uma alternativa menos elaborada é a de utilizar simultaneamente todas as transcrições obtidas com um descodificador fonético convencional. Rose et al. compararam resultados com a utilização de unigramas, bigramas e trigramas obtidos de um corpus de seis milhões de palavras da "Associated Press News Wire" (Rose et al., 1996; Rose e Lleida, 1997). Neste sistema, o orador repete três vezes cada palavra nova, ficando, portanto, registadas três transcrições por palavra. Este tipo de solução não integra a informação respeitante à transcrição num único modelo. Tal facto implica, entre outros, os seguintes inconvenientes: na inspecção directa das transcrições é difícil antever o comportamento conjunto dos diversos modelos; por outro lado estes modelos separados não integram generalizações sobre o comportamento do sinal. Para concretizar este último aspecto, considere-se que através de um procedimento presumivelmente ideal, se determinavam as transcrições fonéticas $\theta^{(1)}$ e $\theta^{(2)}$ de duas locuções $y^{(1)}$ e $y^{(2)}$ de uma mesma palavra. Considere-se também que a transcrição $\theta^{(1)}$ difere de $\theta^{(2)}$, podendo-se obter uma da outra à custa de duas operações do tipo substituição, supressão ou inserção de fones. Considere-se, por exemplo, um reconhecedor que utilize exclusivamente duas transcrições para a palavra "Saturday": $\theta^{(1)} = \{$s ae t er d ey$\}$ e $\theta^{(2)} = \{$s ae dh er dh ey$\}$. Este reconhecedor não se encontra devidamente preparado para o surgimento de uma locução $y^{(3)}$ cuja transcrição $\theta^{(3)}$ possa ser obtida a partir de $\theta^{(1)}$ ou $\theta^{(2)}$ através de apenas uma das duas operações anteriormente referidas (por exemplo $\theta^{(3)} =$ s ae dh er d ey). Ou seja, se não houver uma relação de causalidade entre as duas substituições ocorridas, será de interesse

manter em aberto a possibilidade de poderem ocorrer separadamente. Este tipo de situações podem ser consideradas em termos probabilísticos num único modelo de transcrição, nomeadamente com a utilização das redes de transcrição descritas na secção seguinte.

Desde a altura em que se identificou a importância do modelamento da variação de pronúncia no contexto desta dissertação (Teixeira et al., 1996), têm surgido inúmeras publicações que utilizam este tipo de modelamento no âmbito do reconhecimento de fala em geral[4] (Strik e Cucchiarini, 1998). Além disso e conforme já se referiu, a determinação da transcrição pode ser fundamental para a inclusão de palavras novas nos reconhecedores automáticos.

Recentemente usaram-se modelos de pronúncia específicos para grupos de oradores com diferentes sotaques regionais do inglês (Humphries et al., 1996; Humphries et al., 1997; Humphries e Woodland, 1998). O pressuposto utilizado considera que as variações de pronúncia mais importantes, determinadas pelos sotaques regionais, são mais evidentes nas vogais do que nas consoantes. Com base neste pressuposto, Humphries e Woodland constroem uma rede de reconhecimento para cada palavra que permite duas substituições de vogais a partir de uma forma de citação nativa (secção 2.5.4). Para transcrever uma nova locução, determina as três melhores transcrições obtidas na referida rede. Estas transcrições são comparadas com a anterior forma de citação de modo a gerar listas de substituições de vogais dependentes do contexto. Posteriormente, esta informação é generalizada aos restantes contextos, utilizando uma técnica de agrupamento baseada em árvores de decisão. A abordagem descrita necessita da disponibilidade prévia de uma forma de citação pelo que não é adequada à inclusão de palavras novas. Por outro lado, o pressuposto relativo à substituição de vogais poderá não ser extensível ao caso dos sotaques não nativos.

Bonaventura et al. realizaram um estudo de reconhecimento com oradores e vocabulários espanhóis, ingleses e italianos (Bonaventura et al., 1998). Analisaram as possíveis alterações introduzidas por oradores não nativos nas pronúncias nativas do inglês e do italiano. Por exemplo, o grupo ortográfico /*vogal*/ + /r/ seguido por uma consoante numa palavra italiana é usualmente proferido pelos oradores ingleses como uma única vogal (a palavra italiana "Portogallo" com a transcrição nativa [p o r t o g a l o] é alterada para [p o t o g a l o u]). A partir deste conhecimento fonético a priori determinaram pronúncias alternativas que foram utilizadas em experiências de reconhecimento. Num outro conjunto de experiências substituíram estas transcrições alternativas por outras, obtidas por análise dos resultados do reconhecimento automático. O pequeno acréscimo de desem-

[4] "Publications on pronunciation variation and ASR", "URL: http://lands.let.kun.nl/Tspublic/strik/pronvar/references.html".

penho obtido relativamente ao uso das primeiras transcrições alternativas, foi atribuído a fenómenos marginais específicos do corpus utilizado e não a efeitos não nativos mais generalizados, tais como os determinados a priori.

5.5 Método de determinação de redes de transcrição fonémicas

No problema do reconhecimento com oradores estrangeiros assume-se à partida as diferenças entre as diversas nacionalidades e, em contrapartida, não se pretende acentuar diferenças dentro de cada grupo de determinada nacionalidade. Esta perspectiva é intrínseca à estratégia de identificação do sotaque apresentada no capítulo 6. Por outro lado, pretende-se valorizar tanto quanto possível o conhecimento disponível sobre o sotaque nativo, nomeadamente adoptando como referência o conjunto de fones nativo obtido na secção 5.3.

De acordo com as limitações apresentadas pelos métodos descritos na secção anterior, pretende-se obter modelos de transcrição que integrem informação probabilística relativa à variabilidade das pronúncias possíveis. Por forma a isolar os aspectos referentes a cada sotaque optou-se pela obtenção de modelos separados para cada sotaque. Por outro lado, procurou-se uma abordagem que não necessitasse de qualquer forma de citação a priori. Deste modo, fica em aberto a possibilidade de se utilizar esta abordagem num sistema de reconhecimento que permita a detecção e inclusão de palavras novas, na sequência do trabalho descrito no capítulo 4.

De seguida descreve-se um método que permite determinar para cada palavra, uma rede probabilística de transcrições baseada em quaisquer elementos subpalavra.

5.5.1 Dados do problema

O método automático utilizado na determinação das redes de transcrição probabilísticas para uma dada palavra assume os seguintes dados de entrada:

- O conjunto das repetições da palavra a transcrever $\{y^{(1)}, y^{(2)}, \ldots, y^{(N_p)}\}$, com N_p tão grande quanto possível. Qualquer dos caminhos alternativos na rede de transcrições que se pretende construir deverá representar um número significativo de ocorrências.

- O conjunto de modelos dos elementos subpalavra

$$\{\lambda^{(i)} = (A^{(i)}, B^{(i)}, \Pi^{(i)}), i = 1, \ldots, N_s\}.$$

  Utilizaram-se, uma vez mais, os modelos HMMs subpalavra anteriormente descritos ($N_s = 46$, subsecção 2.5.5).

- O número máximo N_f de elementos subpalavra para cada palavra. No futuro, espera-se encontrar uma solução formal que permita eliminar este dado do problema. Existem contudo alternativas de índole prática, tais como medidas baseadas na duração média das repetições da palavra ou na transcrição ortográfica. Conforme se referiu, procura-se evitar o uso de uma forma de citação pré-definida.

5.5.2 Descrição do modelo

O método aqui proposto pode ser explicado a partir de um descodificador fonético convencional (subsecção 5.4). Este descodificador dispõe de um estado inicial que se confunde com o estado final. De facto, o único ramo de saída do estado final termina no estado inicial, não se considerando neste trajecto a emissão de qualquer observação. Como se referiu, as restantes probabilidades de transição permitem integrar as informações estatísticas equivalentes a um modelo linguístico baseado em unigramas, ou quanto muito em bigramas, no caso de se optar pelo modelo ergódico. Contudo, para a obtenção de uma transcrição de uma palavra com N_f unidades subpalavra, o tipo de informação necessária tem um horizonte temporal correspondente a um modelo linguístico baseado em N_f-*gramas*. Para representar esta informação é necessária uma topologia que permita estruturar ao longo do tempo a ocorrência das sucessivas unidades subpalavra. A topologia linear (figura 2.1) é tradicionalmente indicada para este efeito, pelo que se procurou associar esta topologia com a do descodificador fonético. O resultado desta associação é uma macrotopologia linear de descodificadores fonéticos desprovidos das transições dos respectivos estados finais para os estados iniciais. Pretende-se deste modo que a passagem por cada um destes descodificadores corresponda à emissão de uma única unidade subpalavra. Mais ainda, pretende-se assim dispor de uma relação entre o instante da ocorrência dessa unidade e um ramo de transição na referida topologia linear. As probabilidades de transição associadas a estes ramos, entre unidades subpalavra, permitem descrever a pretendida rede de transcrição probabilística.

Considere-se agora uma série ordenada R_n (para $n = 1, \ldots, N_f$) de conjuntos de modelos subpalavra. Permitem-se todas as transições, aqui designadas por *transições interfones*, de cada um dos elementos de um dado conjunto R_n para todos os elementos do

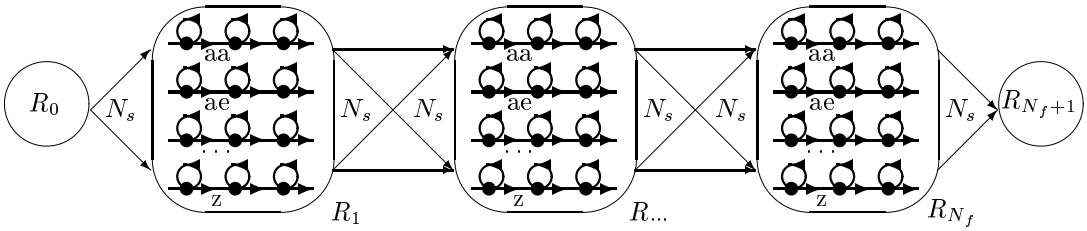


Figura 5.9: Interligação dos elementos da série ordenada R_n para $N_f = 3$. As ligações representadas entre os elementos $R_n : 1 \leq n \leq N_f$ representam $N_s^2 = 2116$ transições (N_s=46) entre modelos subpalavra.

conjunto seguinte R_{n+1}. Os estados não emissores inicial e final são inseridos na série como R_0 e R_{N_f+1}, respectivamente (figura 5.9). Obtém-se deste modo um novo modelo HMM, $\lambda_{\text{tr}} = (A_{\text{tr}}, B_{\text{tr}}, \Pi_{\text{tr}})$ que se designará por *modelo de transcrição*. No descodificador fonético convencional cada modelo subpalavra podia gerar repetidamente a unidade de fala correspondente, num número indeterminado de vezes, em diversos instantes da produção da palavra. No modelo de transcrição, por sua vez, cada modelo subpalavra gera a unidade correspondente uma única vez e num horizonte temporal específico.

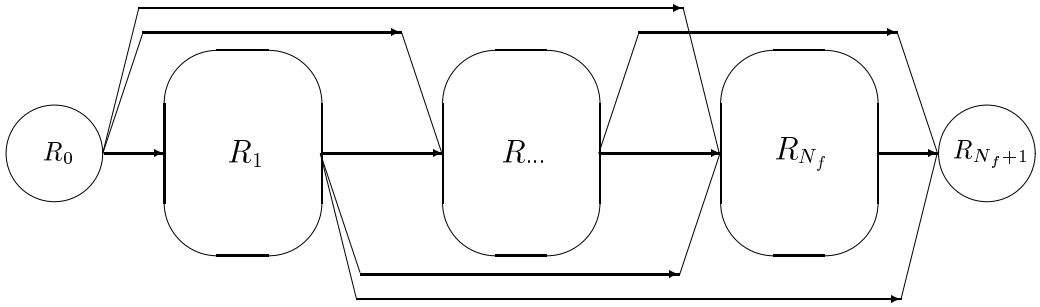


Figura 5.10: Representação esquemática de um modelo de transcrição de três fones. Cada trajecto de ligação entre conjuntos de fones R_n representam $N_s^2 = 2116$ transições. Cada trajecto de ligação entre os estados não emissores (R_0 e R_{N_f}) e os conjuntos de fones representam $N_s = 46$ transições.

As transições permitidas entre conjuntos consecutivos R_n e R_{n+1}, impõem uma estrutura temporal adequada à obtenção das descrições. Contudo, fenómenos como os de inserção ou de supressão fonémica não são modelados por esta estrutura e podem causar estimativas deficientes. Considere-se, por exemplo, o caso da supressão. A unidade subpalavra posterior à supressão será identificada no conjunto R_{n+1} destinado, em princípio, à unidade subpalavra suprimida, em vez de ser identificada no conjunto R_{n+2}. A solução encontrada foi a de permitir transições dos elementos de R_n, para os elementos dos conjuntos R_m tais que $m > n$ (figura 5.10). Assume-se ainda que a probabilidade de transição

decresce na razão do aumento da distância topológico-temporal $(m - n)$ conforme será explicado em detalhe na subsecção 5.5.3.

Os parâmetros dos modelos subpalavra utilizados no modelo de transcrição, são estimados tal como foi descrito na secção 5.3. Os restantes parâmetros do modelo de transcrição, que interessam agora determinar, são as probabilidades de transição entre esses mesmos modelos de subpalavra. Para tal, utiliza-se uma vez mais o algoritmo de Baum-Welsh, o qual permite determinar estas probabilidades maximizando a verosimilhança do modelo de transcrição, dadas as repetições disponíveis da palavra em causa.

Os modelos subpalavra encontram-se repetidos N_f vezes ao longo do modelo de transcrição:

$$R_i = R_j, \qquad 1 \leq i, j \leq N_f. \tag{5.1}$$

Assim, qualquer um dos conjuntos de modelos subpalavra R_i pode representar os restantes. Considerou-se que R_i permanece inalterável durante o treino do modelos de transcrição, ou seja, não altera nenhum modelo subpalavra, reestimando apenas as probabilidades de transição entre estes. Existem duas razões que justificam esta procedimento: primeiro, mantém-se desta forma a igualdade inicial expressa na equação 5.1 no sentido de garantir a manutenção de modelos subpalavra independentes do contexto; segundo, a obtenção de novos modelos subpalavra não constitui um objectivo do procedimento descrito. Em relação a esta segunda justificação, acrescenta-se que os modelos subpalavra foram previamente obtidos a partir de um processo de treino comprovado pelo estado da investigação actual (secção 5.3). Não se pretende, para já, alterar estes modelos por um processo de treino que é utilizado pela primeira vez, como é o caso daquele aqui descrito.

5.5.3 Inicialização do modelo

Uma vez que se pretende reestimar o modelo de transcrição com o algoritmo de Baum-Welsh, há que ter em consideração a respectiva inicialização e finalização. O problema da inicialização foi resolvido utilizando a mesma estratégia adoptada para os HMMs mais convencionais, considerando um modelo inicial com todas as características topológicas do modelo que se pretende determinar (subsecção 2.3.5). Contudo, o problema é agora mais complexo, uma vez que a dimensão do modelo de transcrição é muito superior à dos HMMs mais comuns e se pretende manter algum significado para determinados conjuntos dos respectivos estados. O problema da finalização é por sua vez resolvido de forma igual à considerada para outros modelos HMM mais convencionais. Efectivamente, o cálculo da verosimilhança continua a ser um subproduto do método aqui descrito e é com medidas de estabilização deste valor que tradicionalmente se determina a paragem deste processo

iterativo.

Conforme se referiu, os modelos subpalavra são dados do problema e determinam o tipo de transcrição a obter, ou seja, o tipo de unidades de fala a que se refere. A topologia dos modelos HMM subpalavra apresenta um único estado de entrada e um único estado de saída (figura 5.4). Nestas condições, cada modelo subpalavra pode ser considerado, no seu todo, como uma espécie de estado que se designa aqui por macroestado, por forma a distingui-lo dos estados elementares dos HMMs.

Define-se agora a *matriz de transcrição* Γ como sendo a matriz de transição do modelo de transcrição, quando descrito em termos dos referidos macroestados. Esta matriz contém toda a informação que se pretende obter com o presente método. Designe-se o número total de macroestados por $N_{fs} = N_f.N_s + 2$. A matriz de transcrição inclui N_{fs}^2 elementos. Na matriz seguinte, anularam-se as probabilidades de transição entre conjuntos R_n e R_m tais que $m > n + 2$, por forma a simplificar a sua representação:

$$\Gamma = \begin{bmatrix} 0 & \Gamma_{1,2} & \Gamma_{1,3} & \cdots & \Gamma_{1,2N_s+1} & \cdots & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & \Gamma_{2,2} & 0 & \cdots & 0 & \cdots & \Gamma_{2,N_s+2} & \cdots & \Gamma_{2,3N_s+1} & 0 & \cdots & 0 \\ 0 & 0 & \Gamma_{3,3} & \cdots & 0 & \cdots & \Gamma_{3,N_s+2} & \cdots & \Gamma_{3,3N_s+1} & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & \cdots & 0 & \cdots & 0 & 0 & \cdots & \Gamma_{N_{fs}-1,N_{fs}} \\ 0 & 0 & 0 & \cdots & 0 & \cdots & 0 & \cdots & 0 & 0 & \cdots & 1 \end{bmatrix}$$

Seguidamente determinam-se apenas os elementos não nulos da diagonal desta matriz ($\Gamma_{i,i} \neq 0$). A matriz de transição de cada modelo subpalavra $\lambda^{(i)} = (A^{(i)}, B^{(i)}, \Pi^{(i)})$, correspondente à topologia representada na figura 5.4, com $N = 3$ estados emissores e dois não emissores (o estado inicial e o final). Os respectivos elementos são aqui designados por *probabilidades de transição intrafones.*

$$A^{(i)} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & A_{1,1}^{(i)} & A_{1,2}^{(i)} & 0 & 0 \\ 0 & 0 & A_{2,2}^{(i)} & A_{2,3}^{(i)} & 0 \\ 0 & 0 & 0 & A_{3,3}^{(i)} & 1 - A_{3,3}^{(i)} \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

A matriz de transição A_{tr} do modelo de transcrição $\lambda_{\text{tr}} = (A_{\text{tr}}, B_{\text{tr}}, \Pi_{\text{tr}})$, de dimensão $(N_f.N_s.N+2)^2$, é construída a partir da matriz Γ e das matrizes de transição A_i de cada um

dos modelos subpalavra $\{\lambda^{(i)}, i = 1, \ldots, N_s\}$. Assim, a matriz A_{tr} pode ser representada da seguinte forma:

$$\begin{bmatrix}
0 & \Gamma_{1,2} & 0 & 0 & \Gamma_{1,3} & 0 & 0 & \cdots & 0 & \cdots & \Gamma_{1,2N_s+1} & \cdots & 0 & \cdots & 0 \\
0 & A^{(1)}_{1,1} & A^{(1)}_{1,2} & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & \cdots & 0 & 0 & \cdots & 0 \\
0 & 0 & A^{(1)}_{2,2} & A^{(1)}_{2,3} & 0 & 0 & 0 & \cdots & 0 & 0 & \cdots & 0 & 0 & \cdots & 0 \\
0 & 0 & 0 & A^{(1)}_{3,3} & 0 & 0 & 0 & \cdots & 0 & \Gamma_{3,N_s+2} & \cdots & \Gamma_{1,3N_s+1} & 0 & \cdots & 0 \\
0 & 0 & 0 & 0 & A^{(2)}_{1,1} & A^{(2)}_{1,2} & 0 & \cdots & 0 & 0 & \cdots & 0 & 0 & \cdots & 0 \\
0 & 0 & 0 & 0 & 0 & A^{(2)}_{2,2} & A^{(2)}_{2,3} & \cdots & 0 & 0 & \cdots & 0 & 0 & \cdots & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & A^{(2)}_{3,3} & \cdots & 0 & \Gamma_{3,N_s+2} & \cdots & \Gamma_{1,3N_s+1} & 0 & \cdots & 0 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & \cdots & 0 & 0 & \cdots & A^{(N_s)}_{3,3} \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & \cdots & 0 & 0 & \cdots & 1
\end{bmatrix}$$

Na inicialização do processo de reestimação do modelo de transcrição, dever-se-á obviamente utilizar a melhor estimativa possível para a matriz Γ. A disponibilidade de léxicos de pronúncia pode contribuir para a elaboração desta estimativa, embora o problema seja exactamente a inexistência ou a inexactidão desses mesmos léxicos (Roach e Arnfield, 1998). No caso presente, dispunha-se de um léxico de pronúncia nativo, podendo assim impor-se probabilidades de transição altas no trajecto através dos modelos subpalavra que descrevem a transcrição conhecida. Contudo, pretende-se modelar a existência de outras transcrições plausíveis, nomeadamente não nativas. Para tal deve-se prever a supressão de fones na transcrição conhecida e a inserção ou substituição de fones, reservando-se parcelas das probabilidades de transição para os fones alternativos mais prováveis. Esta solução não chegou a ser testada, uma vez que se conseguiram obter resultados de interesse sem ser necessária outra fonte de informação (subsecções 5.5.6 e 5.5.7).

Seguidamente, descreve-se o método utilizado no cálculo da matriz de transcrição inicial. Antes de mais, define-se a estrutura das matrizes de transição Γ e A_{tr} em termos do ordenamento dos respectivos estados e macroestados, respectivamente.

Os N_{fs} macroestados da matriz de transcrição Γ são ordenados de acordo com as anotações da figura 5.10:

- estado inicial não emissor R_0;

- macroestados representando os modelos subpalavra dos sucessivos conjuntos

$$R_1, R_2, \ldots, R_{N_f};$$

- estado final não emissor $R_{(N_f+1)}$.

Em cada conjunto da série R_n os macroestados são concatenados por ordem alfabética dos símbolos dos fones.

A ordem para os $N_f.N_s.N+2$ estados na matriz de transição A_{tr} decorre da ordenação anterior, onde se substituem os macroestados pelos estados de cada modelos subpalavra (N=3 no presente estudo). Para cada modelo subpalavra, estes estados são ordenados de acordo com a ordem temporal implícita na respectiva topologia linear (figura 5.4).

A partir dos índices dos macroestados utilizados na matriz Γ ($2 \leq i \leq N_{fs}-1$) definem-se duas funções auxiliares:

$$set(i) = int[(i-2)/N_s] + 1$$

$$fon(i) = i - 1 - [set(i) - 1].N_s$$

Estas funções determinam, respectivamente, o índice $j = set(i)$ do conjunto R_j ($1 \leq j \leq N_f$) e o índice $k = fon(i)$ do modelo subpalavra $\lambda^{(k)}$ ($1 \leq k \leq N_s$) associados a cada índice i da matriz Γ. Para a construção da primeira linha de Γ utilizaram-se as expressões seguintes:

$$\Gamma_{1,j} = \begin{cases} \frac{\varepsilon^{set(j)}}{N_s}, & 2 \leq set(j) \leq N_f; \\ \frac{1-\varepsilon-\varepsilon^2+\varepsilon^{N_f+1}}{N_s(1-\varepsilon)}, & set(j) = 1; \end{cases}$$

A probabilidade de se transitar entre os conjuntos R_1 e R_j foi determinada a partir da probabilidade de se transitar entre R_1 e R_2, afectada de um factor de atenuação exponencial ε =0,5. O valor deste factor foi fixado de forma arbitrária e é, obviamente, passível de correcção por via empírica ou outra. Contudo, tal não se veio a justificar, uma vez que os resultados obtidos a posteriori foram satisfatórios. A transição para cada um dos modelos subpalavra, dentro de cada um destes conjuntos, é equiprovável (factor N_s no denominador). As probabilidades de transição para o primeiro conjunto R_1 são determinadas por forma a que se verifique

$$\sum_{j=1}^{N_{fs}} \Gamma_{i,j} = 1, \qquad 1 \leq i \leq N_{fs}. \tag{5.2}$$

As restantes linhas correspondem às probabilidades de transição de saída dos modelos subpalavra. Consideram-se agora os valores na diagonal principal de Γ. Uma vez que

não se pretende que as unidades subpalavra se repitam no mesmo conjunto R_i de modelos subpalavra, esta probabilidade deveria ser nula. Contudo, Γ constitui uma matriz auxiliar para a construção da matriz das probabilidades de transição A_{tr}. Nesta matriz verifica-se que os valores de Γ surgem exclusivamente na primeira e última linha, e nas linhas associadas ao último estado de cada modelo subpalavra. A transição para fora do modelo subpalavra é feita a partir desse estado e fica condicionada à possibilidade desse estado transitar para si próprio (no caso presente $A_{3,3}^{(i)}$). Assim, os modelos subpalavra que lhe sucedam deverão partilhar a restante probabilidade. De acordo com o anteriormente exposto, obtêm-se os restantes elementos da matriz de transcrição para i ($2 \leq i \leq N_{fs}-1$):

$$\Gamma_{i,j} = \begin{cases} 0, & j < i; \\ A_{3,3}^{(fon(i))}, & j = i; \\ 0, & set(j) = set(i) \text{ e } j \neq i; \\ (1 - A_{3,3}^{(fon(i))}) \frac{\varepsilon^{[set(j)-set(i)]}}{N_s}, & set(j) > set(i) \text{ e } j \neq N_{fs}; \\ \frac{1-2.\varepsilon+\varepsilon^{[N_f-set(i)+1]}}{1-\varepsilon}, & j = N_{fs}; \end{cases}$$

As probabilidades de transição para o estado final $\Gamma_{i,N_{fs}}$ são calculadas por forma a que se verifique a equação (5.2).

5.5.4 Limitação do número de macroestados

A partir do modelo de transcrição λ_{tr} é possível gerar $N_s^{N_f} = 97336$ transcrições com a topologia representada na figura 5.9. Com a topologia efectivamente utilizada (figura 5.10) este número é ainda maior. A utilização deste modelo para palavras com um número superior de fones N_f, implica custos computacionais exagerados que podem ser substancialmente reduzidos. De facto, é possível verificar a posteriori que a maioria dos estados do modelo de transcrição não são determinantes no desempenho do modelo.

A probabilidade no instante t do estado q_i do modelo λ $f(s_t = q_i|O^n, \lambda)$ é obtida a partir da equação 2.22. O modelo HMM λ foi estimado com r locuções $O^R = \{O^1, \ldots, O^r\}$. A média do número de observações emitidas pelo estado q_i é dada por:

$$\bar{n}_o(q_i) = \frac{1}{r} \sum_{n=1}^{r} \sum_{t=1}^{T_n} f(s_t = q_i|O^n, \lambda). \tag{5.3}$$

Por sua vez, considera-se Q_j o j-ésimo macroestado, de acordo com a ordem definida na secção anterior. A média do número de observações emitidas por um macroestado

$Q_j = \{q_{1,j}, \ldots, q_{N,j}\}$ é determinada a partir das médias equivalentes dos respectivos $N = 3$ estados internos:

$$\bar{n}_M(Q_j) = \sum_{i=1}^{N} \bar{n}_o(q_{i,j}). \tag{5.4}$$

Para o reconhecimento consideraram-se modelos de transcrição com um subconjunto $\mathcal{Q}$ dos macroestados iniciais:

$$Q_j \in \mathcal{Q} \Leftarrow \bar{n}_M(Q_j) \geq \zeta. \tag{5.5}$$

Em geral, mais de 60% dos macroestados apresentam $\bar{n}_M = 0$. Uma vez que a média dos valores $\bar{n}_o > 0$ é de cerca de 0,2, utilizou-se ζ=0,05. Considerando apenas os macroestados com um valor de $\bar{n}_M$ acima deste limiar, conseguem-se eliminar cerca de 80% dos estados iniciais nos modelos treinados exclusivamente com locuções de oradores nativos.

5.5.5 Uso do conceito de ligação de parâmetros

A existência de informação repetida na estrutura do modelo de transcrição, (equação 5.1) pode ser explorada para uma implementação mais eficiente, utilizando o conceito de ligação de parâmetros (subsecção 2.4.4). Os programas do pacote HTK, pelo menos na versão aqui utilizada (Young et al., 1996), não permitem determinar selectivamente quais os parâmetros que podem ser partilhados. Essa selecção podia ser feita em relação a classes globais de parâmetros tais como as probabilidades de transição, as médias ou covariâncias das misturas, etc. mas não a cada elemento dentro dessas classes. Para a obtenção dos resultados que seguidamente se apresentam utilizaram-se ainda as rotinas básicas do HTK. Contudo, foi necessário o desenvolvimento de uma quantidade substancial de código por forma a implementar o presente paradigma.

Expansão do modelo para oradores de ambos os sexos

É reconhecida a vantagem do uso de modelos de fala separados para cada sexo (secção 6.3). Tal facto pressupõe diferenças a níveis acústico-fonéticos inferiores e não ao nível da transcrição fonémica ou de outros níveis superiores da construção da fala, tais como o sintáctico ou semântico. Pretende-se, por isso, utilizar todo o material disponível num corpus que foi equilibrado em termos de representação de ambos os sexos, por forma a obter modelos de transcrição independentes em termos desta característica dos oradores.

Como ponto de partida dispõe-se de dois modelos de transcrição inicializados com os modelos subpalavra referentes a cada sexo. Pretende-se associar estes dois modelos de

transcrição por forma a competirem em paralelo. Tal solução corresponderia à adopção de uma topologia com o dobro de parâmetros e da carga computacional anterior. Além destes aspectos negativos, obter-se-ia uma duplicação inútil do número de transcrições pretendidas, sem se aumentar a consistência das respectivas estimativas.

Pode-se, uma vez mais, eliminar estes inconvenientes com a utilização do conceito de ligação de parâmetros (subsecção 2.4.4). As probabilidades de transição entre macroestados de cada um dos dois modelos de transcrição são ligadas duas a duas, isto é, cada uma com a respectiva homóloga no modelo do sexo posto. Assim, as locuções de treino afectam estas transições independentemente do sexo do orador.

5.5.6 Resultados de reconhecimento

Os resultados de reconhecimento obtidos com as redes de transcrição encontram-se representados nas seis últimas linhas da tabela 5.1 (designadas por *rede de fones*). Conjuntamente, apresentam-se alguns resultados comparáveis, obtidos com os modelos HMM convencionais, referentes às figuras 5.2 e 5.6 (as duas primeiras linhas de valores (palavra) e as quatro seguintes, respectivamente (fones)). Os valores apresentados referem-se a taxas de reconhecimento (%) e foram obtidos exclusivamente com oradores masculinos.

Na tabela 5.1, cada linha corresponde aos resultados obtidos com um reconhecedor ou conjunto de reconhecedores diferente das restantes linhas. De seguida, explica-se o significado das primeiras quatro colunas destinadas à descrição de cada um destes reconhecedores. A primeira coluna define o tipo de modelos utilizados em cada reconhecedor: palavra, fone (para os modelos subpalavra) e redes de fones (para o modelo de transcrição). A segunda coluna indica a forma como são determinadas as transcrições, no caso dos reconhecedores baseados em modelos subpalavra. A terceira coluna indica qual o material de fala utilizado no treino das probabilidades de transição (apenas as de intrafones no caso dos modelos de transcrição) e das densidades de probabilidade de observação. O número de componentes gaussianas utilizadas nestas densidades encontra-se na quarta coluna. Sublinha-se o facto de a designação de *não nativo*, utilizada nesta tabela, se referir a um único grupo de oradores com um determinado sotaque e não a todos os oradores estrangeiros disponíveis simultaneamente. Assim, se numa dada linha alguma destas colunas apresentar a designação de *não nativo*, utilizam-se reconhecedores específicos para cada sotaque testado. No caso contrário será utilizado um único reconhecedor. Nas seis colunas seguintes apresentam-se os resultados obtidos com cada grupo de oradores. A última coluna representa o valor da média ponderada pelo número de oradores utilizados em cada grupo. Uma vez que estes números são aproximadamente iguais, este valor é

modelos	transcr.	observ.	M	da	de	en	es	it	pt	média
palavra	não utiliza	nativos	1	71,4	19,3	98,6	54,1	35,1	56,2	53,2
		não nat.	3	86,6	76,3	98,3	89,9	81,6	92,1	86,9
fones	léxico	nativos	3	70,4	11,0	94,8	60,8	32,3	62,1	52,9
			6	62,1	15,4	92,0	58,2	29,1	60,6	50,6
		não nativos	3	84,1	66,9	94,8	83,2	63,2	91,4	79,3
			6	85,1	72,6	92,0	81,7	68,1	90,1	80,5
rede de transcr.	nativos	nativos	3	69,7	40,8	95,3	57,5	41,6	62,8	59,0
			6	65,0	27,4	92,5	50,8	42,3	65,0	54,8
	não nativos		3	76,3	58,6	95,3	79,5	63,0	78,0	74,1
			6	74,1	57,5	92,5	77,8	63,3	80,5	73,3
		não nativos	3	83,1	61,4	95,3	82,5	70,5	87,9	79,1
			6	88,0	72,6	92,5	84,9	78,1	90,9	83,2

Tabela 5.1: Taxas de reconhecimento (%) (da 5ª à última coluna) obtidas com diferentes modelos: tipo de modelo (1ª coluna); treino das probabilidades de transição interfones (2ª coluna); treino das funções densidade de probabilidade de observação e das probabilidades de transição intrafones (3ª coluna); número de componentes gaussianas utilizadas no modelamento das referidas funções densidade de probabilidade (4ª coluna) (Teixeira et al., 1997).

muito próximo da média aritmética.

Devido aos elevados custos computacionais associados à utilização do modelo de transcrição e tendo por referência os melhores resultados obtidos com os reconhecedores de modelos subpalavra convencionais, (figuras 5.6 e 5.7) realizaram-se experiências apenas com três e seis componentes gaussianas por estado. Além dos referidos resultados da secção 5.3, incluíram-se ainda na tabela 5.1 alguns resultados da secção 5.2 (figuras 5.2 e 5.3). Na análise dos novos valores obtidos considere-se, em primeiro lugar, a utilização exclusiva de material de treino dos oradores nativos. O modelo de transcrição apresenta, neste caso, valores ligeiramente superiores aos obtidos quer com modelos subpalavra, ou mesmo com os modelos de palavra. A explicação provável para tal facto é a de que o modelo de transcrição consegue reunir as vantagens que são exclusivas de cada um dos modelos anteriores. Por um lado, os modelos subpalavra são treinados com mais repetições do que os modelos de palavra. Por outro, ao contrário dos reconhecedores subpalavra

convencionais, conseguem-se integrar várias transcrições alternativas da mesma palavra num só modelo. Em certa medida, o mesmo acontece nos modelos de palavra, embora de forma implícita.

Consideram-se agora as experiências em que se dispôs de material não nativo para o treino das probabilidades de transição interfones. Os ganhos de desempenho obtidos são claros: em média cerca de 20% na taxa de reconhecimento. Estes ganhos suportam o pressuposto inicial de que os oradores estrangeiros apresentam transcrições específicas diferentes dos oradores nativos.

Comparando finalmente os resultados obtidos com reconhecedores específicos para cada sotaque, para os quais se dispôs de material de treino adequado, verifica-se que o modelo de transcrição apresenta (com seis componentes gaussianas) um desempenho intermédio: ligeiramente superior aos reconhecedores subpalavra convencionais; um pouco inferior aos modelos de palavra. Este resultado está de acordo com o facto do reconhecedor com modelos de transcrição partilhar características dos reconhecedores dependentes do vocabulário (as probabilidades de transição interfones são treinadas com o mesmo vocabulário) e dos reconhecedores independentes do vocabulário, uma vez que os modelos subpalavra não foram obtidos com o mesmo vocabulário.

Os ganhos de desempenho obtidos com o grupo de oradores nativos, foram menos significativos do que os dos oradores não nativos. Assim, se os oradores nativos não fossem considerados no cálculo da média ponderada, (última coluna da tabela 5.1) seriam ainda mais evidentes os ganhos de desempenho obtidos com os oradores não nativos. Estes resultados indiciam, no essencial, a necessidade de se dispor de léxicos de multipronúncia, eventualmente com alguma descrição de índole probabilística, em particular, no caso do reconhecimento de oradores estrangeiros. Existem disponíveis alguns léxicos deste tipo para oradores nativos (subsecção 2.5.5). Procedimentos automáticos, tais como o baseado no modelo de transcrição, poderão facilitar a obtenção destes léxicos para outros grupos de oradores que apresentem maior variabilidade de pronúncia.

5.5.7 Verificação do modelo

O modelo de transcrição foi desenvolvido com base no pressuposto de que os sinais de fala de oradores não nativos apresentam uma maior variabilidade nas transcrições subpalavra do que os de oradores nativos. Os resultados de reconhecimento apresentados na subsecção anterior permitiram verificar que este modelo parece ser, de facto, mais adequados aos sinais de fala de oradores não nativos do que os modelos HMM convencionais.

A estrutura do modelo de transcrição deveria ainda permitir identificar, de uma forma evidente, a referida variabilidade das transcrições. Para tal, pretende-se identificar algumas das transcrições mais prováveis, implícitas no modelo de transcrição. Estes elementos deverão permitir confirmar os pressupostos assumidos na concepção do modelo e a geração de léxicos multipronúncia a partir de grupos específicos de oradores. Estes léxicos poderão ser utilizados, com vantagem, nos reconhecedores independentes do vocabulário mais tradicionais.

Geração de transcrições a partir da matriz de transcrição

nativa				não nat. c/ modelos nat.				não nativa			
probab.	transcrição			probab.	transcrição			probab.	transcrição		
3,42e-03	n	ow	sil	5,56e-03	n	aw	ao	1,46e-02	n	ow	
	0,84	0,62	0,80		0,72	0,15	0,75		0,38	0,64	
9,07e-04	n	er	uw	5,43e-03	n	aa	ao	1,07e-02	n	ow	sil
	0,84	0,00	0,78		0,72	0,14	0,75		0,38	0,64	0,47
5,47e-04	m	ow	sil	4,33e-03	n	aw	el	8,61e-03	n	ow	
	0,62	0,62	0,80		0,72	0,15	0,57		0,38	0,78	
4,93e-04	n	uw	sil	3,35e-03	n	aa		8,50e-03	n	ow	
	0,84	0,81	0,80		0,72	0,14			0,27	0,78	
4,55e-04	m	b	ow	3,25e-03	n	aw	w	8,25e-03	n	uw	sil
	0,62	0,04	0,86		0,72	0,15	0,63		0,38	0,62	0,47

Tabela 5.2: Transcrições geradas a partir da matriz de transcrição da palavra inglesa "no".

Uma transcrição pode ser descrita por um conjunto de elementos subpalavra ordenados numa determinada sequência. Considerem-se as transcrições com $n = N_t$ elementos que foram numerados da forma utilizada para a matriz Γ (de 2 a N_s+1):

$$\theta_{N_t} = \{f_1, f_2, \ldots, f_{N_t}\}.$$

É possível estabelecer-se uma relação entre θ_n e uma sequência de elementos da matriz Γ

$$\Gamma_{\theta_n} = \{\Gamma_{1,f_1}, \Gamma_{f_2,f_3}, \cdots, \Gamma_{f_n,N_{fs}}\}.$$

De acordo com esta descrição é possível calcular a probabilidade condicional:

$$Pr(\theta_n|\Gamma_{\theta_n}) = \Gamma_{1,f_1}\Gamma_{f_n,N_{fs}}\prod_{i=1}^{n-1}\Gamma_{f_i,f_{i+1}}. \tag{5.6}$$

Desta forma é possível determinar uma transcrição a partir de

$$\arg\max_{\theta_n:1\leq n\leq N_f} Pr(\theta_n|\Gamma_{\theta_n}).$$

Os valores da diagonal principal Γ_{f_i,f_i} não foram utilizados na expressão 5.6. Em contrapartida, podem ser representados em conjunto com transcrições obtidas por este processo. As transcrições apresentadas em cada coluna das tabelas 5.2 e 5.3 são referentes aos cinco valores mais elevados obtidos para $Pr(\theta_n|\Gamma_{\theta_n})$, considerando todos os valores de n tal que $1 \leq n \leq N_f$.

Representação gráfica da matriz de transcrição

De acordo com o exposto nos parágrafos anteriores e com os resultados exemplificados nas tabelas 5.2 e 5.3, a matriz Γ contém informação relevante sobre o modelo de transcrição. Contudo, devido à sua dimensão, esta matriz é de difícil leitura (subsecção 5.5.4). Uma vez que Γ é uma matriz de transição entre macroestados, pode ser útil analisar uma representação gráfica semelhante às que são habitualmente utilizadas para os modelos baseados numa máquina de estados.

A etiquetagem numérica da probabilidade de cada transição interfones sobrecarregava a representação, não permitindo uma observação adequada. Procurou-se, por isso, uma forma de representação estritamente de natureza gráfica. Efectuou-se uma correspondência entre o valor de cada probabilidade de transição e uma graduação de níveis de cinzentos com os quais se traçaram os ramos de transição. Assim, os ramos correspondentes às probabilidades de transição de valor mais elevado foram traçados a negro, enquanto que os referentes aos valores mais baixos tendem a diluir-se no fundo branco do papel (figura 5.11). A identificação das transcrições mais prováveis, deverá de ter em consideração, para além das probabilidades de transição entre macroestados diferentes, as probabilidades de transição para os próprios estados. Estas probabilidades informam da importância relativa da inserção ou da supressão do respectivo fone na transcrição, à semelhança do que foi feito nas tabelas 5.2 e 5.3.

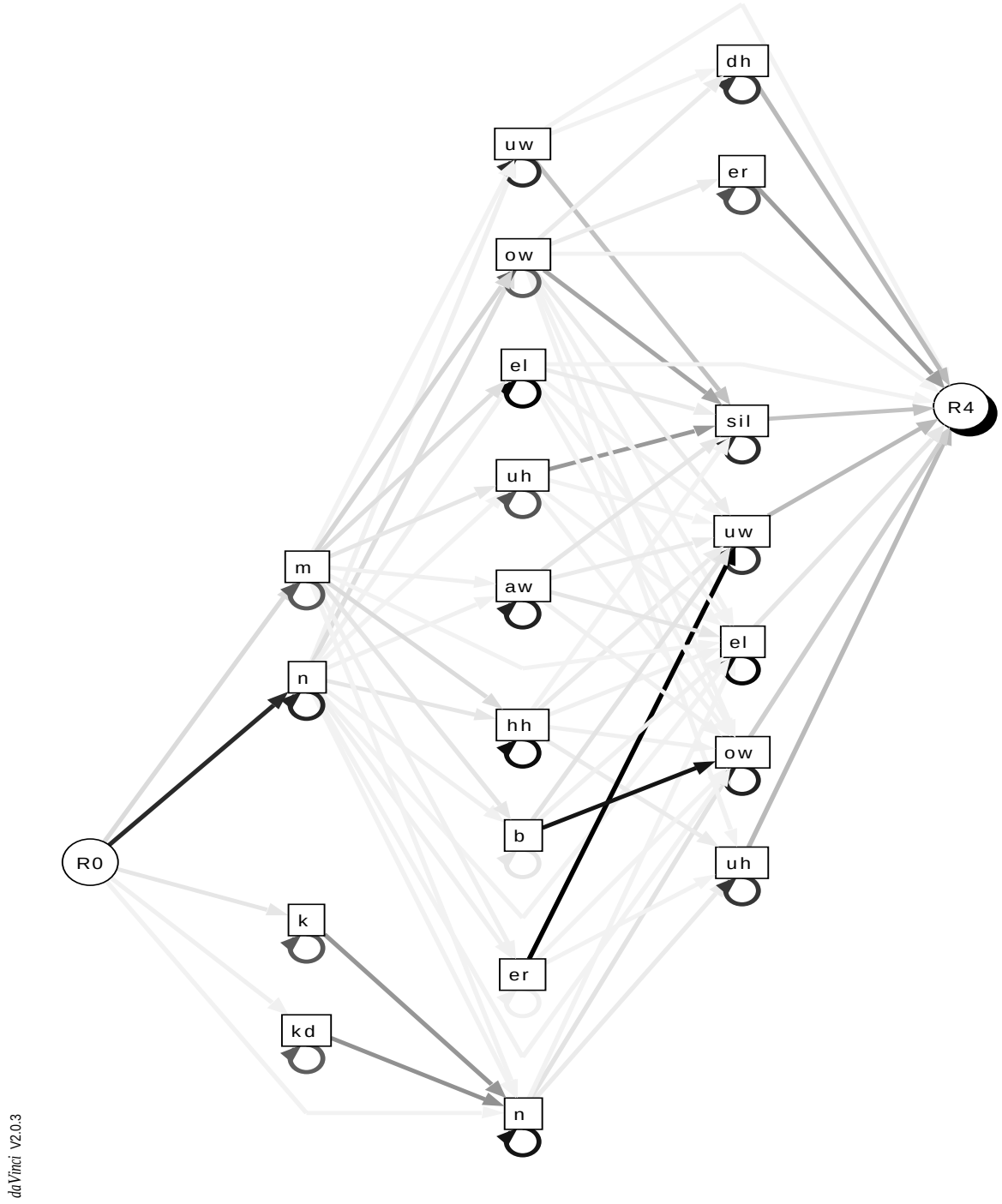


Figura 5.11: Rede probabilística obtida com o modelo de transcrição da palavra "no" treinada exclusivamente com oradores nativos (en) (Teixeira et al., 1997).

Análise de alguns exemplos

Inicialmente analisam-se os modelos de transcrição obtidos para uma palavra inglesa de curta duração: "no". Posteriormente, repete-se o mesmo tipo de análise com uma palavra com maior duração.

Na figura 5.11 representou-se o modelo obtido exclusivamente com sinais de fala de oradores nativos. Para a construção da figura 5.12 utilizaram-se os mesmos modelos de fones, contudo, as probabilidades de transição interfones foram treinadas com um subcorpus de fala de um grupo de oradores não nativos (es). O modelo de transcrição representado na figura 5.13 foi exclusivamente obtido a partir do referido subcorpus não nativo. Em cada coluna da tabela 5.2 representaram-se as cinco transcrições, implícitas em cada modelo, que apresentaram os valores mais elevados do produto das probabilidades de transição.

A primeira transcrição apresentada para o modelo dos oradores nativos corresponde à transcrição fonotípica nativa conhecida /n/ /ow/ com um silêncio no final. Pela análise da figura 5.11, verifica-se ser possível obter esta transcrição sem o referido silêncio, que é substituído por uma transição directa entre R_2 e R_4. Esta transcrição apresenta, contudo, uma probabilidade inferior, não se encontrando representada na tabela 5.2. Na verdade, o valor de probabilidade apresentado para a primeira transcrição demarca-se substancialmente dos restantes valores obtidos. A segunda transcrição apresenta a inserção do fone /er/ não tendo este, contudo, grande significado, pelo menos em termos de duração, uma vez que a respectiva probabilidade de transitar para si mesmo é muito baixa[5]. O mesmo se passa com a inserção da oclusiva /b/ na quinta transcrição apresentada. Verificam-se ainda algumas substituições indesejáveis entre as nasais /m/ e o /n/, entre as vogais /ow/ e /uw/ e outras de significado estatístico aparentemente inferior. Na figura 5.11 é possível identificar algumas transcrições, menos prováveis, em que ocorre uma inserção de uma consoante oclusiva inicial /k/ ou /kd/, a qual poderá ser devida ao surgimento de pequenos artefactos próprios do sinal de fala no início das locuções.

Na segunda coluna da tabela 5.2 apresentam-se algumas transcrições não nativas geradas com modelos subpalavra nativos. A primeira transcrição substitui o fone /ow/ na transcrição nativa conhecida, por dois fones consecutivos /aw/ e /ao/, em que o último apresenta maior duração. Partindo desta transcrição é possível descrever mais facilmente as seguintes em termos de substituições: o fone /aw/ é substituído na segunda e quarta transcrição por /aa/ e o fone /ao/ por /el/ na terceira e por /w/ na quinta transcrição.

[5] O valor apresentado na tabela 5.2 (0,00) resulta de arredondamento, sendo, obviamente, superior a zero.

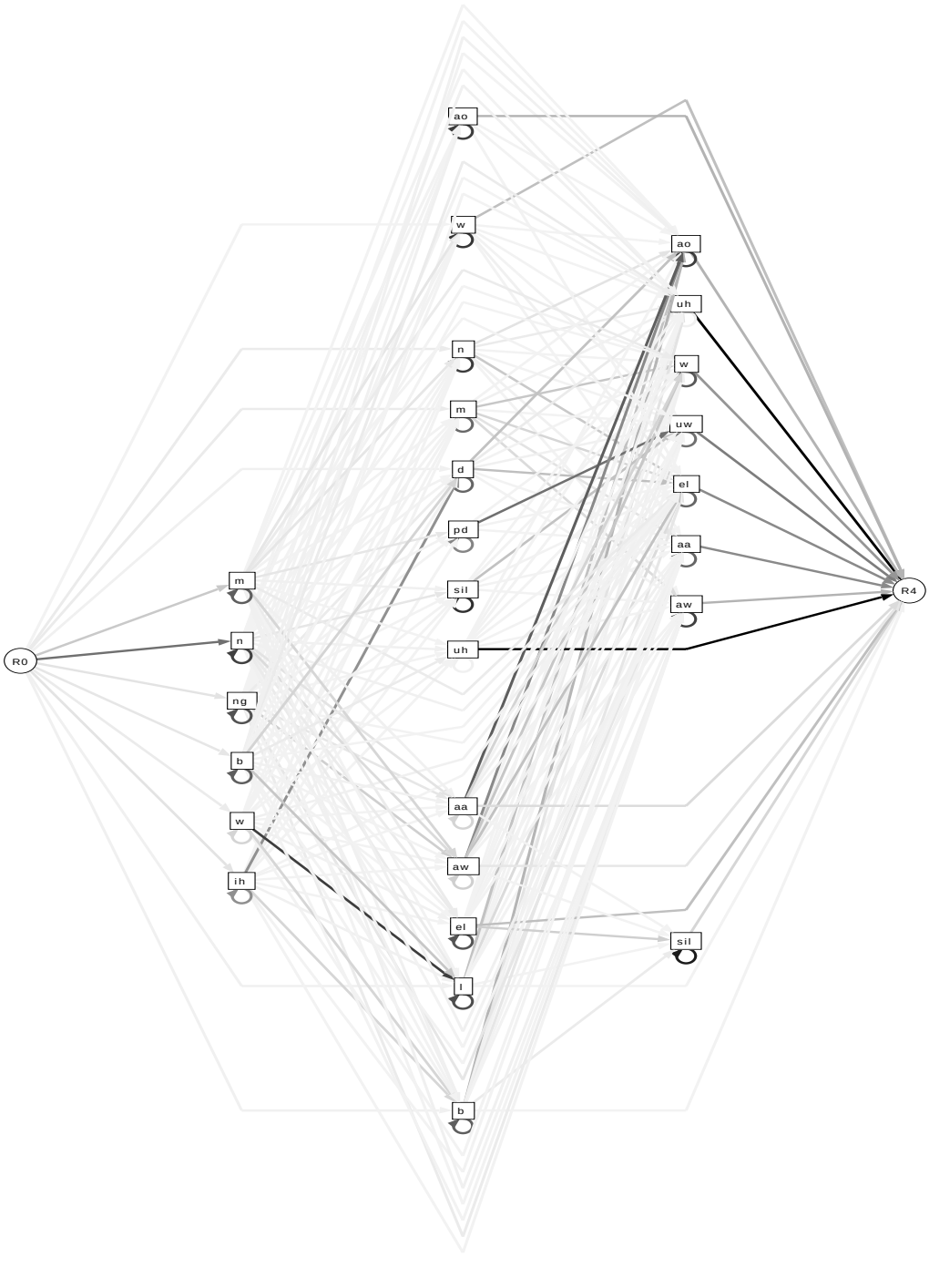


Figura 5.12: Rede probabilística obtida com o modelo de transcrição da palavra "no". Os modelos subpalavra foram treinados com os oradores nativos (en) e as transições entre estes foram treinadas com os oradores não nativos (es) (Teixeira et al., 1997).

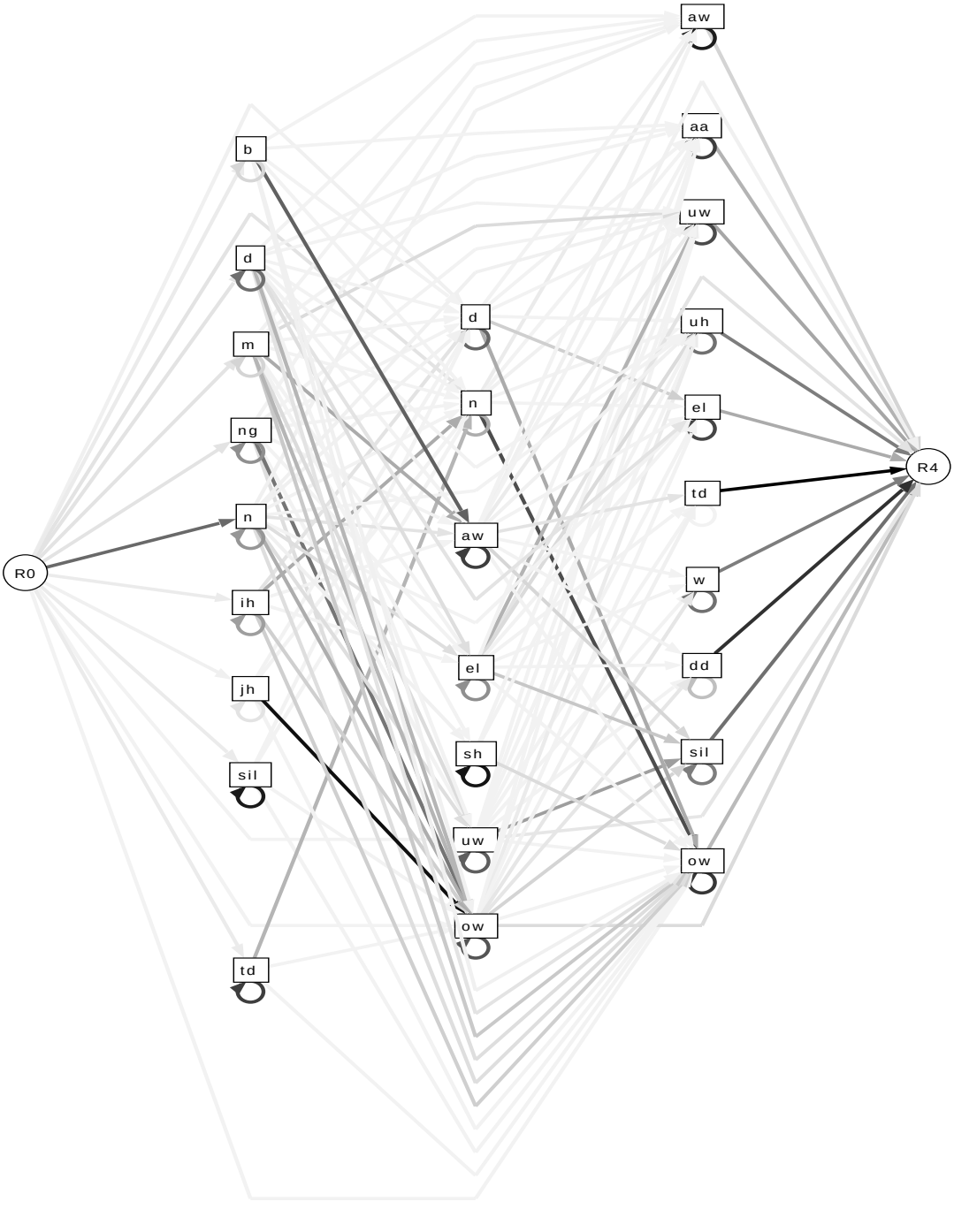


Figura 5.13: Rede probabilística obtida com o modelo de transcrição da palavra "no" treinada exclusivamente com oradores não nativos (es) (Teixeira et al., 1997).

Com excepção da inserção do fone /el/ todas estas transcrições podem ser verificadas, de forma aproximada, por audição das locuções que lhes deram origem. De realçar que, ao contrário das transcrições nativas, as probabilidades calculadas para cada uma delas não diferem de forma significativa. Aliás, a lista de todas as transcrições obtidas (apenas parcialmente aqui representada) apresenta o mesmo decréscimo gradual desta probabilidade. Este facto é evidenciado na apreciação da figura 5.12, quando comparada com a anterior (figura 5.11): existe maior dispersão de trajectos alternativos através de um número superior de macroestados (27 em vez de 20 macroestados). Este efeito pode ser atribuído à maior variabilidade das transcrições subjacentes aos sinais de fala não nativos ou à desadaptação provocada pelos modelos subpalavra treinados com outro grupo de oradores. A análise da figura 5.13 parece confirmar o contributo destes dois factores, em proporções semelhantes. De facto, no modelo de transcrição correspondente elimina-se o segundo factor, uma vez que este modelo é inteiramente obtido a partir do subcorpus de fala não nativa e determina um diagrama ligeiramente mais simples (26 em vez de 27 macroestados). A complexidade acrescida em relação à primeira figura é justificada pelo facto de a fala não nativa apresentar em geral maior variabilidade em relação às transcrições da fala nativa.

Considerando agora o modelo de transcrição exclusivamente treinado com locuções de oradores não nativos, reencontra-se a transcrição fonotípica nativa /n/ /ow/ através de diferentes trajectos incluindo ou não o silêncio final. A quinta transcrição obtida apresenta, tal como acontecia com as transcrições nativas, uma substituição do fone /ow/ pelo fone /uw/. Tal como acontece com as transcrições representadas na segunda coluna da tabela 5.2, as probabilidades calculadas não diferem entre si de forma significativa. Estes factos indicam que estes oradores não nativos cometem pequenos erros de pronúncia com o ditongo /ow/, substituindo-o de forma consistente em diversos contextos. Esta consistência refere-se aos modelos subpalavra independentes do contexto, que integram os erros cometidos noutras palavras que incluem o mesmo fone.

Analisam-se de seguida algumas transcrições geradas para uma palavra um pouco mais longa: a palavra inglesa "undo". A primeira transcrição obtida com as locuções nativas, tal como aconteceu com a palavra "no", é igual à do léxico de pronúncia apresentado no apêndice A, acrescida de um silêncio final. De acordo com primeira coluna da tabela 5.3, também esta primeira transcrição se demarca, em termos do valor de probabilidade assinalado, das restantes transcrições obtidas. Treinando as transições interfones com locuções não nativas, as respectivas transcrições apresentam essencialmente substituições das vogais /ah/ e /uw/, sendo esta última, por vezes, acompanhada pela supressão da consoante /d/. A vogal /ah/ é substituída, pelo menos nas primeiras quarenta transcri-

ções determinadas, (não representadas na tabela) pela vogal /ae/. A consoante nasalada /n/ é por vezes substituída pela sua congénere /ng/. Em relação à vogal /uw/, a respectiva substituição é feita pela associação dos fones /aa/ e /ng/, ou, no caso da supressão da consoante /d/, pela associação dos fones /ax/ e /el/ ou /w/ e /ng/.

nativa						não nat. c/ modelos nat.						não nativa					
probab.	transcrição					probab.	transcrição					probab.	transcrição				
1,51e-3	ah	n	d	uw	sil	3,79e-3	ae	n	d	aa	ng	4,74e-3	ae	n	d	uw	sil
	0,48	0,06	0,82	0,75	0,85		0,53	0,13	0,50	0,20	0,06		0,49	0,48	0,17	0,60	0,23
9,81e-4	t	aa	n	d	uw	3,72e-3	ae	n	ax	el		2,71e-3	ae	n	t	uw	sil
	0,64	0,24	0,06	0,58	0,87		0,53	0,13	0,04	0,24			0,49	0,48	0,11	0,60	0,23
8,01e-4	ah	ng	dd	ix	uw	2,56e-3	ae	ng	d	aa	ng	2,35e-3	ae	n	l	uw	sil
	0,48	0,01	0,29	0,00	0,87		0,53	0,06	0,50	0,20	0,06		0,49	0,48	0,10	0,60	0,23
7,83e-4	ah	n	ax	uw	sil	2,44e-3	ae	n	ax	el		2,09e-3	ae	ng	d	uw	sil
	0,48	0,06	0,33	0,75	0,85		0,53	0,16	0,04	0,24			0,49	0,51	0,17	0,60	0,23
7,58e-4	ah	n	dd	ix	uw	2,41e-3	ae	n	w	ng		1,81e-3	ae	n	l	uh	
	0,48	0,06	0,29	0,00	0,87		0,53	0,13	0,17	0,06			0,49	0,48	0,02	0,38	

Tabela 5.3: Transcrições geradas a partir da matriz de transcrição da palavra inglesa "undo".

Considerando agora o modelo de transcrição exclusivamente treinado com locuções de oradores não nativos, verifica-se a obtenção de uma primeira transcrição com quase o dobro da probabilidade das transcrições que lhe sucedem. Ao contrário do que sucedeu com o ditongo /ow/ na palavra "no", a vogal inicial /ae/ permanece agora em substituição da vogal /ah/ da transcrição nativa. Por outro lado, a vogal /uw/ reaparece seguida de silêncio, apenas pontualmente substituído pela vogal /uh/. A consoante /n/ mantém-se, sendo embora por vezes substituída pela sua congénere nasal /ng/. A oclusiva alveolar /d/ perde por vezes o vozeamento, transformando-se na sua congénere /t/. Uma alternativa vozeada surge de outra consoante alveolar, a líquida /l/. Todas estas substituições constituem alterações de pronúncia que poderão advir da proximidade fonética dos fones substituídos.

De forma semelhante à efectuada em relação ao ditongo /ow/ na palavra "no", pode-se agora concluir que a vogal /uw/ é consistentemente substituída por este grupo de oradores não nativos por /aa/ /ng/, ou por /ax/ /el/ ou /w/ /ng/ com supressão da

consoante /d/. Pelo contrário, a substituição da vogal /ah/ por /ae/ não deverá ocorrer, pelo menos de forma sistemática, em outras palavras diferentes da aqui considerada, uma vez que aparenta ocorrer apenas ao nível da transcrição e não da troca de características dos respectivos modelos de fones (tabela 5.3).

Analisem-se as figuras referentes aos modelos de transcrição da palavra "undo" (figuras 5.14, 5.16 e 5.15). A complexidade destes modelos pode ser quantificada em termos do número de transcrições possíveis ou do número de macroestados e de transições entre estes. Sobressai o facto da figura 5.16 (29 macroestados) apresentar uma complexidade intermédia entre as figuras 5.14 e 5.15 (21 e 43 macroestados, respectivamente). Este facto já tinha sido detectado nas figuras referentes à palavra "no", mas é, no caso presente, mais evidente. O modelo de transcrição em que o material não nativo é utilizado exclusivamente para o treino das probabilidades de transição interfones é, portanto, representado pela figura mais complexa (5.15). Este modelo é o único que não gera a transcrição nativa conhecida, contudo, pertence ao grupo que produz os ganhos mais significativos nas taxas de reconhecimento apresentadas na tabela 5.1. Este facto atesta, uma vez mais, a desadaptação da transcrição nativa para o caso dos oradores estrangeiros.

A análise efectuada nos parágrafos anteriores para as palavras "no" e "undo" foi repetida em outras das palavras de teste (apêndice A.2). As transcrições e figuras obtidas não revelaram factos substancialmente diferentes dos aqui exemplificados para as duas palavras referidas.

A existência de silêncio é detectada de forma sistemática no final das locuções, nos modelos de transcrição da generalidade das palavras utilizadas. Tal facto é devido ao detector de início e fim de palavra que poderia ser sujeito a um pequeno ajuste por forma a determinar o final da palavra cerca de 20 a 30ms mais cedo. Também alguns macroestados iniciais terão sido inseridos ou, mais raramente, excluídos devido a erros na determinação do início da palavra. Surgem ainda, por vezes, algumas oclusivas, tais como /b/, /d/, /t/ e /k/, aparentemente induzidas pela existência de pequenos estalidos causados pela língua ou lábios do orador ("smacks").

Além dos referidos silêncios, as inserções de fones detectadas não são relevantes, uma vez que a probabilidade de durarem mais do que 10ms é muito baixa. Assim, uma alteração prevista para experiências futuras, consiste na diminuição da dimensão da sequência de fones N_f, à custa da introdução de modelos de silêncio antes e depois do modelo de transcrição propriamente dito (figura 5.17).

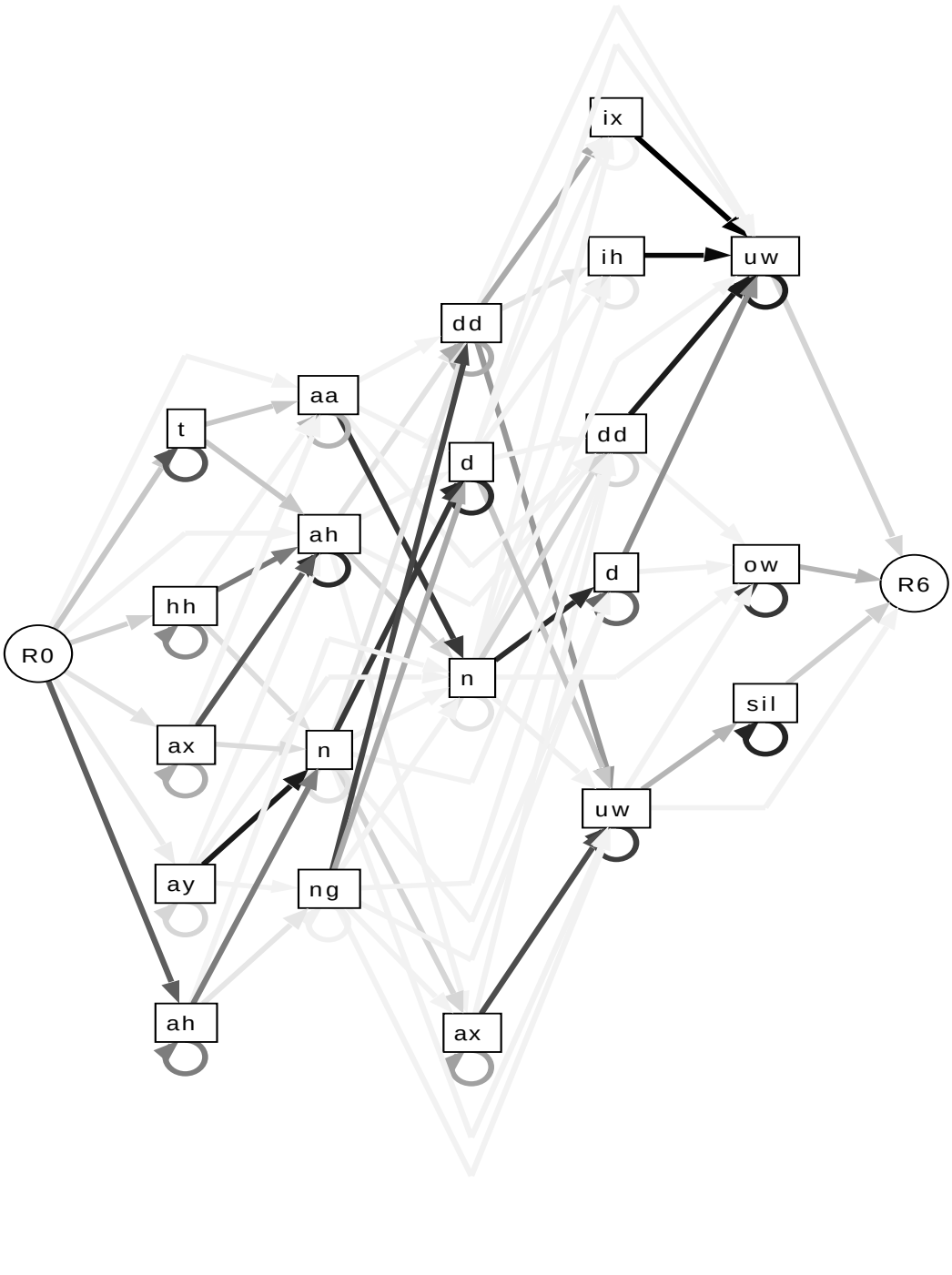


Figura 5.14: Rede probabilística obtida com o modelo de transcrição da palavra "undo" treinada exclusivamente com oradores nativos (en) (Teixeira et al., 1997).

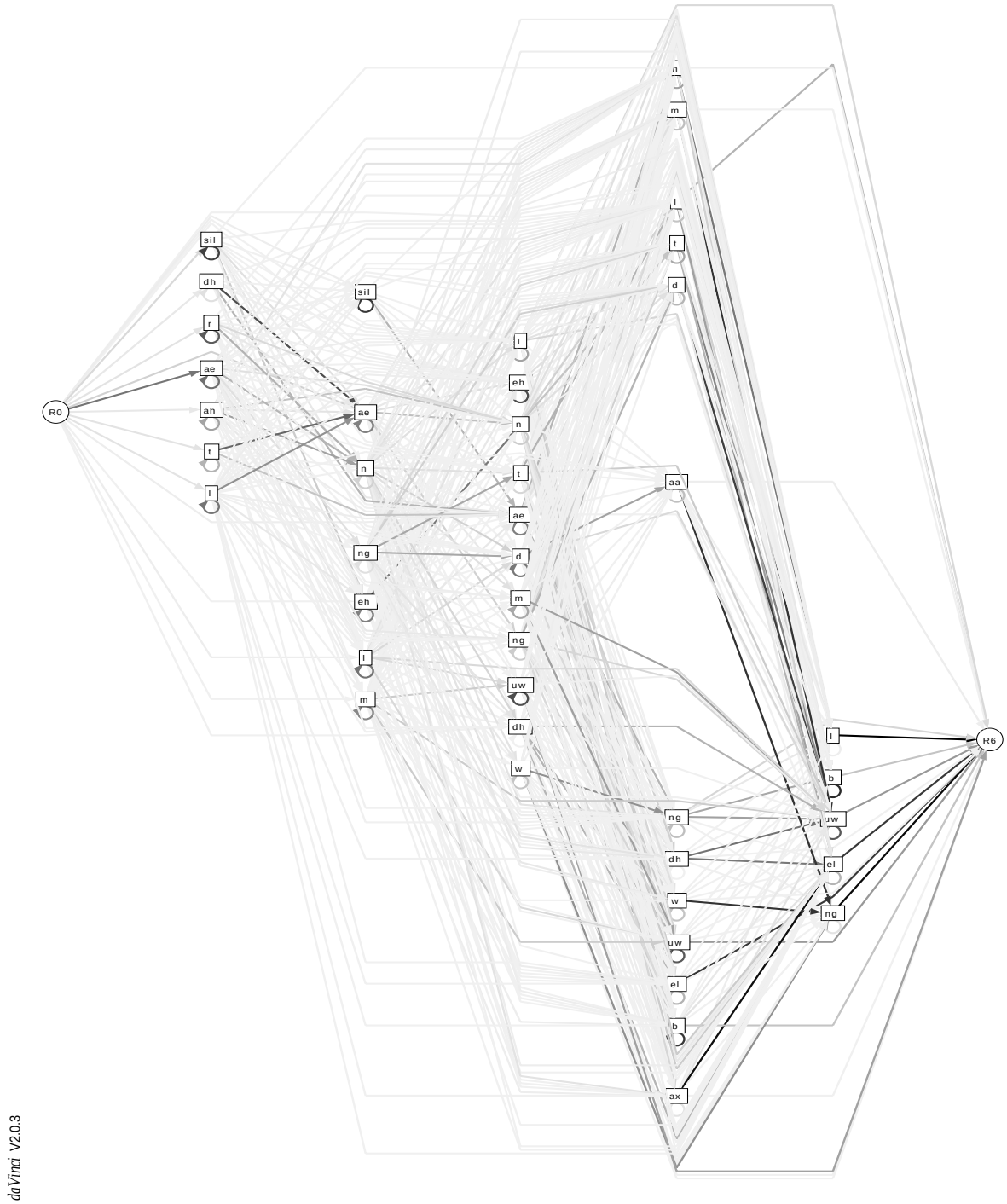


Figura 5.15: Rede probabilística obtida com o modelo de transcrição da palavra "undo". Os modelos subpalavra foram treinados com os oradores nativos (en) e as transições entre estes foram treinadas com os oradores não nativos (es) (Teixeira et al., 1997).

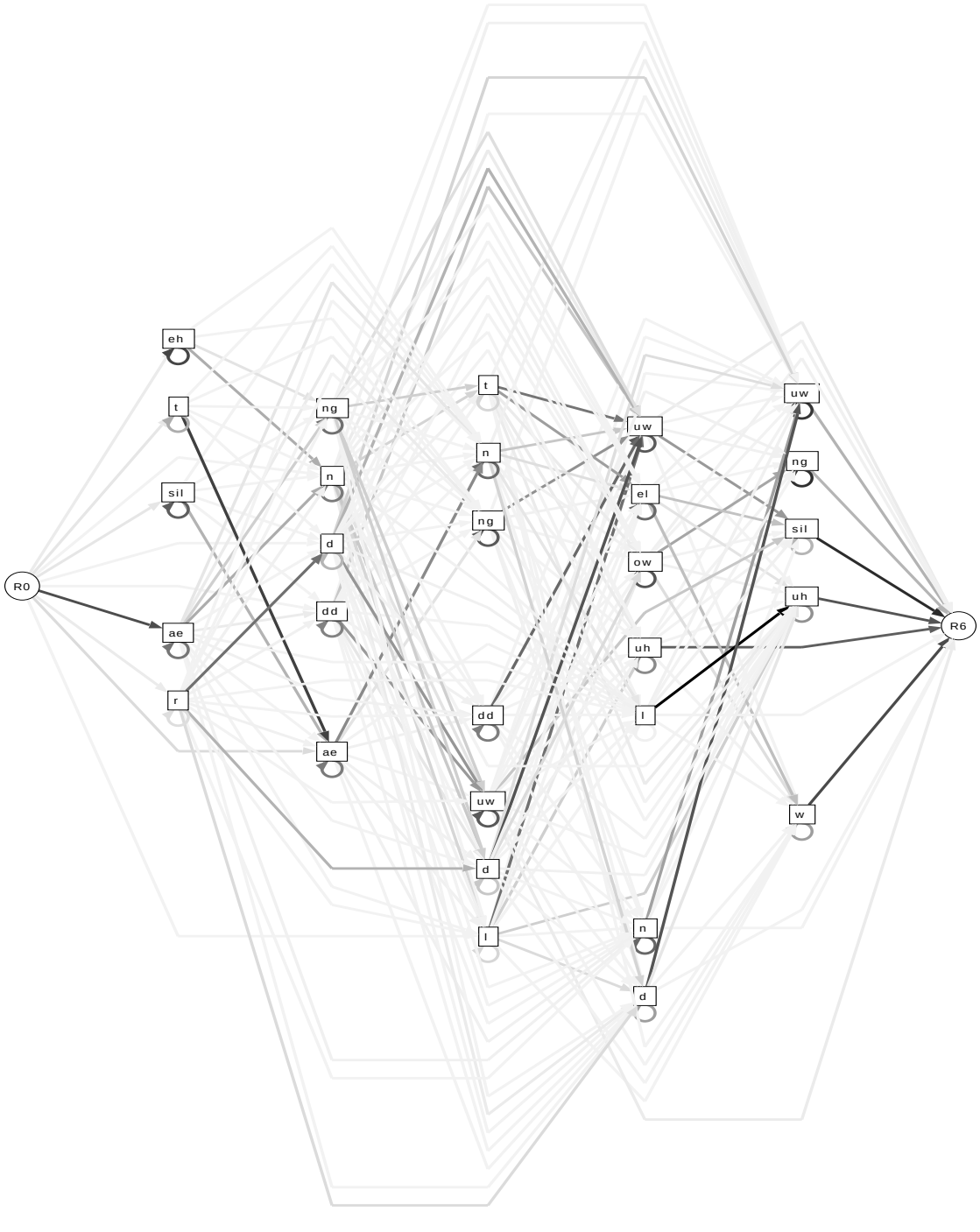


Figura 5.16: Rede probabilística obtida com o modelo de transcrição da palavra "undo" treinada exclusivamente com oradores não nativos (es) (Teixeira et al., 1997).

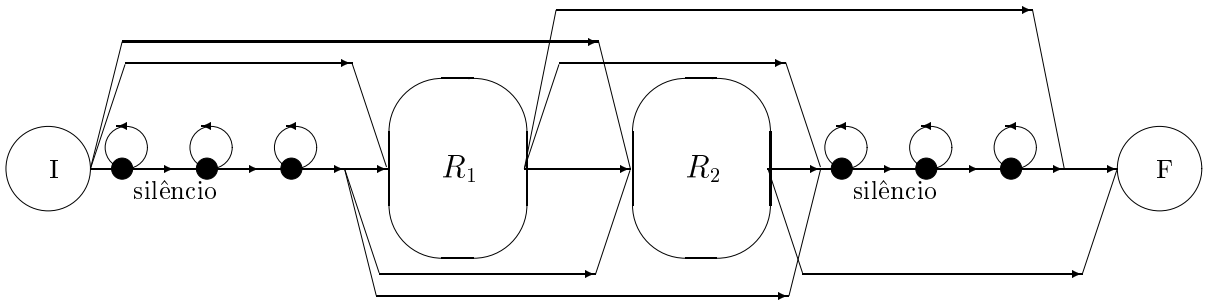


Figura 5.17: Modificação de um modelo de transcrição para palavras isoladas com dois fones, de modo a corrigir a detecção de início e fim de palavra.

Determinação das transcrições mais prováveis

Pretende-se agora determinar as transcrições mais prováveis representadas no modelo de transcrição. Considere-se uma sequência de estados $S = \{s_1, \ldots, s_T\}$ no modelo λ, tal que $S \in \mathcal{S}$ (secção 2.3.4). A probabilidade associada a esta sequência $Pr(S|\lambda)$ é dada pela equação 2.15.

A partir do modelo de transcrição é possível considerar diversos trajectos através dos respectivos estados. Cada trajecto é aqui definido pela sequência de n_θ estados que visita $\theta = \{q_1, \ldots, q_{n_\theta}\}$ (diferente da sequência de estados S definida para cada instante de tempo discreto). Cada um destes trajectos é equivalente a uma transcrição com N_t fones, tal que $n_\theta = 2 + N_t.N$ (com $N = 3$ estados). Conforme se referiu anteriormente, o estado inicial $q_1 = 1$ e o final $q_{n_\theta} = 2 + N_f.N_s.N$ são não emissores. Uma vez que os estados emissores do modelo de transição apresentam uma probabilidade não nula de transitarem para si próprios, o número de observações emitidas por cada um pode ser superior à unidade. A média do número de observações emitidas por cada estado ($\bar{n}_o(q_i)$) foi determinado na subsecção 5.5.4 (equação 5.3). Assim, a probabilidade de um trajecto θ ser gerada pelo modelo λ_{tr} pode ser determinada de forma semelhante à da probabilidade $Pr(S|\lambda)$ com a seguinte expressão

$$Pr(\theta|\lambda_{tr}) = a_{q_1,q_2} \prod_{i=2}^{n_\theta - 1} a_{q_i,q_i}^{\bar{n}_o(q_i)} . a_{q_i,q_{i+1}} \tag{5.7}$$

na qual se considera $\{a_{i,j}\} = A_{tr}$. Calculando $Pr(\theta|\lambda_{tr})$ para todos os trajectos possíveis no modelo λ_{tr}, pode determinar-se qual a transcrição mais provável: $\arg\max_{\theta \in \Theta} Pr(\theta|\lambda_{\mathrm{tr}})$.

A transcrição assim obtida, não é, por si só, de grande interesse, não sendo geralmente adequado adoptá-la como uma transcrição fonotípica. Assim, é possível determinar um número elevado de transcrições muito semelhantes às obtidas a partir da matriz Γ, (tabelas

5.2 e 5.3) mas com probabilidades $Pr(\theta|\lambda_{\mathrm{tr}})$ de valor muito próximo do máximo, mesmo nos modelos de transcrição nativos. Tal facto deve-se, sobretudo, aos motivos seguintes:

detecção imperfeita do início e fim da palavra — É habitualmente identificado um fone correspondente ao silêncio antes ou depois da palavra. Contudo, surge também com alguma frequência, a inserção ou substituição de uma consoante inicial ou final por outras. O surgimento de algumas consoantes, nomeadamente, fricativas e oclusivas, parece ser devido à existência de locuções de treino com pequenos ruídos, ou de outras, mais raras, que foram seccionadas excluindo pequenas porções do sinal de fala;

substituição e supressão de consoantes — Para além das consoantes iniciais ou finais, referidas no item anterior, outras consoantes são por vezes substituídas, embora com menos frequência. Estas substituições ocorrem sobretudo nas oclusivas. De facto, as consoantes têm em geral uma duração curta, quando comparadas com outros fones e são ainda por vezes articuladas de modo deficiente, em particular pelos oradores não nativos;

substituição de vogais — A substituição de vogais é menos frequente do que a de consoantes e quando ocorre é feita entre vogais com sons semelhantes. No caso dos oradores estrangeiros estas substituições ocorrem por vezes nas transcrições mais prováveis de alguns sotaques. Estes casos são corroborados pela audição das locuções de treino podendo assim serem considerados como características de determinado sotaque. Por exemplo menciona-se a substituição da vogal /ow/ na palavra "no" pela vogal /ao/, /aa/ ou /aw/, também verificada na tabela 5.2.

5.6 Conclusões

Neste capítulo, apresentaram-se resultados de experiências obtidas com diversos tipos de reconhecedores que foram conduzidas de forma a se poder, tanto quanto possível, comparar os respectivos resultados.

A necessidade de se estudarem modelos mais detalhados, nomeadamente a nível da transcrição das palavras em unidades elementares, conduziu ao desenvolvimento de um método para a determinação de um modelo probabilístico das transcrições fonotípicas. Este método constrói uma espécie de máquina probabilística com número finito de estados, que emite símbolos correspondentes às unidades subpalavra. O algoritmo de reestimação

Baum-Welsh revelou-se adequado ao treino deste modelo com diversas locuções de cada palavra.

Utilizando-se exclusivamente material de fala de oradores nativos no treino dos modelos subpalavra, obtiveram-se acréscimos de desempenho de mais de 20% com a utilização deste modelo no reconhecimento de fala de oradores não nativos.

Através da análise das transcrições e dos gráficos da subsecção 5.5.7, ambos gerados a partir do modelo de transcrição, é possível uma melhor compreensão da eficácia deste modelo para o reconhecimento de fala de oradores estrangeiros.

Capítulo 6

Identificação do sotaque do orador

6.1 Introdução

Este capítulo é dedicado ao problema da identificação automática do sotaque estrangeiro (Teixeira e Trancoso, 1993; Teixeira et al., 1996). Para se definir o contexto em que se pretende estudar este problema e consequentemente os objectivos deste capítulo, convém distinguir dois tipos de questões: a detecção do sotaque e a classificação ou identificação do sotaque.

Na detecção do sotaque estrangeiro, pretende-se determinar se um dado orador fala a sua língua materna ou uma segunda língua.

A identificação do sotaque deverá distinguir entre oradores, pertencentes a comunidades linguísticas com características específicas, que falam uma mesma língua. No âmbito dos sotaques estrangeiros, a que se refere o presente trabalho, cada uma destas comunidades linguísticas partilha uma primeira língua diferente das restantes comunidades. A solução deste problema resolve de forma implícita o problema da detecção do sotaque pelo que este último, por si só, aparenta menor dificuldade.

6.1.1 Identificação do sotaque no reconhecimento de fala

É um exercício comum para quem encontra um interlocutor a quem detectou um sotaque estrangeiro, tentar descobrir a sua nacionalidade. Este exercício toma como pressuposto um indivíduo com uma história pessoal típica: nasceu e viveu a maior parte do tempo no país de que herdou a nacionalidade e uma língua materna sem grandes variações dialectais. Contudo, esta história sofre geralmente grandes distorções. Por exemplo,

as línguas que se encontram representadas no corpus multissotaque SUNSTAR, têm importantes variações dialectais (essas variações foram, tanto quanto possível, limitadas). Existem também muitos indivíduos que por razões diversas, tais como hábitos nómadas, frequência de escolas de línguas durante a infância ou apenas o contacto sistemático com os meios de comunicação estrangeiros, não são passíveis de serem abrangidos por estes pressupostos.

É com base em pressupostos semelhantes que existem em certos países, nomeadamente em departamentos de criminologia da polícia alemã, especialistas na identificação de dialectos. O pressuposto anterior, no que se refere às variações dialectais, deverá agora ser alterado. São precisamente as diferenças dialectais, por vezes relacionadas com as regiões a que são afectas, que permitem um trabalho útil a estes especialistas. Assim, as limitações referidas para as variações em relação à língua materna são neste caso substituídas por limitações em relação a comunidades dialectais, necessariamente mais reduzidas. O problema da identificação do sotaque no contexto do presente trabalho tem objectivos diferentes. O objectivo principal é o de minimizar as perdas de desempenho de reconhecedores automáticos devidas a oradores que falam uma segunda língua.

No capítulo 5, verificou-se que um reconhecedor treinado com oradores representativos de um dado sotaque, apresentava prestações significativamente superiores aos obtidos com um reconhecedor treinado com oradores representativos de qualquer outro sotaque, nomeadamente o nativo. O problema que fica por resolver é o da selecção do reconhecedor mais adequado para um orador pertencente a uma população que inclui indivíduos com diversos sotaques. Portanto, não se pretende determinar em particular a nacionalidade, a língua materna ou a região onde determinado orador habita. Por um lado, qualquer destes dados, uma vez disponíveis, poderiam ser úteis na resolução deste problema. Considere-se, por exemplo, que se solicita ao orador para começar por dizer o nome da região onde habita. Contudo, tal não dispensa o uso de métodos adicionais para a determinação do reconhecedor adequado. Assim, na sequência deste exemplo, o orador pode ter mudado de residência há pouco tempo, sendo originário de uma região com hábitos linguísticos muito diferentes da região que indicou. Deste modo, seria seleccionado um reconhecedor eventualmente desadequado para este orador. Naturalmente que se pode elaborar mais este procedimento, solicitando adicionalmente outro tipo de dados. Não se pretende explorar neste trabalho este tipo de abordagem, a qual é desadequada para algumas situações típicas de interesse. Tal é o caso dos serviços de emergência que é apresentado na subsecção 6.1.2.

6.1.2 Identificação da língua versus sotaque

Numa situação ideal, um utilizador de um qualquer serviço de atendimento público internacional deveria poder utilizar a sua própria língua, ou seja, deveria dispor de um serviço multilíngua. De facto, uma parte significativa da população mundial desconhece outra língua além da sua língua materna e talvez ainda menos uma língua franca. Mesmo que a conheça, é sempre mais cómodo e menos sujeito a erros, utilizar a língua materna do que ter de conhecer, falar e perceber uma segunda língua. Estes factos parecem implicar uma desvalorização do interesse prático do problema dos sotaques. Contudo, não existem na prática sistemas com um número suficientemente grande de línguas de modo a eliminar uma percentagem significativa de oradores não nativos. Além disso, a maioria dos serviços com reconhecimento automático de fala são ainda sistemas muito simples que trabalham com uma única língua.

Nos serviços de atendimento tradicionais é muito dispendioso dispor-se de um departamento com pessoas capazes de atenderem a um número significativo de línguas. De modo semelhante, o desenvolvimento de um serviço automático multilíngua acrescenta um esforço de desenvolvimento quase proporcional ao número de línguas disponíveis.

Em qualquer destes serviços multilíngua, tradicional ou automático, é necessário ainda atribuir cada utilizador ao atendedor apropriado. Este primeiro passo pode, em geral, ser dado pelo próprio utilizador sem grandes inconvenientes, em particular se existir um número relativamente pequeno de línguas de atendimento disponíveis. Por exemplo, em alguns serviços telefónicos nos E.U.A. é comum dispor-se de um número de telefone diferente para cada língua: espanhol, inglês e eventualmente mais uma ou duas línguas europeias.

Para o desenvolvimento do reconhecimento automático multilíngua, é necessário um número considerável de dados de cada língua e a obtenção dos respectivos modelos. Dispondo-se destes dados, fica facilitado o desenvolvimento de um mecanismo de identificação da língua. Contudo, mesmo nos serviços tradicionais, para os quais não existem ainda reconhecedores automáticos de fala, tem sido referida, por vários autores, a necessidade de um sistema automático de identificação de língua (Muthusamy et al., 1994a; Zissman, 1995). Nomeadamente, nos serviços de chamadas de emergência, têm sido verificados casos em que o operador não consegue entender a língua de um utilizador em estado de choque ou sob grandes tensões nervosas. A companhia AT&T introduziu recentemente o serviço "Language Line" para negócios, público em geral e para os departamentos de polícia que suportam as emergências 911[1]. Este serviço utiliza tradutores humanos trei-

[1] O 911 é o número de telefone de emergência nos E.U.A.

nados, permitindo atender 140 línguas diferentes. Contudo, uma grande responsabilidade recai sobre o operador humano que tem de proceder ao encaminhamento da chamada para o tradutor adequado. Uma chamada para a "Language Line" feita por alguém que fale exclusivamente tamil, resulta num atraso de três minutos antes que a língua seja identificada e que fique disponível na linha um tradutor de tamil. O atraso é devido ao operador que tenta sem sucesso três tradutores de línguas do Sudoeste Asiático e toca registos pré-gravados de outras línguas na tentativa de descobrir o tradutor adequado. O atraso ainda pode ser maior se o utilizador tentar ser cooperante e conseguir pronunciar o nome "tamil" em inglês em vez de o fazer na própria língua tamil (Muthusamy et al., 1994a). Espera-se que, num futuro próximo, um sistema automático possa vir a substituir este operador, reduzindo o tempo de identificação em uma ou duas ordens de grandeza (Zissman, 1995).

O problema da identificação da língua tem sido recentemente estudado por vários grupos de investigação no âmbito das telecomunicações (House e Neuburg, 1977; Hazen e Zue, 1993; Zissman, 1993; Berkling et al., 1994; Muthusamy et al., 1994a; Zissman, 1995; Caseiro, 1998). As características da fala que permitem obter a identificação da língua com algum sucesso incidem sobre o posicionamento e a concentração de fonemas e, obviamente, nas características lexicais e sintácticas (Arslan, 1996; Arslan e Hansen, 1996).

No capítulo 3 foram referidos alguns dos aspectos que influenciam a produção da fala estrangeira. As fronteiras de decisão no problema dos sotaques são difusas comparativamente com a identificação da língua. Embora possam existir línguas que se poderiam considerar intermédias, tais como os *pidgin*, estas serão pelo menos enumeráveis. O sotaque, por sua vez, pode ser afectado em graus diversos por várias línguas em simultâneo e por muitos outros factores decorrentes do processo de aprendizagem. Desta perspectiva decorre que o problema da identificação do sotaque seja considerado mais difícil do que o da identificação da língua (Arslan, 1996; Arslan e Hansen, 1996; Teixeira et al., 1996).

6.2 Identificação de características não linguísticas no sinal de fala

Ao fazer uma descrição integrada dos métodos preconizados para a identificação automática da língua, Zissman distingue os métodos baseados em HMMs, que modelam as características sequenciais da produção da fala, dos restantes métodos, que considera executarem essencialmente uma classificação estática, de acordo com determinados

parâmetros de segmentos de fala (Zissman, 1993). O primeiro sistema de identificação da língua baseado em HMMs (House e Neuburg, 1977) utilizava observações discretas em modelos ergódicos (subsecção 5.4). As sequências de símbolos de entrada eram resultado de transcrições fonéticas de textos publicados. Zissman apresenta um método semelhante, adaptado para o sinal de fala. Utiliza duas séries temporais de características do sinal e o conceito de ligação das respectivas funções de densidade de probabilidade das observações (gaussianas) através dos estados dos modelos de Markov (secção 2.4.4). Para cada língua, treinam-se um ou mais HMMs que são utilizados na descodificação do sinal de fala a ser classificado. No final do processo de descodificação, determina-se qual o modelo que produz o valor mais elevado de verosimilhança.

Os modelos HMM ergódicos podem igualmente ser utilizados no desenvolvimento de procedimentos que visam a identificação automática de outras características não linguísticas da fala, tais como o sexo ou a própria identidade do orador. Deste modo constituiu-se uma formulação mais geral para este tipo de problemas que tem como vantagem principal a facilidade de desenvolvimento e correspondente integração com os reconhecedores actuais baseados em modelos de fones (Lamel e Gauvain, 1993):

1. Treino de um modelo de Markov não observável λ_i para cada classe i da característica não linguística a ser identificada. Esse modelo é, nesta formulação, um descodificador fonético convencional (subsecção 5.4).

2. A identificação de um sinal de entrada x é feita pelo cálculo da verosimilhança $f(x/\lambda_i)$ para cada modelo λ_i. A classe escolhida será a associada ao modelo que tenha obtido o valor de verosimilhança mais elevado.

Esta formulação pode ser generalizada a outros tipos de unidades subpalavra diferentes do fone, embora raramente tenham sido utilizadas. As unidades que integram contexto, tais como os trifones, são necessariamente muito numerosas comprometendo a aplicação prática do método. A escolha de outro tipo de unidades sem carácter linguístico, não permite geralmente dispor de igual quantidade de informação a priori o que, de acordo com Lamel e Gauvain, impede a obtenção de modelos mais discriminativos.

Referem-se algumas vantagens desta formulação:

- possibilita efectuar a identificação das referidas características não linguísticas independentemente do texto da mensagem oral;

- garante maior precisão do que os métodos baseados em estatísticas de longa duração, tais como o espectro de longa duração, dicionários de codificação vectorial e

mapas acústicos probabilísticos (Tseng et al., 1992);

- facilita a integração de restrições fonotácticas, o que é particularmente útil na identificação da língua uma vez que estas restrições são em geral bem conhecidas e substancialmente diferentes.

Com base na última vantagem, admite-se a possibilidade de se obterem outras alternativas ao descodificador fonético que, ao integrarem mais informação sobre o problema que se pretende resolver, permitam obter melhores desempenhos. Uma destas alternativas consiste em impor restrições na rede ergódica de acordo com as transcrições fonéticas correspondentes a um determinado vocabulário (Teixeira et al., 1996). Assim, cada rede ergódica é substituída pelo conjunto de sequências de modelos de fones em paralelo ou seja, por um reconhecedor de palavras isoladas com um determinado vocabulário. Deste modo, obtém-se simultaneamente um identificador de determinada característica não linguística e um reconhecedor automático de fala, ambos adaptados ao contexto de determinada aplicação.

Um descodificador fonético pode, por sua vez, fornecer dados que permitam o reconhecimento da fala. Para tal, bastaria associar um segundo sistema a jusante do primeiro, capaz de interpretar as sequências de fones e recorrendo, nomeadamente, a um léxico de pronúncia. O desenvolvimento de um sistema deste tipo revela-se, contudo, complexo, uma vez que as sequências obtidas na saída de um descodificador fonético são muito irregulares apresentando inúmeras inserções e supressões. A utilização de *N-gramas* pode filtrar algumas destas irregularidades (secção 2.6). Adicionalmente, se for possível determinar *N-gramas* de fones específicas de cada classe a identificar, consegue-se integrar mais informação útil em termos discriminativos. Para *N-gramas* com N suficientemente elevado obtêm-se transcrições que podem ser múltiplas para a mesma palavra:

- Na secção 6.5 são apresentadas experiências realizadas com reconhecedores que utilizam um léxico de pronúncia com uma única transcrição fonémica por palavra. Estes sistemas têm a capacidade de, simultaneamente com o reconhecimento, realizarem a identificação do sotaque do orador.

- Na secção 6.6 descrevem-se algumas experiências de reconhecimento, também com identificação do sotaque, nas quais se utiliza um modelo que descreve a ocorrência de diversas transcrições: o modelo de transcrição apresentado no capítulo 5.

Outra alternativa, óbvia no contexto do reconhecimento de fala, consiste na substituição das transcrições com modelos de fones por modelos de palavras. Em termos de

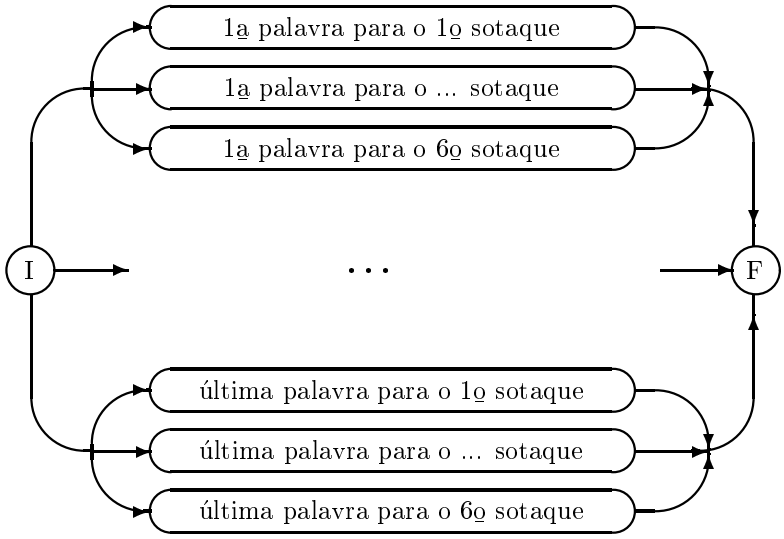


Figura 6.1: Representação esquemática de um reconhecedor de palavras isoladas com modelos múltiplos.

reconhecimento aplica-se o que foi referido na subsecção 2.5.1, a propósito da utilização de modelos de palavra e subpalavra. Em relação à identificação da característica não linguística, verificam-se alguns factos equivalentes, tal como a dependência do vocabulário e antevêem-se outros semelhantes, tal como a melhoria do desempenho devida à incorporação de informação referente à coarticulação dos fones. Nas secções 6.3 e 6.4 apresentam-se algumas experiências de reconhecimento com modelos de palavra, nas quais é possível identificar, simultaneamente e com algum sucesso, o sexo e o sotaque do orador. Na figura 4.1a representava-se esquematicamente o reconhecedor de palavras isoladas tradicional, utilizado nessas experiências para um único sotaque. Na figura 6.1 apresenta-se uma topologia que permite o uso vários modelos para cada palavra do vocabulário. Se cada um destes modelos for treinado com sinais de fala apresentando sotaques diferentes, é possível, deste modo, obter-se um mecanismo de identificação do sotaque.

6.3 Identificação do sexo versus sotaque

É geralmente reconhecido (Gupta e Mermelstein, 1982; Teixeira e Trancoso, 1993; Hansen e Arslan, 1995; Arslan, 1996; Arslan e Hansen, 1996) que o sexo é uma característica do orador mais importante do que o respectivo sotaque no reconhecimento de fala independente do orador. A importância desta característica do orador na produção do sinal de fala tem levado a generalidade dos investigadores a utilizar modelos independentes para cada sexo por forma a obter os melhores resultados de reconhecimento (Vergin

et al., 1996). Este procedimento enquadra-se na formulação geral apresentada na secção 6.2 e foi também adoptado no capítulo 5.

A determinação do sexo do orador baseada numa única locução e repetida para outras locuções do mesmo orador não é aconselhável em sistemas de reconhecimento com tarefas complexas que impliquem muita interacção com cada utilizador. Deste modo, a duplicação do número de modelos implica simultaneamente a duplicação do tempo de processamento. Nestes casos, pode ser vantajoso caracterizar o orador uma única vez, utilizando-se a partir daí apenas o conjunto de modelos referente ao sexo determinado (Lamel e Gauvain, 1993). O mesmo raciocínio pode ser aplicado ao problema do sotaque. Obtém-se desta forma um sistema hierarquizado em três níveis de decisão: o sexo, o sotaque, a palavra (ou a frase). A decisão do sexo e do sotaque deverá ocorrer numa fase inicial, de forma semelhante a um processo de adaptação ao orador, que identifica o reconhecedor mais adequado. Para além de reduzir os custos de computação, este tipo de estratégia permite explorar de forma independente eventuais melhorias de cada um dos níveis de decisão.

6.3.1 Identificação do sexo no sinal de fala

O problema da identificação do sexo do orador tem sido investigado utilizando classificadores relativamente simples. Childers et al. utilizaram diversos esquemas de identificação baseados na comparação de padrões de referência e de teste (Childers et al., 1988). Realizaram experiências com diversos tipos de métricas e de parâmetros do sinal determinados em tramas de 25,6 milissegundos. Utilizaram parâmetros de LPC, de autocorrelação, de reflexão, do cepstrum e mel-cepstrum. Em alternativa, utilizaram ainda um vector com a frequência fundamental mais as frequências, larguras de banda e amplitudes das primeiras quatro formantes. As métricas utilizadas foram: a euclidiana, a de Mahalanobis, a função densidade de probabilidade, a distância de Itakura (só para os LPC) e uma medida para os parâmetros do cepstrum equivalente a d_c (equação 2.9). Analisaram em separado os resultados obtidos com vogais e fricativas vozeadas e não vozeadas, sustentadas durante aproximadamente 2 segundos. Os resultados obtidos com as fricativas não vozeados foram, obviamente, os mais baixos, ainda assim com taxas de identificação acima dos 80%. Para as fricativas vozeadas obteve-se 98% e para as vogais 100%. Posteriormente utilizaram um outro método baseado em diversas decisões com limiares. Estas decisões são determinadas com base no cálculo de LPC em cada período da fundamental das vogais anteriormente analisadas. Nestas condições, obtiveram-se taxas de identificação superiores a 98%. Eliminando 2 dos 52 oradores utilizados (27 de cada sexo) conseguiu-se obter 100% nas mesmas circunstâncias.

Fussel (Fussel, 1991) realizou um estudo com sinais de fala lida do corpus DARPA TIMIT (Keating et al., 1994). Utilizou parâmetros do cepstrum e do delta-cepstrum num classificador gaussiano (Duda e Hart, 1973) calculados em tramas de 16 milissegundos. Os resultados obtidos confirmaram a superioridade das vogais para o desempenho da identificação do sexo, quando comparadas com outros fonemas. São contudo em geral inferiores aos obtidos por Childers et al., o que pode ser justificado pela duração substancialmente inferior das vogais pronunciadas de forma natural em fala lida. Além disso, acresce o facto da existência de coarticulação com os fonemas circundantes.

Um outro grupo de investigadores desenvolveu um método de identificação para determinar modelos de reconhecimento específicos de cada sexo (Vergin et al., 1996). Este método baseia-se nas diferenças da posição da primeira e da segunda formante entre homens e mulheres. Os testes foram efectuados com 10 oradores que proferiram o total de 201 frases, seleccionados do corpus ATIS ("Air Travel Information System"). Embora a duração do segmento de fala utilizado na identificação fosse muito superior ao dos trabalhos anteriormente referidos, a taxa de identificação obtida foi de apenas 90%. No reconhecimento de palavras obteve-se uma redução do erro de 14% comparativamente aos resultados obtidos com modelos independentes do sexo.

6.3.2 Sexo versus sotaque no reconhecimento de fala

Seguidamente apresentam-se algumas experiências preliminares que procuraram determinar qual dos factores referentes ao orador, sexo ou sotaque, tem maior influência no reconhecimento automático de fala. Nestas experiências utilizaram-se exclusivamente modelos de palavra, para um vocabulário de quarenta palavras. Para o treino e teste destes modelos utilizou-se todo o material de fala recolhido de oradores britânicos e dinamarqueses do corpus SUNSTAR multissotaque (capítulo 3). Utilizaram-se 70% destes oradores para o treino dos modelos e os restantes 30% para os respectivos testes. O número de oradores utilizados para treino e para teste continham igual número de oradores nativos e não nativos, e igual número de oradores de cada sexo.

oradores	treino	teste
femininos	35%	15%
masculinos	35%	15%

Tabela 6.1: Distribuição do material de fala de acordo com o sexo dos oradores utilizados

Testes com um modelo por cada palavra do vocabulário

Os oradores de cada sexo foram utilizados em conjuntos separados para efeito do treino de modelos e na determinação dos resultados dos testes (tabela 6.1). Os resultados de reconhecimento apresentados na tabela 6.2 foram obtidos com três reconhecedores diferentes (um por cada linha). O primeiro utilizou no treino os oradores femininos disponíveis para esse fim (nativos e não nativos). O segundo utilizou exclusivamente os oradores masculinos e o terceiro, todos os oradores disponíveis para o treino.

treino/teste	feminino	masculino	total
feminino	92,5	56,5	74,5
masculino	87,5	92,3	89,9
masc.+fem.	94,4	90,8	92,6

Tabela 6.2: Taxas de reconhecimento (%) obtidas com três reconhecedores treinados com diferentes combinações de oradores, agrupados de acordo com o respectivo sexo (Teixeira e Trancoso, 1993).

Os modelos treinados com oradores de ambos os sexos permitiram obter resultados globalmente superiores do que os obtidos com os modelos específicos de cada sexo. Os oradores femininos registaram taxas de reconhecimento superiores aos obtidos com os oradores masculinos.

Posteriormente, procedeu-se a uma divisão do material de fala anteriormente utilizado de acordo com o carácter nativo ou não nativo do sotaque de cada um dos oradores. Manteve-se uma proporção idêntica à da tabela 6.1 (substituindo a distinção feminino/masculino por nativo/não nativo).

teste	nativo		não nativo		total	
treino	CHMM	SCHMM	CHMM	SCHMM	CHMM	SCHMM
nativo	94,4	94,8	67,7	72,7	81,1	83,8
não nativo	82,9	82,1	93,7	91,2	88,3	86,7
nat. + não nat.	94,8	94,2	90,4	89,4	92,6	91,8

Tabela 6.3: Taxas de reconhecimento (%) obtidas com três reconhecedores treinados com diferentes combinações de oradores, agrupados de acordo com o respectivo sotaque (Teixeira e Trancoso, 1993).

A tabela 6.3 refere-se a experiências equivalentes às descritas nas tabela 6.2 quando se substitui a divisão dos oradores em relação ao sexo pela divisão relativa ao uso da primeira ou da segunda língua. Alternativamente ao reconhecedor de observações contínuas utilizou-se um reconhecedor de observações semicontínuas (SCHMM, secção 2.4). Assim, na tabela 6.3 apresentam-se os resultados obtidos com seis reconhecedores diferentes, existindo em cada linha da tabela resultados referentes ao reconhecedor de observações contínuas e ao seu equivalente de observações semicontínuas. A primeira linha de resultados refere-se a modelos exclusivamente treinados com fala de oradores nativos, a segunda, com fala de oradores não nativos e na terceira os modelos de palavra foram treinados com todos os oradores disponíveis para o treino. No caso dos modelos contínuos, este reconhecedor é o mesmo que o utilizado nos testes referentes à terceira linha de resultados na tabela 6.2.

treino / teste	nativo	não nativo	total
nativo	96,9	79,7	88,3
não nativo	86,9	96,9	91,9
nat.+não nat.	96,9	95,9	96,4

Tabela 6.4: Taxas de reconhecimento (%) obtidas com CHMMs treinados com 80% do material de fala disponível (Teixeira e Trancoso, 1992).

A análise da tabela 6.3 confirma alguns dos resultados dos capítulos 4 e 5. Por forma a facilitar a comparação de resultados com as experiências do capítulo 4 apresentam-se na tabela 6.4 os resultados relevantes no presente contexto[2]. Verifica-se, nomeadamente, que as quebras de desempenho obtidas são maiores quando o grupo de treino é nativo e o de teste é não nativo, do que na situação inversa. Verifica-se também o aumento do desempenho global do reconhecedor com ambos os grupos de oradores de teste quando estes se encontram simultaneamente representados no grupo de treino dos modelos. A única diferença entre os reconhecedores referentes à tabela 6.4 e os reconhecedores baseados em observações contínuas referentes à tabela 6.3, reside na repartição do corpus de fala para o treino e o teste dos modelos. De facto, nas primeiras experiências utilizaram-se 80% dos oradores para o treino dos modelos em vez dos 70% agora utilizados (20% para o teste em vez dos 30% agora utilizados de acordo com a tabela 6.1). Esta diferença é tanto mais significativa, se se considerar que o número total de oradores é relativamente baixo (40 oradores). Tal facto justifica o decréscimo de desempenho verificado em relação aos resultados da tabela 6.4.

[2]Resultados seleccionados das tabelas 4.4 e 4.5.

Comparando agora os resultados da tabela 6.3 com os obtidos na tabela 6.2, verificam-se relações semelhantes entre cada um dos grupos de três reconhecedores utilizados. Sobressai apenas o mau resultado obtidos pelos oradores masculinos com os modelos treinados com oradores femininos. Apenas esta diferença indicia a preponderância do factor sexo sobre o factor sotaque, a qual é verificada de forma mais clara nas experiências seguintes.

Após a realização do reconhecedor de observações semicontínuas, pretendeu-se verificar o seu desempenho em tarefas com palavras isoladas. Nas experiências descritas na secção 4.7 obtiveram-se desempenhos ligeiramente inferiores aos conseguidos com o reconhecedor de observações contínuas. A justificação aí avançada permanece válida no presente contexto, uma vez que esta tendência pode igualmente ser comprovada nas experiências referentes às tabelas 6.3 e 6.6.

Testes com dois modelos por cada palavra do vocabulário

modelo/teste	feminino	masculino
feminino	86,2	6,7
masculino	4,6	87,9
total	90,8	94,6

Tabela 6.5: Taxas de reconhecimento (%) obtidas com um modelo para cada sexo e por cada palavra do vocabulário (Teixeira e Trancoso, 1993).

Na tabela 6.5 apresentam-se os resultados obtidos com um único reconhecedor construído com todos os modelos específicos utilizados nos reconhecedores referentes às primeiras duas linhas de resultados da tabela 6.2. Utiliza-se assim um reconhecedor com uma topologia do mesmo tipo da que se encontra representada na figura 6.1. Nesta nova tabela as duas primeiras linhas de resultados têm um significado diferente e referem-se à percentagem de palavras correctamente identificadas com cada um dos grupos de modelos referentes a cada sexo. Deste modo a taxa de reconhecimento para cada sexo, é obtida na terceira linha e corresponde à soma das linhas anteriores. A média dos valores na terceira linha corresponde à taxa de reconhecimento global, o qual é de 92,7%. Este valor é praticamente igual ao obtido com o reconhecedor referente à terceira linha de resultados da tabela 6.2 e no qual se utilizou igualmente todo o material de treino (embora com um único modelo por palavra do vocabulário). Curiosamente os oradores masculinos obtêm agora um desempenho superior ao dos oradores femininos, ao contrário

do que acontecia com o reconhecedor anterior. Assim, os oradores femininos do grupo de teste conseguem melhores resultados com um único modelo por palavra, enquanto que algumas locuções de oradores masculinos obtiveram valores de verosimilhança superiores com modelos treinados com oradores femininos.

Realça-se o facto de este reconhecedor permitir desenvolver a identificação do sexo do orador, bastando para tal determinar a que sexo se refere o modelo que obtém maior verosimilhança dada uma determinada locução. Com os valores obtidos na tabela 6.5 é possível calcular a taxa de identificação do sexo sobre as palavras correctamente reconhecidas obtendo-se o valor de 87,1%.

teste	nativo		não nativo	
treino	CHMM	SCHMM	CHMM	SCHMM
nativo	83,1	81,9	13,5	18,3
não nativo	11,0	11,5	80,0	70,2
total	94,1	93,4	93,5	88,5

Tabela 6.6: Taxas de reconhecimento (%) obtidas com um modelo para cada sotaque e por cada palavra do vocabulário (Teixeira e Trancoso, 1993).

A tabela 6.6 relaciona-se com a tabela 6.5 de forma semelhante à referida relação entre a tabela 6.3 e a tabela 6.2. Ou seja, refere-se a experiências equivalentes às descritas na tabela 6.5 quando se substitui a divisão dos oradores em relação ao sexo, pela divisão relativa ao uso de primeira ou segunda língua. Assim, utilizam-se em simultâneo os modelos treinados com oradores nativos e os modelos treinados com oradores não nativos, num único reconhecedor. De forma semelhante ao efectuado com a tabela 6.5, é possível determinar-se uma taxa de reconhecimento global de 93,8% para o reconhecedor com modelos contínuos e de 91,0% para o reconhecedor com modelos semicontínuos. Estes valores são muito próximos dos obtidos com os reconhecedores referentes à terceira linha de resultados da tabela 6.3.

À semelhança do mecanismo de identificação do sexo utilizado na experiência anterior, (tabela 6.5) é agora possível obter-se uma forma de detecção do sotaque. Para tal, determina-se a que grupo de oradores (nativo ou não nativo) se refere o modelo que apresenta maior verosimilhança para uma dada locução. Assim, determinam-se, a partir da tabela 6.6, as taxas de detecção do sotaque sobre as palavras correctamente reconhecidas, obtendo-se os valores de 80,6% e de 76,1%, respectivamente para os reconhecedores com modelos contínuos e semicontínuos. Por fim, comparam-se estes resultados com o valor

obtido na identificação do sexo dos oradores (87,1%). A diferença é justificada com base no modo como as características do sinal da fala mais especificamente dependentes do sexo, ou do sotaque, influenciam o funcionamento dos reconhecedores utilizados. Assim, as características dependentes do sexo, parecem ser as mais influentes neste funcionamento.

6.3.3 Conclusões

Alguns dos métodos desenvolvidos no âmbito desta tese permitem caracterizar um dado orador em termos de sexo. Os resultados obtidos são ligeiramente inferiores aos obtidos por Vergin et al. (90%) em termos de taxas de identificação (Vergin et al., 1996). De facto, seria de esperar uma diferença ainda maior. Nesse estudo os modelos foram treinados com uma quantidade muito superior de sinais de fala (285 oradores proferiram um total de 9269 frases) e cada decisão foi baseada em duas frases completas. Estes resultados parecem indicar que o uso exclusivo da informação relativa às duas primeiras formantes, adoptado por Vergin et al., não conduz aos melhores resultados de identificação do sexo. Contudo, em ambos os casos, estes resultados deveriam ser superiores aos obtidos por Fussel, se se considerar que este utilizou segmentos com apenas 16 milissegundos de duração (Fussel, 1991). Este contra-senso aparente pode ser justificado por dois tipos de argumentos:

- Todos os sinais utilizados por Fussel são segmentos de sequências acústico-fonéticas etiquetadas por peritos do MIT. Os resultados foram obtidos para cada fonema ou classe de fonemas, utilizando os respectivos segmentos no treino e no teste. Em contrapartida, os sinais utilizados neste trabalho foram exclusivamente segmentados com métodos automáticos, sendo o teste restringido a um vocabulário que, não sendo muito grande, apresenta, contudo, uma perplexidade superior à tarefa de Fussel. Em relação aos resultados obtidos com modelos subpalavra, acresce ainda o facto de se referirem exclusivamente a experiências independentes do vocabulário, o que contribui para a obtenção de mais erros de reconhecimento e de identificação.

- O objectivo final das experiências efectuadas no decorrer do presente trabalho, é o de obter melhores resultados de reconhecimento de fala e não o de distinguir objectivamente o sexo. Esta distinção é estudada apenas no sentido de se poder compreender os processos envolvidos no reconhecimento e de forma a cumprir o referido objectivo. A maioria dos métodos estudados correspondem a reconhecedores de fala que adicionalmente permitem a identificação do sexo do orador. Não foram alterados para se obterem melhores resultados em relação a este último aspecto.

Com os resultados obtidos nas experiências descritas nesta secção verificou-se o papel preponderante do sexo relativamente ao sotaque, em termos das características do orador. Tal facto reflecte-se numa maior facilidade de distinguir os oradores em termos de sexo e na conveniência de se utilizarem modelos de reconhecimento específicos, para cada um dos respectivos grupos de oradores.

6.4 Identificação do sotaque com modelos de palavra

As experiências de reconhecimento que são descritas seguidamente são a extensão das experiências apresentadas na subsecção 6.3.2 que permitiam a detecção do sotaque com modelos de palavra. Pretende-se com estas experiências identificar os seis sotaques existentes no corpus de fala multissotaque SUNSTAR.

Os resultados de reconhecimento e a taxa de identificação do sotaque referentes a estas experiências encontram-se representados na primeira linha de resultados da tabela 6.7. A repartição de vocabulário e de oradores utilizada para a definição do conjunto de treino e de teste é idêntica à utilizada no capítulo 5. A topologia global adoptada para o único reconhecedor utilizado na presente secção é a representada na figura 6.1. O mesmo acontece com os restantes reconhecedores descritos no presente capítulo e cujos resultados também se encontram representados na tabela 6.7. Por isso, é possível comparar os resultados de reconhecimento de palavras apresentados nesta tabela com os da tabela 5.1, uma vez que a tarefa de reconhecimento é a mesma. O significado das colunas da tabela 6.7 é o mesmo do que foi descrito na secção 5.5.6 com excepção da última coluna (id.) agora adicionada e que se refere à taxa de identificação do sotaque. Esta taxa é também apresentada em termos percentuais e é calculada através da média ponderada dos elementos da diagonal da matriz de confusão. Ao longo do restante capítulo apresentam-se igualmente as matrizes de confusão referentes às experiências que permitiram obter os valores mais elevados da taxa de identificação do sotaque.

Comparando os resultados apresentados na primeira e na segunda linha respectivamente das tabelas 5.1 e 6.7, verificam-se valores de reconhecimento muito semelhantes. De acordo com as topologias descritas na figura 6.1, o presente reconhecedor resulta da associação em paralelo dos reconhecedores utilizados nas experiências da tabela 5.1. A diferença mais relevante consiste no facto de, enquanto que nas primeiras experiências se seleccionava a priori o grupo de oradores a ser testado com cada reconhecedor, o mesmo é agora feito de forma automática e revelando-se igualmente eficiente. Os erros de identificação (14,7%) poderão ser justificados pela existência de alguns oradores, ou de algumas

modelos	transcr.	observ.	M	da	de	en	es	it	pt	tot.	id.
palavra	não usa	não nat.	3	88,0	74,3	98,9	88,4	82,8	92,4	86,8	85,3
fones	léxico	não nativos	3	85,3	66,7	95,0	82,5	63,3	89,2	79,0	82,0
			6	84,1	72,8	93,9	81,8	68,6	90,2	80,8	83,2
rede de transcr.	não nativos	nativos	3	76,5	55,0	90,9	80,0	63,2	77,1	73,0	59,1
			6	71,4	49,3	89,8	75,6	63,3	78,3	70,4	57,0
		não nativos	3	82,2	61,6	95,3	81,7	70,5	85,7	78,6	84,5
			6	86,3	72,8	91,7	84,7	78,6	91,1	83,6	85,8

Tabela 6.7: Taxas (%) de reconhecimento (da 5ª à penúltima coluna) e de identificação do sotaque (última coluna) obtidas com diferentes modelos: tipo de modelo (1ª coluna); treino das probabilidades de transição interfones (2ª coluna); treino das probabilidades de transição intrafones e das funções densidade de probabilidade de observação (3ª coluna); número (M) de componentes gaussianas utilizadas no modelamento destas funções (4ª coluna).

locuções destes oradores, que são pronunciadas com um sotaque mais próximo de outro grupo de oradores.

A matriz de confusão referente à identificação de sotaque está representada na tabela 6.8. A respectiva leitura revela que os oradores dinamarqueses obtiveram a menor percentagem de locuções identificadas nos respectivos modelos específicos. As restantes locuções foram identificadas principalmente em dois grupos de modelos: os dos oradores nativos, o que pode ser justificado pelo facto do dinamarquês médio ter uma pronúncia razoável do inglês; os dos oradores espanhóis, o que pode ser justificado pelo facto da maior parte dos oradores dinamarqueses utilizados (da Jutlândia) apresentarem uma forte tendência para transformarem as fricativas alveolares (/s/) em palatais (/sh/) (capítulo 3). Embora de forma menos significativa, os modelos treinados com oradores portugueses tiveram um efeito semelhante. Os oradores alemães obtiveram a mais alta taxa de identificação com os seus próprios modelos. Este facto está provavelmente relacionado com a segmentação destas locuções que determinaram taxas de reconhecimento baixas, conforme foi descrito no capítulo 5. O mesmo se aplica, embora com menor incidência, em relação aos oradores italianos. Os erros de identificação dos oradores espanhóis são essencialmente devidos à existência dos modelos treinados com oradores portugueses. Estes últimos obtiveram a parcela mais significativa dos erros com os modelos nativos. De facto os oradores portugueses apresentaram uma pronúncia próxima da nativa para a maior parte das palavras.

sotaque	da	de	en	es	it	pt
da	63,6	1,0	7,6	19,3	2,0	6,6
de	0,0	99,6	0,0	0,0	0,0	0,4
en	2,2	0,6	88,1	5,0	0,3	3,9
es	3,7	1,7	4,9	83,7	0,0	6,0
it	0,0	3,7	1,2	0,5	91,9	2,6
pt	2,0	2,7	7,4	3,7	2,5	81,8

Tabela 6.8: Matriz de confusão (%) resultante do processo automático de identificação do sotaque. Utilizaram-se modelos de palavras treinados para cada sotaque ($M = 3$).

6.5 Identificação do sotaque com transcrição fixa

Repetiram-se experiências semelhantes às realizadas na secção anterior, mas substituindo na topologia da figura 6.1 os modelos de palavra por modelos subpalavra. Utilizou-se um vocabulário de treino diferente do de teste, de forma a obterem-se testes independentes do vocabulário, nas mesmas circunstâncias das utilizada na secção 5.3 (usando nomeadamente o mesmo léxico de pronúncia).

Os resultados de reconhecimento obtidos, expressos na segunda e terceira linhas da tabela 6.7, podem ser comparados com os das quinta e sexta linha da tabela 5.1, respectivamente. Estes resultados foram praticamente iguais aos obtidos com os reconhecedores específicos de cada sotaque. Tal facto reforça a ideia, verificada na secção anterior, de que a selecção de modelos com maior verosimilhança é efectuada de modo eficiente. Por outro lado, a taxa de identificação do sotaque obtida é agora ligeiramente inferior (2%) à da secção anterior. O facto destas experiências serem, ao contrário das anteriores, independentes do vocabulário, pode justificar este pequeno decréscimo de desempenho.

Da análise da matriz de confusão (tabela 6.9) tiram-se conclusões essencialmente semelhantes às obtidas com os modelos de palavra. Contudo, tornaram-se mais evidentes as relações entre as locuções dos oradores espanhóis e as dos portugueses, quer entre si, quer com a identificação com os modelos dos oradores dinamarqueses e ingleses, respectivamente.

Verifica-se a existência de valores mais elevados na zona superior à diagonal da matriz de confusão. Esta tendência já era possível de detectar na matriz de confusão da secção anterior mas de forma pouco acentuada. Esta distribuição de valores decorre directa-

sotaque	da	de	en	es	it	pt
da	63,3	0,2	3,4	23,5	1,7	7,8
de	0,0	97,6	0,0	0,0	0,4	2,0
en	4,1	3,0	76,5	6,6	1,7	8,0
es	5,5	2,7	3,4	81,3	0,5	6,5
it	0,2	2,1	0,2	0,7	92,8	4,0
pt	1,7	2,2	4,4	5,9	3,5	82,3

Tabela 6.9: Matriz de confusão (%) resultante do processo automático de identificação do sotaque. Utilizaram-se transcrições fonémicas nativas com modelos de subpalavra treinados para cada sotaque ($M = 6$).

mente do ordenamento escolhido para os grupos de oradores de treino e de teste e que corresponde à ordem alfabética. Uma justificação possível para esta tendência é o facto de as línguas germânicas e as línguas românicas terem ficado ordenadas consecutivamente. Tal facto poderia ser explorado no desenvolvimento da identificação automática da família linguística da língua materna do orador. Sendo as línguas germânicas referentes a países do Norte da Europa e as línguas românicas utilizadas nos países mais a Sul, verifica-se ainda uma distribuição mais rigorosa desta ordem de natureza aparentemente geográfica: a Dinamarca no extremo norte e Portugal no extremo sul do grupo de países seleccionados, ocupam os dois extremos do referido ordenamento.

6.6 Identificação do sotaque com modelos de transcrição

Os modelos de palavra inicialmente utilizados na topologia da figura 6.1 foram substituídos por modelos de transcrição (secção 5.5). Obtiveram-se dois tipos de modelos de transcrição distintos que permitem efectuar a identificação automática do sotaque: os que utilizaram os modelos subpalavra treinados com oradores nativos e aqueles que utilizaram os modelos subpalavra treinados com oradores de cada sotaque específico (nativos inclusive). Naturalmente que, com estes últimos, são esperados melhores desempenhos na identificação do sotaque.

6.6.1 Uso do conjunto de modelos subpalavra nativo

Começa-se por analisar os resultados obtidos com os modelos de transcrição que usam os modelos subpalavra treinados com os oradores nativos. Os resultados obtidos, expressos no penúltimo par de linhas da tabela 6.7, podem ser comparados com os apresentados no penúltimo par de linhas da tabela 5.1. Os resultados de reconhecimento apresentam um pequeno decréscimo de desempenho, embora mais acentuado no caso em que se utilizaram mais componentes gaussianas no modelamento das funções densidade de probabilidade de observação ($M = 6$).

Utilizando também os resultados apresentados na secções anteriores, é agora possível verificar que a taxa de identificação do sotaque apresenta uma relação directa com as taxas de reconhecimento de fala. Tal facto é justificado pela relação entre estas taxas e a qualidade dos modelos utilizados, em termos da representação dos sinais de fala com que foram treinados.

sotaque	da	de	en	es	it	pt
da	43,5	3,2	12,5	25,2	4,9	10,8
de	2,4	59,2	3,1	5,9	19,3	10,1
en	16,9	4,1	52,8	8,6	4,7	13,0
es	12,1	2,0	4,7	67,2	2,2	11,8
it	3,3	5,6	2,6	4,9	76,3	7,2
pt	7,4	8,6	7,9	16,5	15,0	44,6

Tabela 6.10: Matriz de confusão (%) resultante do processo automático de identificação do sotaque. Utilizaram-se modelos de transcrição. As probabilidades de transição interfones foram treinadas para cada sotaque. Os modelos subpalavra foram treinados com oradores nativos ($M = 3$).

Na tabela 6.10, apresenta-se a matriz de confusão referente à experiência que obteve melhor desempenho ($M = 3$). Os elementos fora da diagonal não revelam aspectos diferentes dos detectados nas experiências anteriores. Contudo, os valores mais elevados foram obtidos na detecção dos oradores dinamarqueses por modelos treinados com oradores espanhóis e nos oradores italianos detectados por modelos treinados com oradores alemães. Uma vez que os modelos subpalavra utilizados são os mesmos, estes resultados podem indicar a existência de semelhanças no modo como estes oradores transformam as transcrições nativas. Do mesmo modo se poderia então concluir que os valores mais elevados, obtidos nas anteriores matrizes de confusão e que não sobressaem na presente matriz,

podem estar associados a semelhanças acústico-fonéticas entre sotaques, modeladas nos respectivos modelos subpalavra.

Os dois valores máximos seguintes surgem, pela primeira vez, no triângulo inferior à diagonal. Também não revelam, contudo, aspectos realmente diferentes: os oradores ingleses são detectados com modelos treinados com dinamarqueses; os oradores portugueses são detectados com modelos treinados com oradores espanhóis. Em relação às secções 6.4 e 6.5, inverte-se aqui o papel das locuções dos oradores utilizados no treino com as dos oradores utilizados no teste.

Os valores da diagonal, ao contrário dos restantes, apresentam grandes variações em termos absolutos e relativos, quando comparados com os das experiências anteriores. Ao contrário da relação directa existente entre as taxas de reconhecimento e de identificação do sotaque, verifica-se que a relação entre as taxas de reconhecimento obtidas para cada grupo de oradores e os respectivos elementos da diagonal principal é praticamente inversa. Assim, os oradores dinamarqueses, ingleses e portugueses obtiveram os melhores resultados de reconhecimento e os três valores correspondentes, na diagonal principal da matriz de confusão, são os mais baixos entre os seis existentes. Este aspecto não era pelo menos tão evidente nas experiências anteriores. As locuções dos oradores nativos foram identificadas em menor número com os modelos de transcrição correspondentes. Este facto parece indicar algumas semelhanças, em termos acústico-fonéticos e das transcrições adoptadas, entre um número significativo de locuções dos três grupos de oradores anteriormente referidos.

6.6.2 Uso dos modelos subpalavra de cada sotaque

Por fim, substituíram-se os modelos de transcrição utilizados nas experiências da subsecção anterior por outros treinados a partir de modelos subpalavra específicos de cada sotaque, ou seja, todos aqueles que foram utilizados na secção 6.5.

Os resultados expressos nas últimas duas linhas da tabela 6.7, podem, mais uma vez, ser comparados com os apresentados em posição semelhante na tabela 5.1, uma vez que os modelos de transcrição utilizados são os mesmos, bem como os oradores de teste. Os resultados de reconhecimento são aproximadamente iguais mas, para $M = 6$, obteve-se um ligeiro acréscimo de desempenho. Embora estatisticamente irrelevante, este resultado foi o melhor alcançado sem a utilização de modelos de palavra. A taxa de identificação do sotaques obtida é, por sua vez, ligeiramente superior à taxa obtida com modelos de palavra. Este é um dos resultados mais significativos até agora obtido com

os modelos de transcrição. Estes resultados, conjuntamente com os obtidos na secção 5.5, parecem indicar que estes modelos são capazes de descrever os aspectos relacionados com o problema dos sotaques estrangeiros de forma mais adequada do que as transcrições fonémicas simples.

sotaque	da	de	en	es	it	pt
da	69,2	0,2	3,7	19,3	2,4	5,1
de	0,2	99,1	0,0	0,0	0,0	0,7
en	2,8	0,3	76,5	6,9	0,8	12,7
es	6,2	0,5	2,5	84,5	0,3	5,9
it	0,2	1,8	0,0	0,2	94,7	3,2
pt	1,5	1,0	4,4	4,9	3,0	85,2

Tabela 6.11: Matriz de confusão (%) resultante do processo automático de identificação do sotaque. Utilizaram-se modelos de transcrição treinados para cada sotaque ($M = 6$).

Na tabela 6.11, apresenta-se a matriz de confusão referente à experiência que obteve melhor desempenho ($M = 6$). Esta matriz apresenta algumas diferenças em relação às anteriormente obtidas. Apenas dois valores fora da diagonal são claramente superiores aos restantes, vindo contudo confirmar anteriores tendências: um desvio das locuções dos oradores ingleses para os modelos específicos do sotaque português e o desvio das locuções dos oradores dinamarqueses para os modelos do sotaque espanhol.

6.7 Testes perceptuais

É importante ter-se uma medida de comparação para os resultados obtidos com procedimentos de identificação automática do sotaque. Tal como acontece com outros procedimentos que procuram imitar tarefas habitualmente executadas directamente pelo homem, a comparação com o desempenho humano é inevitável. Estas comparações são complexas devido à grande diversidade de aspectos que podem afectar os executantes humanos, nomeadamente, se se tratam de tarefas que abranjam grande diversidade de aptidões, como aquelas que se relacionam com a fala. No caso da identificação do sotaque, o papel destes executantes é o de ouvintes. É assim conveniente controlar os ouvintes de forma semelhante ao que se fez com os oradores na criação do corpus de fala.

No caso do reconhecimento de fala, é habitualmente assumida a superioridade absoluta

das capacidades humanas. Em testes de inteligibilidade (Leeuwen et al., 1995), os ouvintes apresentaram uma taxa de erro de reconhecimento de palavras de 2,6% e detectou-se algumas dificuldades no reconhecimento de frases mais longas. O reconhecedor automático apresenta para a mesma taxa o valor de 12,6% e evidenciou problemas essencialmente com as frases com grande perplexidade. Um ouvinte comum dispõe de uma experiência intensiva com a sua língua materna, mesmo antes de ingressar no ensino escolar básico. Outras tarefas relacionadas, tais como a identificação do sexo ou do orador, contam com uma experiência semelhante, contudo, não sendo nestes casos tão clara a eficácia do ouvinte humano. O reconhecimento humano do orador é em geral praticado exclusivamente num grupo de oradores reduzido provenientes do ambiente afectivo ou profissional do ouvinte.

Em tarefas que envolvem línguas estrangeiras, tais como a identificação da língua e do sotaque, a experiência do ouvinte comum é, com raras excepções, pontual. Deste modo, existe grande expectativa em relação a procedimentos automáticos capazes de incorporar uma quantidade considerável de informação sobre essas línguas. Tal é tanto mais verdade quanto se verifica, na generalidade dos ouvintes humanos, o desconhecimento parcial ou total dessas línguas e que o contacto com os oradores nativos dessas línguas é diminuto ou inexistente (Muthusamy et al., 1994b).

Nos estudos destas tarefas designam-se por *testes perceptuais* os testes realizados com ouvintes humanos. O termo perceptual sublinha o envolvimento dos fenómenos da percepção humana e contrasta sobretudo com a designação de testes auditivos, destinados à avaliação das capacidades fisiológicas do ouvido.

Testes muito informais (Teixeira et al., 1996) foram realizados com um subconjunto reduzido do corpus SUNSTAR multissotaque. Seleccionaram-se palavras isoladas proferidas por seis oradores do sexo feminino, um de cada um dos países representados. Os ouvintes avaliados eram cinco portugueses e um francês, todos eles participantes frequentes de projectos e conferências internacionais e, como tal, habituados a ouvir o inglês falado por estrangeiros. Todos os ouvintes foram informados sobre as nacionalidades disponíveis no corpus de teste. A sessão de audição consistiu no encadeamento de sequências virtualmente infinitas de palavras, do mesmo orador, interrompidas assim que o ouvinte tomasse uma decisão. Duas conclusões principais foram determinadas com base neste pequeno teste informal:

(1) foi em geral muito difícil obter qualquer resposta dos ouvintes antes de estes ouvirem cerca de 6 a 12 palavras consecutivas do mesmo orador;

(2) o sotaque nativo, inglês britânico, foi sempre identificado correctamente.

Observaram-se igualmente outros factos que podem, contudo, estar dependentes dos oradores seleccionados não sendo, por isso, generalizáveis:

(3) os sotaques italianos e espanhóis eram, em geral, objecto de confusão entre si;

(4) os sotaques alemães e dinamarqueses eram igualmente confundidos entre si e por vezes com o inglês nativo, embora com menos frequência;

(5) o sotaque dos portugueses foi, em geral, difícil de caracterizar.

Considerando que a conclusão (2) se verificou em todos os testes e que os erros com inglês nativo mencionados em (4) são pouco frequentes, pode-se concluir que a taxa de detecção do sotaque é muito alta.

Arslan e Hansen realizaram testes perceptuais, com o corpus recolhido na Universidade de Duke, (secção 3.3) para quatro sotaques de inglês americano: neutro, (nativo) chinês, turco e alemão; num total de 48 oradores do sexo masculino (Arslan, 1996; Arslan e Hansen, 1996). Compararam os resultados com os obtidos por um sistema automático que podia efectuar a decisão, em relação a um dado orador, com base em uma ou mais locuções. Verificou-se que o desempenho do sistema de identificação automático aumentava com o número de palavras pronunciadas, tendo-se obtido 93% de identificações e 96% de detecções do sotaque correctas com o uso de 7–8 palavras, contra 68,7% e 83,3% respectivamente, quando a decisão se baseou em apenas uma única palavra. Nos testes perceptuais realizados permitia-se a repetição da mesma palavra até que o orador apresentasse uma decisão. Trata-se, portanto, de uma decisão com base numa única palavra, mas possibilitando a concentração do ouvinte nos detalhes que considerar relevantes para a tarefa. Utilizaram 12 ouvintes nativos e 9 não nativos do inglês. Os ouvintes apresentaram os resultados médios de 52,3% e 77,2%, respectivamente, para a identificação e para a detecção do sotaque, os quais são significativamente inferiores, em particular a taxa de identificação, aos resultados obtidos por via automática.

Este estudo incluiu ainda testes de reconhecimento de fala com os modelos específicos de cada sotaque. Tal como se obteve nos resultados apresentados no capítulo 5, quanto pior o desempenho de um sotaque com os modelos nativos, maior a redução da taxa de erros com a utilização dos modelos específicos correspondentes a esse sotaque. Também aqui, estas reduções não foram suficientes para alterar as posições relativas em termos de taxas de reconhecimento. Os oradores nativos de chinês apresentaram os piores resultados de reconhecimento, seguidos dos de turco, dos de alemão e finalmente, como seria de esperar, os dos nativos de inglês. Curiosamente, esta ordem poderia ser estabelecida

pelas distâncias entre as respectivas primeiras línguas em termos geográficos. Existem de facto relações profundas entre a geografia, a história e os elementos constitutivos e estruturais de cada língua (Saussure, 1916; Akmajian et al., 1990b).

Mais recentemente (Kumpf e King, 1997b) realizaram-se testes perceptuais com o corpus ANDSOL (secção 3.3). Utilizaram-se 22 oradores australianos, 26 oradores líbios e 24 do Vietname do Sul. Estes oradores são todos do sexo masculino, têm mais de 17 anos de idade, residiram pelo menos 4 anos na Austrália e pronunciaram um total de 6450 frases em inglês com uma duração média de 4,4 segundos. Utilizaram-se 20 ouvintes australianos. Realizou-se um estudo estatístico para avaliar a influência da:

- experiência do orador com o inglês;
- duração dos segmentos de fala;
- duração da experiência na percepção do ouvinte.

Foram igualmente realizadas entrevistas após os testes perceptuais, em que se procurou descobrir quais as marcas características do sotaque utilizadas pelos ouvintes. O sotaque nativo foi sistematicamente identificado correctamente, de forma confiante, por parte dos oradores, à semelhança do que foi verificado entre os sotaques europeus do inglês (Teixeira et al., 1996). Os erros mais significativos verificaram-se com as frases proferidas pelos oradores líbios que foram confundidos com oradores vietnamitas ou australianos consoante a predominância do seu próprio sotaque.

Sublinha-se aqui a inevitabilidade de surgirem desvios na identificação de um orador estrangeiro no sentido de o confundir com um orador nativo. A classificação de referência inicial é essencialmente estabelecida com base na nacionalidade do orador e não atende a capacidades ou experiências particularmente intensivas com a segunda língua. Mesmo quando o orador estrangeiro apresenta hábitos linguísticos bem demarcados da sua língua materna, tal não o impede, de acordo com as suas aptidões e experiências pessoais, de apresentar um sotaque mais próximo do nativo do que do sotaque típico dos seus conterrâneos.

As frases utilizadas por Kumpf e King permitiram a obtenção de uma taxa de identificação do sotaque de 87,4% em segmentos de fala de 6 segundos de duração. Este resultado foi superior ao obtido com o procedimento automático que desenvolveram: 84,7% para segmentos com 8,1 segundos de duração. A diferença foi justificada pelo facto de os ouvintes humanos combinarem o processamento das características de baixo nível, semelhantes às utilizadas pelo procedimento automático, com o conhecimento morfológico e sintáctico

da língua. Contudo, para segmentos com durações de 40 segundos, o mesmo procedimento permitiu-lhes obter uma taxa de 88,1%.

Por outro lado, estudos anteriores, realizados com a língua francesa, demonstraram que a detecção do sotaque não é influenciada pelo facto da fala resultar de uma leitura de palavras ou frases isoladas ou de um contexto de fala espontânea (Flege, 1984). O mesmo estudo revelou que determinados ouvintes eram capazes de detectar sotaques em segmentos com apenas 30 milissegundos de duração (aproximadamente o tempo de duração da relaxação da oclusiva referente ao fonema /t/).

6.8 Conclusões

No presente capítulo procurou-se abordar o problema do reconhecimento automático de fala de oradores não nativos, por forma a compreender as interacções existentes, neste contexto, entre os diversos grupos de oradores.

No capítulo 5 treinaram-se e testaram-se vários modelos para as palavra de determinado vocabulário. Alguns deles foram treinados com todos os oradores disponíveis independentemente do seu sotaque ou nacionalidade. Obtiveram-se, por outro lado, conjuntos de modelos treinados exclusivamente com oradores de cada nacionalidade, ou, em alternativa, com alguns parâmetros treinados com locuções de oradores nativos (secção 5.5.6). Ou seja, estes conjuntos de modelos deveriam de distinguir-se entre si através de algumas características específicas dos sotaques com que foram treinados. No presente capítulo associaram-se estes modelos em paralelo de acordo com uma estratégia geral, descrita na secção 6.2. A topologia obtida (figura 6.1) permite não só obter resultados satisfatórios de reconhecimento com oradores nativos e não nativos, como determinar com algum sucesso qual a língua materna de cada orador, com base numa única locução com uma só palavra.

Analisando as matrizes de confusão referentes ao processo de identificação do sotaque, notaram-se algumas relações entre os sotaques testados. Estes resultados poderão ter algum interesse prático, nomeadamente no dimensionamento de futuros corpus de fala para o treino de reconhecedores multissotaque. Outro aspecto importante do presente capítulo foi a confirmação da eficiência dos modelos de transcrição apresentados no capítulo 5, quer no reconhecimento de fala, quer na identificação do sotaque.

Os testes perceptuais realizados, dado o seu carácter preliminar e informal, não foram suficientemente concludentes. Ainda assim, as confusões detectadas entre sotaques não

foram, no geral, semelhantes às obtidas pela identificação automática. O aspecto aparentemente mais concordante refere-se à dificuldade em caracterizar o sotaque apresentado pelos oradores portugueses. Isto pode estar relacionado com os valores de verosimilhança elevados que apresentaram os modelos treinados com oradores portugueses na descodificação dos restantes sotaques. Contudo, a justificação mais provável para os resultados perceptuais obtidos, deverá estar associada ao facto da maioria dos ouvintes serem de nacionalidade portuguesa.

Capítulo 7

Conclusões e trabalho futuro

A presente dissertação resultou de um trabalho de procura de soluções mais robustas para o reconhecimento de fala a partir de um cenário de aplicações práticas que poderia ser descrito da seguinte forma:

- reconhecimento de um vocabulário de palavras isoladas de dimensão média;
- utilização de uma língua franca por parte de oradores nativos e não nativos pertencentes a um conjunto de nacionalidades conhecidas;
- aplicação de reconhecimento independente do orador acedida através da rede telefónica pública;
- inexistência de ruído no ambiente acústico do orador e no canal de transmissão;
- filtragem passa-banda típica de um canal de transmissão telefónico, não sendo considerada qualquer outra distorção devida ao microfone ou ao restante canal de transmissão;
- reconhecimento baseado numa única palavra proferida por cada orador, ou seja, não se considerou qualquer mecanismo de adaptação ao orador;
- utilização de reconhecedores baseados em modelos de Markov não observáveis.

A partir deste cenário e identificados os factores em condições de causarem maior degradação no desempenho deste sistema, estabeleceram-se duas prioridades:

1. rejeição de palavras estranhas;

2. robustez em relação à utilização por parte de oradores não nativos ou estrangeiros.

A estratégia empregue na rejeição de palavras estranhas, baseou-se no estudo da utilização de modelos múltiplos de escoamento (capítulo 4):

- usando diversos procedimentos para o seu treino com um corpus de sinais de fala de pequena dimensão (Teixeira e Lindberg, 1992);
- alterando a dimensão do vocabulário de palavras-chave. Na sequência deste estudo realizaram-se experiências de reconhecimento de fala ligada, nas quais se determinaram compromissos entre o uso de modelos linguísticos integrando modelos de escoamento e o número de frases incorrectas (Teixeira et al., 1992);
- comparando o uso de modelos HMM contínuos e semicontínuos. Contudo, não se verificou qualquer vantagem na utilização destes últimos (Teixeira et al., 1993a);
- utilizando-os com oradores que são não nativos. Para este caso ensaiaram-se duas estratégias (Teixeira e Trancoso, 1992):

  1. modelos de escoamento treinados em separado com oradores de cada nacionalidade — esta estratégia permite uma adaptação incremental para novos grupos de oradores de outras nacionalidades e o estabelecimento de um mecanismo de identificação automática do sotaque;
  2. modelos de escoamento treinados com oradores de todas as nacionalidades — apresenta como principal vantagem o facto de utilizar um número menor de modelos, com as consequentes reduções no armazenamento e processamento dos dados. Esta vantagem pode ser anulada pela adopção de modelos mais complexos, dada a disponibilidade de maior quantidade e variabilidade dos dados de treino para cada modelo. Uma vez que estes modelos podem ser considerados *independentes do sotaque*, poderão surgir outras vantagens nas situações em que existam oradores com nacionalidades diferentes das utilizadas no treino.

  A utilização de um modelo de escoamento específico para cada sotaque revelou taxas de rejeição ligeiramente superiores às obtidas com modelos de escoamento treinados simultaneamente com sotaques nativos e não nativos. Contudo, aumentando o número de modelos de escoamento, a capacidade de rejeição aumentou significativamente em ambos os casos, aproximando ainda mais as respectivas taxas de rejeição.

Em relação à segunda prioridade mencionada no início deste capítulo, procurou-se a obtenção de modelos de fala robustos em relação à utilização por parte de oradores não nativos. Consideraram-se inicialmente duas estratégias semelhantes às utilizadas para os modelos de escoamento:

- modelos de fala treinados em separado com oradores de cada nacionalidade. Tal como no caso dos modelos de escoamento, esta estratégia permite uma adaptação incremental para novos grupos de oradores de outras nacionalidades e o estabelecimento de um mecanismo de identificação automática do sotaque;

- modelos de fala treinados com oradores de todas as nacionalidades. Esta é a estratégia mais utilizada no quadro global em que se pode inserir este problema: o do reconhecimento independente do orador. Apresenta os mesmos compromissos referidos a propósito da estratégia adoptada para os modelos de escoamento.

Utilizando estas duas estratégias realizaram-se testes, os quais não revelaram diferenças significativas de desempenho (Teixeira e Trancoso, 1993). De qualquer modo, tal como se referiu em relação aos modelos de escoamento, é mais vantajoso o uso de modelos específicos, uma vez que pode ser considerado um novo sotaque, acrescentando apenas os modelos de fala correspondentes. Do mesmo modo, a vantagem mais significativa de uma estratégia do tipo independente do orador é a obtenção de uma robustez acrescida dos modelos, em relação a novos grupos de oradores de teste. No caso do problema dos sotaques, estes oradores deveriam ter uma nacionalidade diferente das utilizadas no treino dos modelos. A obtenção de um reconhecedor *independente do sotaque* poderá ser objecto de trabalho futuro.

Numa tentativa de compreender o problema dos sotaques de oradores não nativos ao nível das unidades subpalavra, realizaram-se diversas experiências independentes do vocabulário com reconhecedores baseados em modelos de fones. Em particular, procurou-se averiguar as possíveis variações na pronúncia dos oradores estrangeiros em relação à transcrição fonotípica de cada palavra. Para tal, concebeu-se uma estrutura, baseada em modelos HMM, capaz de integrar de forma probabilística transcrições múltiplas, a qual foi designada por modelo de transcrição (Teixeira et al., 1997). A utilização deste tipo de modelos no reconhecimento independente do orador permitiu obter taxas de reconhecimento significativamente superiores às conseguidas com as transcrições fonotípicas. Além disso, utilizando métodos relativamente simples, foi possível determinar, de forma aproximada, quais as variações na pronúncia mais prováveis para cada grupo de oradores (capítulo 5).

O problema da determinação automática da transcrição do sinal de fala em termos de unidades subpalavra é um assunto crucial para o reconhecimento de fala contínua em geral e não apenas no âmbito do reconhecimento de sotaques não nativos. Este problema tem sido recentemente objecto de diversos estudos.

Como subproduto da utilização simultânea dos reconhecedores específicos para cada sotaque, surge a identificação automática do sotaque. Contudo, pode ser interessante dissociar o processo de identificação do de reconhecimento, utilizando metodologias autónomas que podem ser melhoradas separadamente, tendo por objectivo final a obtenção de taxas de reconhecimento mais elevadas. As taxas de identificação do sotaque obtidas (Teixeira et al., 1996) são comparáveis ou mesmo superiores aos resultados publicados em estudos semelhantes (Arslan, 1996). Verificou-se que, no reconhecimento de fala, o sexo do orador é um factor mais determinante do que o sotaque. Como tal, é também mais fácil de identificar do que este último (Teixeira e Trancoso, 1993).

7.1 Desenvolvimentos futuros

A elaboração desta dissertação representou uma oportunidade do autor para organizar a informação disponível referente ao trabalho exposto. De igual modo, permitiu repensar alguns pormenores e estratégias de abordagem dos problemas inicialmente propostos, para os quais se exige uma confirmação experimental. Além disto, dada a intensa produção científica actualmente verificada nestes domínios, é cada vez mais necessário estar atento às propostas de outros autores e confrontar novos resultados experimentais. Assim, o trabalho aqui descrito não se encerra com a conclusão da presente dissertação. Seguidamente descrevem-se alguns dos assuntos a serem estudados futuramente.

Em testes de campo, as palavras estranhas proferidas por oradores estrangeiros poderão ser, em média, mais fáceis de rejeitar do que as palavras estranhas proferidas por oradores nativos. Não deverá ser difícil provar que estas palavras serão expressas de alguma forma ainda menos semelhante às palavras-chave, porventura com mais distorções devidas ao sotaque, podendo mesmo ser palavras pertencentes à língua materna do orador.

Conforme se verificou no capítulo 4, o uso de modelos de escoamento permitiu obter ganhos significativos no desempenho dos reconhecedores, à custa da taxa de rejeição, mas não produziu melhorias nas taxas de reconhecimento, quer dos oradores nativos quer dos não nativos. Contudo, as taxas definidas não contabilizaram de forma positiva a possibilidade de algumas palavras-chave, pronunciadas distorcidamente pelos oradores não nativos, poderem ter sido identificadas como palavras estranhas em vez de outras

palavras-chave de substituição. Como se referiu, esta *supressão* de uma palavra-chave é, na maior parte dos casos, preferível à sua substituição por outra palavra-chave, a qual poderá ter graves consequências no desempenho da aplicação. Uma deficiente aprendizagem da segunda língua, poderá determinar a existência de mais substituições, tornando o uso de modelos de escoamento ainda mais vantajoso. Conforme se referiu no capítulo 3, o corpus de fala disponível foi construído com base numa selecção criteriosa de oradores, a maioria dos quais com bons conhecimentos de inglês. Para a realização de experiências em condições mais próximas das aplicações práticas, será importante dispor-se de um corpora recolhido de forma menos controlada, nomeadamente no que se refere aos conhecimentos do orador sobre a segunda língua.

A curto-prazo, pretende-se realizar algumas experiências com o corpus "Multi-English Speech Database" (COST232). Este corpus possui algumas das propriedades necessárias a este estudo, permitindo experiências em condições muito próximas das aplicações práticas, uma vez que foi recolhido através da rede telefónica.

A médio-prazo, espera-se extender parte deste trabalho ao reconhecimento de fala contínua. Para este efeito, o corpus "Translanguage English Database" (TED) apresenta-se actualmente como o mais adequado.

Na sequência desta dissertação, será interessante utilizar uma extensão dos modelos de transcrição como modelo de escoamento na detecção de palavras-chave. A maior dificuldade reside no dimensionamento do parâmetro N_f, bem como na complexidade do modelo daí resultante. Uma forma de simplificar este problema será a de considerar $N_f = 1$ e permitir transições do estado final para o estado inicial, degenerando o modelo num descodificador fonético convencional (figura 5.8).

Da inspecção das transcrições implícitas no modelo de transcrição é possível concluir que o fone mais provável, no início e no fim da transcrição, é o correspondente ao silêncio. Além disso, as inserções de fones detectadas não são relevantes, uma vez que a probabilidade de durarem mais do que 10ms é muito baixa. Tais factos inspiraram a modificação do modelo de transcrição original por forma a obter-se um modelo mais simples. Este novo modelo será testado em futuras experiências. Outra alteração, que poderá vir a ser introduzida, consiste na utilização da informação recolhida num bigrama para fones, como estimativa inicial para as probabilidades de transição interfones. Com estas alterações, prevê-se uma redução substancial na complexidade do modelo e, consequentemente, no respectivo tempo de processamento. Com vista à construção de léxicos multipronúncia dependentes do sotaque, prevê-se, de igual modo, uma maior simplificação nos processos de análise do modelo descritos na secção 5.5.7.

Não se conhecem alterações notórias nas transcrições fonéticas, directamente relacionadas com o sexo do orador. As experiências realizadas até agora consideraram modelos de transcrição distintos para oradores de cada sexo. Para as experiências seguintes prevê-se o treino em paralelo destes modelos, utilizando o conceito de ligação de parâmetros (subsecção 2.4.4). Este conceito será utilizado para ligar as transições interfones, por forma a serem independentes do sexo.

O corpus de fala utilizado neste trabalho possui dois conjuntos de oradores, de cardinais aproximadamente iguais, cujas línguas maternas pertencem, respectivamente, a duas famílias de línguas do ramo europeu: a românica e a germânica. Em certos casos, em vez de se considerar modelos distintos para cada língua materna poderá ser vantajoso considerar apenas modelos distintos para cada das famílias de línguas. Esta poderá ser uma das estratégias a investigar no sentido de se obterem reconhecedores independentes do sotaque do orador.

Neste trabalho, a identificação do sotaque foi realizada com o uso de modelos específicos para cada sotaque. Cada um destes modelos foi treinado com locuções do vocabulário da língua franca, proferidas por oradores com a mesma língua materna. Uma outra possibilidade seria a de treinar cada um destes modelos com vocabulários próprios de cada língua materna. Esta solução é particularmente importante, se se pretender efectuar a identificação do sotaque com o objectivo de seleccionar um reconhecedor para a correspondente língua materna. Neste caso, pressupõe-se a existência de um corpus de fala para cada uma das línguas, utilizado no treino dos respectivos reconhecedores. Seria assim importante, averiguar a utilidade destes corpora para o desenvolvimento de um identificador de sotaques. Tal poderia ser realizado utilizando as topologias e restantes formalismos descritos nesta dissertação.

Algumas características da fala que deverão ser consideradas em futuros desenvolvimentos, nomeadamente, na identificação do sotaque, são as relacionadas com a prosódia (Delmonte, 1998; Sundström, 1998) e em particular a entoação (Auberg et al., 1998; Jilka e Möhler, 1998; Mennen, 1998).

O ensino da língua estrangeira (Frias, 1992) é uma área com tradição secular na qual as novas tecnologias da fala têm tido dificuldades de afirmação (Bernstein et al., 1990; Franco et al., 1997; Bernstein, 1998; Price, 1998). Contudo, existem actualmente neste domínio alguns sistemas interessantes (Dalby et al., 1998; Neumeyer et al., 1998). Nesta área é importante motivar e avaliar a capacidade dos estudantes articularem correctamente determinadas palavras ou frases. No contexto do presente trabalho, seria interessante a aplicação do modelo de transcrição para este fim. Não se pretendendo aprofundar

aqui este assunto, salientam-se apenas alguns aspectos a serem estudados: a aferição dos reconhecedores, quer com oradores nativos, quer com oradores estrangeiros; a identificação de alternativas ao critério da máxima verosimilhança, mais adequadas ao fim em vista (Townshend et al., 1998; Witt e Young, 1998). Entre as variáveis de controlo possíveis no grupo de avaliados salientam-se: o grau de conhecimento prévio das expressões a reproduzir; a duração do ensino ou a exposição com a língua estrangeira; o tipo de ensino da língua estrangeira empregue, nomeadamente entre os designados métodos tradicionais ou indirectos e os métodos directos.

Para finalizar, procura-se dar uma perspectiva sumária da evolução da tecnologia da fala. O telefone é hoje, para muitas pessoas, praticamente uma extensão dos seus aparelhos fonadores e auditivos e é, por certo, o primeiro e o mais importante marco desta tecnologia. Os sintetizadores e os reconhecedores automáticos deverão ser, no futuro, as extensões dos incapacitados visuais e auditivos, respectivamente. Além disso, permitirão um acesso mais natural da restante população a máquinas cada vez mais sofisticadas. O estudo dos problemas relacionados com a multilingualidade têm um papel fundamental no desenvolvimento da tecnologia da fala, numa sociedade da informação e de comunicação que se adivinha à escala planetária. Tal como o telefone, poderão dar um contributo significativo para o entendimento dos novos problemas desta sociedade, para os quais se desejam soluções atempadas.

Apêndice A

Léxicos de pronúncia utilizados neste trabalho

A.1 Léxico de pronúncia para o treino de modelos de fones

ABBREVIATED_DIALLING	ae b r eh v ih ey td ih dd d ay l ih ng
ACTIVE_BOOK	ae k td ih v b uh kd
ADDRESS-BOOK	ae d r eh s b uh kd
ADDRESSES	ae d r eh s ih z
ALARM	ax l aa r m
ALARM-CLOCK	ax l aa r m kd l aa kd
ALARM_CALL	ax l aa r m k ao l
AUTOMATIC	ao t ah m ae t ih kd
BITMAP-EDITOR	b ih td m ae pd eh d ih td er
BOOKING	b uh k ih ng
BUSY	b ih z iy
CALCULATOR	k aw l k y uh l ey td er
CALENDAR	k ae l eh n d er
CALL	k ao l

CALL_THE_	k ao l dh ax
CANTEEN	k ae n t iy n
CHECK	td ch eh kd
CLOCK	kd l aa kd
COMPUTER_CENTRE	k aa m p y uw t er s eh n t er
CONFERENCE_WITH_	k aa n f er ax n s w ih dh
CONFERENCE_WITH_THE_	k aa n f er ax n s w ih dh dh ax
CONFIGURE	k aa n f ih g y uh r
CONTENTS	k aa n t eh n t z
CONTOUR	k aa n t uh r
CUT	k ah td
D-N-C	dd iy sil eh n sil s iy
D-X-F	d iy sil eh k s sil eh f
DE-ACTIVATE	d iy ae kd t ih v ey td
DIAL	d ay ae l
DIARY	d ay er iy
DISK	d ih s kd
DIVERSE	d ay v er s
DIVERT_TO_	d ay v er td t uw
DIVERT_TO_THE_	d ay v er td t uw dh ax
DIVIDED_BY_	d ih v ay d ih d b ay
DONE	d ow n
DOT	d aa td
DOWN	dd aw n
EDIT	eh d ih td
EDITOR	eh d ih td er
EIGHT	ey td

EQUALS	eh k w ax l z
FACSIMILE	f ae kd s ih m ih l
FILE	f ay l
FILE_SYSTEM	f ay l s ih s t eh m
FIND	f aa iy n dd
FIRST	f er s td
FIRST_AID	f er s td ey dd
FIVE	f iy v
FIXED_DESTINATION	f ih k s td d eh s t ih n ey sh ax n
FOLLOW	f aa l ow
FOLLOW_LINK	f aa l ow l ih ng kd
FOUR	f ao r
GO_BACK	g ow b ae kd
HANG-UP	hh ae ng ah pd
HOSPITAL	hh aa s pd ih t el
HOST	hh ow s td
HYPERLINKS	hh ay p er l ih ng kd z
I-O	ay sil aa
IMMEDIATE	ih m iy dd ih ey td
INDEX	ih n d eh kd s
IN_TRAY	ih n t r ey
KEYPAD	k iy p ae dd
LAST	l ae s td
LEFT	l eh f td
LIBRARY	l ay b r aa r iy
LINK	l ih ng kd
M-I	eh m sil ay

MAIN_MENU	m ey n m eh n y uw
MAKE_LINK	m ey k l ih ng kd
MEETING_ROOM	m iy td ih ng r uw m
MINUS	m ay n ah s
MIRROR	m ih r er
MISCELLANEOUS	m ih s eh l ey n iy ax s
MODE	m ow dd
MOVE	m uw v
MULTIPLIED_BY_	m ah l t ih p l ay d b ay
NEXT	n eh k s td
NINE	n ay n
NOTEPAD	n ow t eh pd ae dd
NO_REPLY	n ow r eh p l ay
NUMBER	n ah m b er
OH	hh ao w
ONE	w ah n
OTHER	ah dh er
OUTGOING_RESTRICTION	aw t g ow ih ng r eh s t r ih k sh ax n
PAGE	p ey jh
PEN	p eh n
PERSON	p er s en
PERSONNEL_DEPARTMENT	p er s aa n eh l d ih p aa r t m eh n td
PHONE	f ow n
PLUS	p l ah s
POINT	p oy n td
PREVIOUS	p r iy v ih ax s
PRINT	p r ih n td

PROGRAM	p r aa g r ae m
PURCHASING_DEPARTMENT	p er ch ey z ih ng d ih p aa r t m eh n td
RAILWAY_STATION	r ey l w ey s t ey sh ax n
RECORDER	r iy k ao r d er
REDIAL	r iy d ay el
REDISPLAY	r iy d ih s p l ey
REDO	r iy d ow
RIGHT	r ay td
RUBBER	r ah b er
RUB_OUT	r ah b aw td
SAVE	s ey v
SEARCH	s er td ch
SECURITY	s iy k y uh r iy t ih
SELECTOR	s iy l eh k td er
SET	s eh td
SETTINGS	s eh td ix ng z
SET_ALARM	s eh td ax l aa r m
SET_CLOCK	s eh td k l aa kd
SEVEN	s eh v eh n
SHIFT_LEFT	sh ih f td l eh f td
SHIFT_RIGHT	sh ih f t r ay td
SHOW	sh ow
SHOW_ROOM	sh ow r uw m
SIX	s ih k s
SOUND	s aw n dd
SPREADSHEET	s p r eh dd sh iy td
STARTING_POINT	s t aa r td ih ng p oy n td

START_MICRO	s t aa r td m ay k r ow
START_VOICE	s t aa r td v oy s
STOCK_EXCHANGE	s t aa k eh k s td ch ey n dd jh
STOP	s t aa pd
STOP_MICRO	s t aa pd m ay k r ow
STOP_VOICE	s t aa pd v oy s
STYLUS	s t ay l ah s
TAXI	t ae kd s ih
TECHNICAL_MANAGER	t eh kd n ih k el m ae n ih jh er
TEN	t eh n
TEXT-EDITOR	t eh k s t eh d ih td er
THAT	dh ae td
THREE	th r iy
TIME	t ay m
TIMES	t ay m z
TRAINING	t r ey n ih ng
TRANSFER	t r ae n s f er
TRANSFER_TO_	t r ae n s f er t uw
TRANSFER_TO_THE_	t r ae n s f er t uw dh ax
TRANSFORMATION	t r ae n s f ao r m ey sh ax n
TRAVEL_AGENCY	t r ae v el ey jh eh n s iy
TWO	t uw
UP	hh ah pd
VARIOUS	v eh r ih ax s
WAKE_UP	w ey k ah pd
WALK	w ao kd
ZERO	z ix r ow

A.2 Léxico de pronúncia utilizado nos testes

ACCEPT	ae k s eh pd td
ACTIVATE	ae kd td ih v ey td
AGAIN	ax g eh n
AIRPORT	eh r p ao r td
ALTERNATIVE	ao l t er n ae t ih v
ANNOTATE	ae n aa t ey td
BACKWARD	b ae k w ao r dd
BIGGER	b iy g er
CANCEL	k ae n s el
CLOSE	k l ow z
CONNECT	k aa n eh k td
CONTINUE	k aa n t ih n y uw
CONVERT	k aa n v er td
COPY	k aa p iy
DELETE	dd ix l iy td
DIRECTORY	d ih r eh kd t ao r iy
DIVERSION	dd ay v er sh ax n
ERASE	eh r ey z
EXTEND	eh k s t eh n dd
FORWARD	f ao r w ao r dd
FRIDAY	f r ay d ey
HELP	hh eh l pd
IMPORT	ih m p ao r td
INSERT	ih n s er td

INTERRUPTION	ih n t er ah p sh ax n
KEYBOARD	k iy b ao r dd
LOAD	l ow dd
MESSAGES	m eh s ix dd jh z
MONDAY	m ah n dd ey
NO	n ow
NOTEBOOK	n ow t eh b uh kd
OKAY	ow kd ey
OPEN	ow pd eh n
OPERATOR	ow p er ey t er
OUTPUT	aw t p uh td
PASTE	p ae s td
QUIT	k w ih td
REPEAT	r iy p iy td
RETURN	r iy td er n
SATURDAY	s ae t er d ey
SEND	s eh n dd
SETUP	s eh t ah pd
SMALLER	s m ao l er
SPLIT	s p l ih td
SUNDAY	s ah n d ey
THURSDAY	th er z d ey
TUESDAY	t uw eh s d ey
UNDO	ah n d uw
WEDNESDAY	w eh n z d ey
YES	y eh s

Bibliografia

Abrantes, A. J. C. (1992). Modelamento híbrido da fala com sinusóides e funções de base de banda estreita. Tese de Mestrado, Instituto Superior Técnico, Lisboa.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, AC-19(6):716–723.

Akmajian, A., Demers, R., Farmer, A., Harnish, R. (1990a). *Language Change*, capítulo 8. Em (Akmajian et al., 1990c).

Akmajian, A., Demers, R., Farmer, A., Harnish, R. (1990b). *Language Varieties*, capítulo 7. Em (Akmajian et al., 1990c).

Akmajian, A., Demers, R., Farmer, A., Harnish, R. (1990c). *Linguistics , An Introduction to Language and Communication*. The MIT Press, Cambridge, Massachusetts, London.

Almeida, L. B. (1982). *Modelamento Espectral Não-estacionário da Fala Vozeada*. Tese de Doutoramento, Instituto Superior Técnico, Lisboa.

Almeida, L. B. (1993). An introduction to multilayer perceptrons. *Técnica, revista de Engenharia*, págs. 67–87. número único de 1992.

Andersen, B., Kristiansen, J., Lindberg, B., Pelaez, C., Rigosi, F., Teixeira, C., Trancoso, I. (1992). Second generation recogniser. Relatório técnico, SUNSTAR Esprit Project 2094, Aalborg.

Andersen, O. e Dalsgaard, P. (1992). SAM-IES-059. DKISALA V1.1 — Users Guide. Relatório técnico, Institute of Electronic Systems, Universidade de Aalborg.

Arslan, L. M. (1996). *Automatic Foreign Accent Classification in American English*. Tese de Doutoramento, Universidade de Duke — Escola de Engenharia, Durham — Carolina do Norte (E.U.A.).

Arslan, L. M. e Hansen, J. H. (1996). Language accent classification in American English. *Speech Communication*, 18(4):353–367.

Asadi, A., Schwartz, R., Makhoul, J. (1990). Automatic detection of new words in a large-vocabulary continuous speech recognition system. Em *Proc. Int. Conf. on Acoustic Speech and Signal Processing*, volume 1, págs. 125–128, Albuquerque.

Asadi, A., Schwartz, R., Makhoul, J. (1991). Automatic modeling for adding new words to a large-vocabulary continuous speech recognition system. Em *Proc. Int. Conf. on Acoustic Speech and Signal Processing*, págs. 305–308, Toronto.

Auberg, S., Correa, N., Rothenberg, M., Shanahan, M. (1998). Vowel and intonation training in an English pronunciation tutor. Em *ESCA Workshop on Speech Techonology in Language Learning (STiLL 98)*, págs. 69–72, Estocolmo.

Bacchiani, M., Ostendorf, M., Sagisaka, Y., Paliwal, K. (1996). Design of a speech recognition system based on acoustically derived segmental units. Em *Proc. Int. Conf. on Acoustic Speech and Signal Processing*, págs. 443–446, Atlanta.

Bahl, L., Brown, P., Sousa, P., Mercer, R., Picheny, M. (1993). A method for the construction of acoustic Markov models for words. *IEEE Transactions on Speech and Audio Processing*, 1(4):443–452.

Baker, J. (1975). The Dragon system — an overview. *IEEE Transactions on Acoustics, Speech and Signal Processing*, ASSP-23(1):24–29.

Barbosa, J. (1994). *Introdução ao Estudo da Fonologia e Morfologia do Português*. Livraria Almedina, Coimbra.

Barry, W. J. e Fourcin, A. J. (1992). Levels of labelling. *Computer Speech and Language*, 6(1):1–14.

Barry, W. J., Hoequist, C., Nolan, F. (1989). An approach to the problem of regional accent in automatic speech recognition. *Computer Speech and Language*, 3:335–366.

Baum, L. (1972). An inequality and associated maximization technique in statistical estimation of probabilistic functions of Markov processes. *Inequalities*, 3:1–8.

Baum, L., Petrie, T., Soules, G., Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Ann. Math. Stat.*, 41:164–171.

Bellegarda, J. R. e Nahamoo, D. (1990). Tied mixture continuous parameter modeling for speech recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing*, ASSP-38(12):2033–2045.

Benoît, C. e Campbell, R., editores (1997). *Proceedings of the Workshop on Audio-Visual Speech Processing*. ESCA / ESCOP, Rodes – Grécia.

Berkling, K. M., Arai, T., Barnard, E. (1994). Analysis of phoneme-based features for language identification. Em *Proc. Int. Conf. on Acoustic Speech and Signal Processing*, volume 1, págs. 289–292, Adelaide.

Bernstein, J. (1998). New uses for speech technology in language education. Em *ESCA Workshop on Speech Techonology in Language Learning (STiLL 98)*, págs. 175–178, Estocolmo.

Bernstein, J., Cohen, M., Murveit, H., Rtischev, D., Weintraub, M. (1990). Automatic evaluation and training in English pronunciation. Em *Proc. Int. Conf. on Spoken Language Processing*, págs. 1185–1188.

Best, C. T. (1995). *A Direct Realistic View of Cross-Language Speech Perception*, capítulo 6, págs. 171–204. Em (Strange, 1995b).

Bűhler, C. (1962). *Psychologie im Leben unserer Zeit*. Dromersche Verlagsanstalt Th. Knaur Nachf, Munique/Zurique, 2a edição. Tradução Portuguesa: Fundação Calouste Gulbenkian, Lisboa.

Blackburn, C., Vonwiller, J., King, R. W. (1993). Automatic accent classification using artificial neural networks. Em *Proc. of the European Conf. on Speech Comm. and Tech.*, volume 2, págs. 1241–1244, Berlim. ESCA.

Bohn, O.-S. (1995). What determines the perceptual difficulty encountered in the acquisition of non-native contrasts? Em *ICPHS*, volume 4, págs. 84–91, Estocolmo.

Boll, S. (1979). Suppression of acoustic noise in speech using spectral subtraction. *IEEE Transactions on Acoustics, Speech and Signal Processing*, ASSP-27:113–120.

Bonaventura, P., Gallocchio, F., Mari, J., Micca, G. (1998). Speech recognition methods for non-native pronunciation variation. Em *Proceedings of the Workshop Modeling Pronunciation Variation for Automatic Speech Recognition*, págs. 17–22, Rolduc.

Bourlard, H. e Wellekens, C. (1988). Links between Markov models and multilayer perceptrons. Em *Proceedings NIPS 88*, págs. 502–510.

Boysson-Bardies, B. e Halle, P. (1995). Some suggestions for future research in speech acquisition. *European Studies in Phonetics and Speech Communication*, págs. 90–92.

Brázio, J. M. L., Moreira, H. F. O., Simões, F. E. R. (1979). Câmara Anecóica do Centro de Análise e Processamento de Sinais das Universidades de Lisboa. Em *1º Simpósio Luso-Espanhol de Acústica Ambiental*, volume 2, pág. 19, Lisboa.

Brousseau, J. e Fox, S. (1992). Dialect-dependent speech recognisers for Canadian and European French. Em *Proc. Int. Conf. on Spoken Language Processing*, volume 2, págs. 1003–1006, Banff – Canadá.

Byrne, W., Knodt, E., Khudanpur, S., Bernstein, J. (1998). Is automatic speech recognition ready for non-native speech? a data collection effort and initial experiments in modeling conversational Hispanic English. Em *ESCA Workshop on Speech Techonology in Language Learning (STiLL 98)*, págs. 37–40, Estocolmo.

Caseiro, D. A. (1998). Identificação automática da língua em fala contínua. Tese de Mestrado, Instituto Superior Técnico, Lisboa.

Chigier, B. (1992). Rejection and keyword spotting algorithm for a directory assistance city name recognition application. Em *Proc. Int. Conf. on Acoustic Speech and Signal Processing*, págs. 93–96, São Francisco.

Childers, D., Wu, K., Bae, K., Hicks, D. (1988). Automatic recognition of gender by voice. Em *Proc. Int. Conf. on Acoustic Speech and Signal Processing*, págs. 603–606, Nova Iorque.

Chollet, G. e Montacie, C. (1987). Evaluating speech recognisers and data bases. Em *NATO ASI*, pág. 4, Bad Windsheim.

Christensen, H. e Lindberg, B. (1992). Design and recording of the SAMOGO speech database. Relatório técnico, SUNSTAR Esprit Project 2094, Aalborg.

Ciocea, S., Dufranne, M., Schoentgen, J., Beeckmans, R. (1998). A multi-modal software interface for teaching phonetic transcription. Em *ESCA Workshop on Speech Techonology in Language Learning (STiLL 98)*, págs. 127–130, Estocolmo.

Clary, G. J. e Hansen, J. H. L. (1992). A novel speech recognizer for keyword spotting. Em *Proc. Int. Conf. on Spoken Language Processing*, volume 1, págs. 13–16, Banff – Canadá.

Cole, R., Carmell, T., Connors, P., Macon, M., Wouters, J., de Villiers, J., Tarachow, A., Massaro, D., Cohen, M., Beskow, J., Yang, J., Meier, U., Waibel, A., Stone, P., Fortier, G., Davis, A., Soland, C. (1998). Intelligent animated agents for interactive language training. Em *ESCA Workshop on Speech Techonology in Language Learning (STiLL 98)*, págs. 163–166, Estocolmo.

Cole, R., Hirschman, L., Atlas, L., Beckman, M., Bierman, A., Bush, M., Clements, M., Cohen, J., Garcia, O., Hanson, B., Hermansky, H., Levinson, S., McKeown, K., Morgan, N., Ostendorf, D. G. N. M., Oviatt, S., Price, P., Silverman, H., Spitz, J., Waibel, A., Weinstein, C., Zahorian, S., Zue, V. (1995). The challenge of spoken language systems: Research directions for the nineties. *IEEE Transactions on Speech and Audio Processing*, 3(1):1–21.

Compernolle, D. V. (1989). Noise adaptation in a hidden Markov model speech recognition system. *Computer Speech and Language*, 3:151–167.

Compernolle, D. V. (1997). Speech recognition in the car — from phone dialing to car navigation. Em *Proc. of the European Conf. on Speech Comm. and Tech.*, págs. 2431—2434, Rodes – Grécia. ESCA.

Cook, G. e Robinson, A. J. (1995). Utterance clustering for large vocabulary continuous speech recognition. Em *Proc. of the European Conf. on Speech Comm. and Tech.*, volume 1, págs. 219–222, Madrid.

Dahl, M., Claesson, I., Nordebo, S. (1997). Simultaneous echo cancellation and car noise supression employing a microphone array. Em *Proc. Int. Conf. on Acoustic Speech and Signal Processing*, Munique.

Dalby, J., Kewley-Port, D., Sillings, R. (1998). Language-specific pronunciation training using the HearSay system. Em *ESCA Workshop on Speech Techonology in Language Learning (STiLL 98)*, págs. 25–28, Estocolmo.

Dalsgaard, P. e Andersen, O. (1992). Identification of mono- and poly-phonemes using acoustic-phonetic features derived by a self-organising neural-network. Em *Proc. Int. Conf. on Spoken Language Processing*, volume 1, págs. 547–550, Banff – Canadá.

Davis, S. B. e Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech and Signal Processing*, ASSP-28(4):67–72.

Delmonte, R. (1998). Prosodic modeling for automatic language tutors. Em *ESCA Workshop on Speech Techonology in Language Learning (STiLL 98)*, págs. 57–60, Estocolmo.

Duda, R. O. e Hart, P. E. (1973). *Pattern Classification and Scene Analysis (Part I)*. John Wiley & Sons.

Dudley, H. (1939). The Vocoder. *Bell Labs Rec.*, (17):122–126.

El meliani, R. e O'Shaughnessy, D. (1995). Lexical fillers for task independent-training based keyword spotting and detection of new words. Em *Proc. of the European Conf. on Speech Comm. and Tech.*, volume 3, págs. 2129–2132, Madrid.

El meliani, R. e O'Shaughnessy, D. (1998). Specific language modelling for new-word detection in continuous speech recognition. Em *Proc. Int. Conf. on Acoustic Speech and Signal Processing*, Seattle.

Elliot, S. J. e Nelson, P. A. (1993). Active noise control. *IEEE Signal Processing Magazine*, págs. 12–35.

Euler, S. (1990). *Clustering of Gaussian densities in hidden Markov models*, capítulo 1, págs. 83–88. Em (Laface e de Mori, 1990).

Feng, M.-W. e Mazor, B. (1992). Continuous word spotting for applications in telecommunications. Em *Proc. Int. Conf. on Spoken Language Processing*, volume 1, págs. 21–24, Banff – Canadá.

Finke, M. e Rogina, I. (1997). Wide context acoustic modeling in read vs. spontaneous speech. Em *Proc. Int. Conf. on Acoustic Speech and Signal Processing*, págs. 1743–1746, Munique.

Flege, J. E. (1984). The detection of French accent by American listeners. *Journal of the Acoustical Society of America*, 76(9):692–707.

Flege, J. E. (1995). *Second Language Speech learning:Theory, Findings and problems*, capítulo 8, págs. 233–277. Em (Strange, 1995b).

Flege, J. E. (1998). Second language learning: The role of subject and phonetic variables. Em *ESCA Workshop on Speech Techonology in Language Learning (STiLL 98)*, págs. 1–8, Estocolmo.

Franco, H., Neumeyer, L., Kim, Y., Ronen, O. (1997). Automatic pronunciation scoring for language instruction. Em *Proc. Int. Conf. on Acoustic Speech and Signal Processing*, Munique.

Franco, H. e Serralheiro, A. (1991). Training HMMs using a minimum recognition error approach. Em *Proc. Int. Conf. on Acoustic Speech and Signal Processing*, págs. 357–360, Toronto.

Frias, M. (1992). *Língua Materna — Língua Estrangeira, Uma Relação Multidimensional.* Colecção Mundo de Saberes. Porto Editora, Lda.

Fukada, T., Bacchiani, M., Paliwal, K. K., Sagisaka, Y. (1996). Speech recognition based on acoustically derived segment units. Em *Proc. Int. Conf. on Spoken Language Processing*, Filadélfia.

Furui, S. (1997). Recent advances in robust speech recognition. Em *Proceedings of the ESCA - NATO Tutorial and Research Workshop on Robust Speech Recognition for Unknown Communication Channels*, págs. 11–22, Nancy. ESCA - NATO.

Fussel, J. (1991). Automatic sex identification from short segments of speech. Em *Proc. Int. Conf. on Acoustic Speech and Signal Processing*, págs. 409–412, Toronto.

Gersh, W. (1970). Spectral analysis of EEG's by autorregressive decomposition of time series. *Mathematical Biosciences*, 7:191–204.

Gibbon, D., Moore, R., Winski, R., editores (1997). *Handbook on Standards and Resources for Spoken Language Systems.* Mouton De Gruyter, Berlim.

Gillik, L., Ito, Y., Young, J. (1997). A probabilistic approach to confidence estimation and evaluation. Em *Proc. Int. Conf. on Acoustic Speech and Signal Processing*, págs. 879–882, Munique.

Gish, H., Ng, K., Rohlicek, J. R. (1992). Secondary processing using speech segments for an HMM word spotting system. Em *Proc. Int. Conf. on Spoken Language Processing*, volume 1, págs. 17–20, Banff – Canadá.

Gupta, V. e Mermelstein, P. (1982). Effects of speaker accent on the performance of a speaker-independent, isolated-word recognizer. *Journal of the Acoustical Society of America*, 71:1581–1587.

Haeb-Umbach, R. (1997). Robust speech recognition for wireless networks and mobile telephony. Em *Proc. of the European Conf. on Speech Comm. and Tech.*, págs. 2427–2430, Rodes – Grécia. ESCA.

Haeb-Umbach, R., Beyerlein, P., Thelen, E. (1995). Automatic transcription of unknown words in a speech recognition system. Em *Proc. Int. Conf. on Acoustic Speech and Signal Processing*, volume 1, págs. 840–843, Detroit.

Hansen, J. H. L. e Arslan, L. (1995). Foreign accent classification using source generator based prosodic features. Em *Proc. Int. Conf. on Acoustic Speech and Signal Processing*, volume 1, págs. 836–839, Detroit.

Harrison, W. A., Lim, J. S., Singer, E. (1984). Adaptive noise cancellation in a fighter cockpit environment. Em *Proc. Int. Conf. on Acoustic Speech and Signal Processing*, volume 2, págs. 18A.4.1–18A.4.4, São Diego.

Hazen, T. J. e Zue, V. W. (1993). Automatic language identification using a segment-based approach. Em *Proc. of the European Conf. on Speech Comm. and Tech.*, volume 2, págs. 1303–1306, Berlim.

Hermanski, H. (1990a). Auditory model for parameterization of speech in real-life environment based on re-integration of temporal derivative of auditory spectrum. Relatório Técnico File Folder ST 04-01, U.S. WEST Advanced Technologies Research Report.

Hermanski, H. (1990b). Perceptual linear predictive (PLP) analysis of speech. *Journal of the Acoustical Society of America*, págs. 1738–1752.

Hetherington, I. L. (1995). New words: effect on recognition performance and incorporation issues. Em *Proc. of the European Conf. on Speech Comm. and Tech.*, volume 3, págs. 1645–1648, Madrid.

Higgins, A. L. e Wohlford, R. E. (1985). Keyword recognition using template concatenation. Em *Proc. Int. Conf. on Acoustic Speech and Signal Processing*, volume 3, págs. 1233–1236, Tampa, Flórida.

House, A. S. e Neuburg, E. P. (1977). Toward automatic identification of the language of an utterance. I. Preliminary methodological considerations. *Journal of the Acoustical Society of America*, 62(3):708–713.

Hu, Z., Schalkwyk, J., Barnard, E., Cole, R. (1996). Speech recognition using syllable-like units. Em *Proc. Int. Conf. on Spoken Language Processing*, Filadélfia.

Huang, X. D. (1998). Spoken language technology research at Microsoft. Em *Proc. ICA/ASA '98*, Seattle.

Huang, X. D., Acero, A., Alleva, F., Hwang, M.-Y., Jiang, L., Mahajan, M. (1995). Microsoft Windows highly intelligent speech recognizer: Whisper. Em *Proc. Int. Conf. on Acoustic Speech and Signal Processing*, Detroit.

Huang, X. D., Ariki, Y., Jack, M. (1990). *Hidden Markov models for speech recognition.* Information Technology Series. Edinburgh University Press.

Huang, X. D. e Jack, M. (1989). Semi-continuous Markov models for speech signals. *Computer Speech and Language*, 3:239–251.

Humphries, J. e Woodland, P. (1998). The use of accent-specific pronunciation dictionaries in acoustic model training. Em *Proc. Int. Conf. on Acoustic Speech and Signal Processing*, Seattle.

Humphries, J., Woodland, P., Pearce, D. (1996). Using accent-specific pronunciation modelling for robust speech recognition. Em *Proc. Int. Conf. on Spoken Language Processing*, Filadélfia.

Humphries, J., Woodland, P., Pearce, D. (1997). Using accent-specific pronunciation modelling for improved large vocabulary continuous speech recognition. Em *Proc. of the European Conf. on Speech Comm. and Tech.*, págs. 2367–2370, Rodes – Grécia. ESCA.

Irion, J., Riccio, A., Renner, T., Balle, T., Dalsgaard, P., Trancoso, I. (1992). Final report — part 1: Technical achievements. Relatório técnico, SUNSTAR Esprit Project 2094, Aarhus.

Itakura, F. (1975). Minimum prediction residual principle applied to speech recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing*, ASSP-23(1):67–72.

Jacobsen, C. (1992). SIRtrain, an open standard environment for CHMM recognizer development. Em *Proc. Int. Conf. on Spoken Language Processing*, volume 2, págs. 1555–1558, Banff – Canadá.

Jelinek, F. (1976). Continuous speech recognition by statistical methods. *Proceedings IEEE*, 64:532–556.

Jelinek, F. (1985). The development of an experimental discrete dictation recognizer. *Proceedings IEEE*, 73:1616–1624.

Jelinek, F., Bahl, L. R., Mercer, R. L. (1975). Design of a linguistic statistical decoder for the recognition of continuous speech. *IEEE Transactions on Information Theory*, 21:250–256.

Jelinek, F., Lafferty, J. D., Mercer, R. L. (1990). *Basic Methods of Probabilistic Context Free Grammars*, capítulo 4, págs. 345–360. Em (Laface e de Mori, 1990).

Jelinek, F., Mercer, R. L., Roukos, S. (1991). *Principles of Lexical Language Modeling for Speech Recognition*, capítulo 3, págs. 651–699. Number 21. Marcel Dekker, Nova Iorque.

Jilka, M. e Möhler, G. (1998). Intonational foreign accent: speech technology and foreign language teaching. Em *ESCA Workshop on Speech Techonology in Language Learning (STiLL 98)*, págs. 115–118, Estocolmo.

Juang, B.-H., Rabiner, L. R., Wilpon, J. G. (1987). On the use of bandpass liftering in speech recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing*, ASSP-35(7):947–954.

Junkawitsch, J., Ruske, G., Höge, H. (1997). Efficient methods for detecting keywords in continuous speech. Em *Proc. of the European Conf. on Speech Comm. and Tech.*, volume 1, págs. 259–262, Rodes – Grécia. ESCA.

Junqua, J.-C. e Morin, P. (1996). Naturalness of the interaction in multimodal applications. Relatório Técnico 5, Speech Transmission Laboratory, Santa Bárbara, Califórnia.

Kay, S. M. e Marple, S. L. (1981). Spectrum analysis — a modern perspective. *Proceedings IEEE*, 69(11):642–652.

Keating, P. A., Byrd, D., Flemming, E., Todaka, Y. (1994). Phonetic analyses of word and segment variation using the TIMIT corpus of American English. *Speech Communication*, 14:131–142.

Kellermann, W. (1997). Strategies for combining acoustic echo cancellation and adaptive beamforming microphone arrays. Em *Proc. Int. Conf. on Acoustic Speech and Signal Processing*, Munique.

Komori, Y. e Rainton, D. (1992). Minimum error classification training for HMM-based keyword spotting. Em *Proc. Int. Conf. on Spoken Language Processing*, volume 1, págs. 9–12, Banff – Canadá.

Kreyszig, E. (1970). *Introductory Mathematical Statistics — Priciples and Methods.* John Wiley & Sons.

Kubala, F. e Schwartz, R. (1991). A new paradigm for speaker-independent training. Em *Proc. Int. Conf. on Acoustic Speech and Signal Processing*, págs. 833–836, Toronto.

Kuhl, P. K. (1995). The acquisition of language and speech. *European Studies in Phonetics and Speech Communication*, págs. 93–103.

Kullback, S. e Leiber, R. (1951). On information and sufficiency. *Ann. Math. Stat.*, 22:79–86.

Kumpf, K. e King, R. W. (1997a). Automatic accent classification of foreign accented Australian English speech. Em *Proc. Int. Conf. on Spoken Language Processing*, Filadélfia.

Kumpf, K. e King, R. W. (1997b). Foreign speaker accent classification using phoneme-dependent accent discrimination models and comparisons with human perception benchmarks. Em *Proc. of the European Conf. on Speech Comm. and Tech.*, págs. 2323–2326, Rodes – Grécia. ESCA.

Kuwano, H., Nomura, K., Ookumo, A., Hiraoka, S., Watanabe, T., Niyada, K. (1992). Speaker independent speech recognition methods using word spotting technique and its application to VCR programming. Em *Proc. Int. Conf. on Spoken Language Processing*, volume 2, págs. 1527–1530, Banff – Canadá.

Laface, P. e de Mori, R., editores (1990). *Speech Recognition and Understanding — Recent Advances, Trends and Applications*, volume 75 da série *F: Computer and Systems Sciences*. Springer-Verlag, Cetraro (Itália).

Lamel, L. F. e Gauvain, J. L. (1993). Identifying non-linguistic speech features. Em *Proc. of the European Conf. on Speech Comm. and Tech.*, volume 1, págs. 23–30, Berlim. Keynote.

Lamel, L. F., Rabiner, L. R., Rosenberg, A. E., Wilpon, J. G. (1981). An improved endpoint detector for isolated word recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing*, ASSP-29(4):777–785.

Lau, R. e Seneff, S. (1997). Providing sublexical constraints for word spotting within the ANGIE framework. Em *Proc. of the European Conf. on Speech Comm. and Tech.*, volume 1, págs. 263–266, Rodes – Grécia. ESCA.

Lee, C.-H., Rabiner, L. R., Pieraccini, R., Wilpon, J. G. (1990a). Acoustic modeling for large vocabulary speech recognition. *Computer Speech and Language*, 4:127–165.

Lee, C.-H., Soong, F. K., Juang, B.-H. (1988). A segmented model based approach to speech recognition. Em *Proc. Int. Conf. on Acoustic Speech and Signal Processing*, págs. 501–504, Nova Iorque.

Lee, K.-F. (1989). *Automatic Speech Recognition: The Development of the SPHINX System*. Kluwer Academic Publishers, Boston.

Lee, K.-F., Hon, H.-W., Reddy, R. (1990b). An overview of the SPHINX speech recognition system. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 38(1):35–45.

Leeuwen, D. A., van den Berg, L.-G., Steeneken, H. J. M. (1995). Human benchmarks for speaker independent large vocabulary recognition performance. Em *Proc. of the European Conf. on Speech Comm. and Tech.*, volume 2, págs. 1461–1464, Madrid.

Levinson, S. E., Liberman, M. Y., Ljolje, A., Miller, L. G. (1989). Speaker independent phonetic transcription of fluent speech for large vocabulary speech recognition. Em *Proc. Int. Conf. on Acoustic Speech and Signal Processing*, págs. 441–444, Glásgua.

Lin, Q., Jan, E.-E., Flanagan, J. (1994). Microphone arrays and speaker identification. *IEEE Transactions on Speech and Audio Processing*, 2(4):622–629. Special Issue: Robust Speech Recognition.

Linde, Y., Buzo, A., Gray, R. M. (1980). An algorithm for vector quantizer design. *IEEE Transactions on Communications*, COM-28:84–95.

Lippmann, R. P. e Gold, B. (1987). Neural-net classifiers useful for speech recognition. Em *IEEE International Conference on Neural Networks*.

Lleida, E., Mariño, J. B., Salavedra, J., Bonafonte, A., Monte, E., Martínez, A. (1993). Out-of-vocabulary word modelling and rejection for keyword spotting. Em *Proc. of the European Conf. on Speech Comm. and Tech.*, volume 2, págs. 1265–1268, Berlim.

Lleida, E. e Rose, R. C. (1996). Likelihood ratio decoding and confidence measures for continuous speech recognition. Em *Proc. Int. Conf. on Spoken Language Processing*, Filadélfia.

Llisterri, J. (1995). Relationships between speech production and speech perception in a second language. Em *ICPHS*, volume 4, págs. 92–99, Estocolmo.

Lopes, P., Santos, B., Bento, M., Piedade, M. (1998). Controlo activo de ruído e vibração. *Anais da Engenharia e Tecnologia Electrotécnica*, (6):13–16. Ano III.

Lucassen, J. e Mercer, R. (1984). An information theoretic approach to the automatic determination of phonemic baseforms. Em *Proc. Int. Conf. on Acoustic Speech and Signal Processing*, number 1 in 42.5, São Diego.

Lyon, R. e Mead, C. (1988). An analog electronic cochlea. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 36(7):1119–1134.

Makoul, J. (1975). Linear prediction: A tutorial review. *Proceedings IEEE*, 63(4):561–580.

Manos, A. S. e Zue, V. W. (1997). A segmented-based wordspotter using phonetic filler models. Em *Proc. Int. Conf. on Acoustic Speech and Signal Processing*, págs. 899–902, Munique.

Marcus, J. N. (1992). A novel algorithm for HMM word spotting, performance evaluation and error analysis. Em *Proc. Int. Conf. on Acoustic Speech and Signal Processing*, São Francisco.

Mariño, J. B., Nadeu, C., Moreno, A., Lleida, E., Monte, E., Bonafonte, A. (1990). *RAMSES: A Spanish Demisyllable Based Continuous Speech Recognition System*, capítulo 1, págs. 113–118. Em (Laface e de Mori, 1990).

Markel, J. D. e Gray, A. H. (1976). *Linear Prediction of Speech.* Springer-Verlag, Nova Iorque.

Markel, J. D. e Gray, A. H. (1980). Implementation and comparison of two transformed reflection coefficient scalar quantization methods. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 28(5):575–583.

Markham, D. J. e Nagano-Madsen, Y. (1996). Input modality effects in foreign accent. Em *Proc. Int. Conf. on Spoken Language Processing*, Filadélfia.

Marques, J. S. (1994). *Classificação de Dados.* Instituto Superior Técnico, Lisboa, 2ª edição.

Marques, J. S., Almeida, L. B., Tribolet, J. M. (1990). Harmonic coding at 4.8 kb/s. Em *Proc. Int. Conf. on Acoustic Speech and Signal Processing*, volume 1, págs. 17–20, Albuquerque.

Martinet, A. (1967). *Éléments de Linguistique Générale.* Librairie Armand Colin, 2ª edição. Tradução Portuguesa: Livraria Sá da Costa Editora, Lisboa.

Martins, C. A. D. (1998a). Modelos de linguagem no reconhecimento de fala contínua. Tese de Mestrado, Instituto Superior Técnico, Lisboa.

Martins, C. A. R. (1998b). *Canceladores Adaptativos de Ruído — aplicações na aquisição de sinais de fala.* Tese de Doutoramento, Instituto Superior Técnico, Lisboa.

Martins, C. R., Almeida, T. M., Piedade, M. S. (1990). An adaptive noise canceller and its implementation on a DSP. Em *Proceedings of the International Symposium on Circuits and Systems — IEEE ISCAS*, págs. 1939–1942, Nova Orleães.

Mateus, M. H. M., Andrade, A., Viana, M. C., Villalva, A. (1990). *Fonética, Fonologia e Morfologia do Português*, volume 28. Universidade Aberta, Lisboa.

McAllister, R. (1995). Perceptual foreign accent and L2 production. Em *ICPHS*, volume 4, págs. 570–573, Estocolmo.

McAllister, R. (1998). Second language perception and the concept of foreign accent. Em *ESCA Workshop on Speech Techonology in Language Learning (STiLL 98)*, págs. 155–158, Estocolmo.

Mengel, A. (1993). Transcribing names — a multiple choice task: mistakes, pitfalls and escape routes. Em *Proc. 1st ONOMASTICA Research Colloquium*, págs. 5–9, Londres.

Mennen, I. (1998). Can language learners ever acquire the intonation of a second language? Em *ESCA Workshop on Speech Techonology in Language Learning (STiLL 98)*, págs. 17–20, Estocolmo.

Meyer, J. e Simmer, K. U. (1997). Multi-channel speech enhancement in a car environment using Wiener filtering and spectral subtraction. Em *Proc. Int. Conf. on Acoustic Speech and Signal Processing*, Munique.

Meyer, P. L. (1965). *Introductory probability and statistical applications*. Addison-Wesley. Tradução Brasileira: Livros Técnicos e Científicos Editora S.A., Rio de Janeiro, 1980.

Miller, L. G. e Levinson, S. E. (1988). Syntactic analysis for large vocabulary speech recognition using a context-free covering grammar. Em *Proc. Int. Conf. on Acoustic Speech and Signal Processing*, págs. 271–274, Nova Iorque.

Mixdorff, H. (1996). Foreign accent in intonation patterns — a contrastive study applying a quantitative model of F0 contour. Em *Proc. Int. Conf. on Spoken Language Processing*, Filadélfia.

Moore, R. K. (1986). The NATO research study group on speech processing: RSG 10. Em *Proceedings Speech Tech'86*, págs. 201–203. Media Dimensions.

Morgan, N. e Bourlard, H. (1990). Continuous speech recognition using multilayer perceptrons with hidden Markov models. Em *Proc. Int. Conf. on Acoustic Speech and Signal Processing*, volume 1, págs. 413–416, Albuquerque.

Morin, P. e Junqua, J.-C. (1993). Habitable interaction in goal-oriented multimodal dialogue systems. Em *Proc. of the European Conf. on Speech Comm. and Tech.*, volume 3, págs. 1669–1672, Berlim.

Morin, P., Junqua, J.-C., Pierrel, J.-M. (1992). A flexible multimodal dialogue architecture independent of the application. Em *Proc. Int. Conf. on Spoken Language Processing*, volume 2, págs. 939–942, Banff – Canadá.

Muthusamy, Y. K., Barnard, E., Cole, R. A. (1994a). Reviewing automatic language identification. *IEEE Signal Processing Magazine*, págs. 33–41.

Muthusamy, Y. K., Jain, N., Cole, R. A. (1994b). Perceptual benchmarks for automatic language identification. Em *Proc. Int. Conf. on Acoustic Speech and Signal Processing*, volume 1, págs. 333–336, Adelaide.

Nadas, A. (1983). A decision theoretic formulation of a training problem in speech recognition and a comparison of training by unconditional versus conditional maximum likelihood. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 31(4):814–817.

Neto, J., Almeida, L., Hochberg, M., Martins, C., Nunes, L., Renals, S., Robinson, T. (1995). Speaker-adaptation for hybrid HMM-ANN continuous speech recognition system. Em *Proc. of the European Conf. on Speech Comm. and Tech.*, volume 3, págs. 2171–2174, Madrid.

Neto, J., Martins, C., Almeida, L. (1996). An incremental speaker-adaptation technique for hybrid HMM-MLP recognizer. Em *Proc. Int. Conf. on Spoken Language Processing*, págs. 1289–1292, Filadélfia.

Neto, J. P. S. (1998). *Reconhecimento de Fala Contínua com aplicação de Técnicas de Adaptação ao Orador*. Tese de Doutoramento, Instituto Superior Técnico, Lisboa.

Neumeyer, L., Franco, H., Abrash, V., Julia, L., Ronen, O., Bratt, H., Bing, J., Digalakis, V., Rypa, M. (1998). WebGrader^{TM}: A multilingual pronunciation practice tool. Em *ESCA Workshop on Speech Techonology in Language Learning (STiLL 98)*, págs. 61–64, Estocolmo.

Ney, H. (1990). *Stochastic Grammars and Pattern Recognition*, capítulo 4, págs. 319–344. Em (Laface e de Mori, 1990).

O'Kane, M. J. e Kenne, P. E. (1993). Word and phrase spotting with limited training. Em *Proc. of the European Conf. on Speech Comm. and Tech.*, volume 2, págs. 1269–1272, Berlim.

Okawa, S., Kobayashi, T., Shirai, K. (1993). Word spotting in conversational speech based on phoneme unit likelihood by mutual information criterion. Em *Proc. of the European Conf. on Speech Comm. and Tech.*, volume 2, págs. 1281–1284, Berlim.

Oliveira, L. C. (1996). *Síntese de Fala a Partir de Texto*. Tese de Doutoramento, Instituto Superior Técnico, Lisboa.

Oppenheim, A. V. (1974). *Digital Signal Processing*. Prentice Hall.

O'Shaughnessy, D. (1987). *Speech Communication, Human and Machine*. Series in Electrical Engineering: Digital Signal Processing. Addison-Wesley.

Papoulis, A. (1984). *Probability, Random Variables, and Stochastic Processes*. McGraw-Hill International Book Company.

Patterson, R. e Litt, D., editores (1996). *New Webester's Expanded Dictionary*. Paradise Press Inc.

Perdigão, F. M. S. (1997). *Modelos do Sistema Auditivo Periférico no Reconhecimento Automático de Fala*. Tese de Doutoramento, Faculdade de Ciências e Tecnologia, Coimbra.

Pfau, T., Beham, M., Reichl, W., Ruske, G. (1997). Creating large subword units for speech recognition. Em *Proc. of the European Conf. on Speech Comm. and Tech.*, volume 3, págs. 1191–1194, Rodes – Grécia. ESCA.

Polka, L. (1995). Linguistic influences in adult perception of non-native vowel contrasts. *Journal of the Acoustical Society of America*, págs. 1286–1296.

Price, P. (1998). How can speech technology replicate and complement skills of good language teachers in ways that help people to learn language? Em *ESCA Workshop on Speech Techonology in Language Learning (STiLL 98)*, págs. 81–86, Estocolmo.

Príncipe, J. C. e Tracey, J. (1993). Reconhecimento de palavras isoladas por redes neuronais com memória recursiva. Em *Actas do Encontro de Processamento da Língua Portuguesa*, págs. 81–85, Lisboa.

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings IEEE*, págs. 257–286.

Rabiner, L. R. e Juang, B.-H. (1993). *Fundamentals of Speech Recognition*. Signal Processing Series. Prentice Hall.

Ribeiro, C. E. M. (1991). Celp: Codificação de fala a baixos ritmos. Tese de Mestrado, Instituto Superior Técnico, Lisboa.

Ribeiro, C. M., Trancoso, I., Viana, M. C. (1993). EUROM.1 Portuguese Database. Report D6. Relatório técnico, INESC — Instituto de Engenharia de Sistemas e Computadores, Lisboa.

Riley, M. D. (1991). A statistical model for generating pronunciation networks. Em *Proc. Int. Conf. on Acoustic Speech and Signal Processing*, págs. 737–740, Toronto.

Roach, P. e Arnfield, S. (1998). Variation information in pronunciation dictionaries. Em *Proceedings of the Workshop Modeling Pronunciation Variation for Automatic Speech Recognition*, págs. 121–124, Rolduc.

Rochet, B. L. (1995). *Perception and Production of Second-Language Speech Sounds by Adults*, capítulo 13, págs. 379–410. Em (Strange, 1995b).

Rohlicek, J. R., Russel, W., Roukos, S., Gish, H. (1989). Continuous hidden Markov modeling for speaker independent word spotting. Em *Proc. Int. Conf. on Acoustic Speech and Signal Processing*, págs. 627–630, Glásgua.

Rose, R. C. e Lleida, E. (1997). Speech recognition using automatically derived acoustic baseforms. Em *Proc. Int. Conf. on Acoustic Speech and Signal Processing*, Munique.

Rose, R. C., Lleida, E., Erhart, G., Grubbe, R. (1996). A user-configurable system for voice label recognition. Em *Proc. Int. Conf. on Spoken Language Processing*, Filadélfia.

Rose, R. C. e Paul, D. B. (1990). A hidden Markov model based keyword recognition system. Em *Proc. Int. Conf. on Acoustic Speech and Signal Processing*, volume 1, págs. 129–132, Albuquerque.

Saussure, F. (1916). *Cours de Linguistique Générale*. Editions Payot, Paris. Tradução Portuguesa: Publicações Dom Quixote, Lisboa.

Schulte-Pelkum, R. (1976). *Interferenzfehler bei deutshsprechenden Japanern*, págs. 59–111. Em (Mixdorff, 1996). (resumo em inglês das págs. 70-73).

Schwartz, R., Kimball, O., Kubala, F., Feng, M.-W., Chow, Y., Barry, C., Makhoul, J. (1989). Robust smoothing methods for discrete hidden Markov models. Em *Proc. Int. Conf. on Acoustic Speech and Signal Processing*, págs. 548–551, Glásgua.

Segura, J. C., Rubio, A. J., Peinado, A. M., García, P., Román, R. (1994). Multiple vq hidden Markov modelling for speech recognition. *Speech Communication*, 14:163–170.

Serralheiro, A., Martins, C., Ribeiro, C., Teixeira, C., Schulz, M., Trancoso, I. (1991). Echo cancellation and noise reduction. Relatório técnico, SUNSTAR Esprit Project 2094, Lisboa.

Serralheiro, A. J. S. R. (1990). *Metodologias Probabilísticas no Reconhecimento de Palavras Isoladas*. Tese de Doutoramento, Instituto Superior Técnico, Lisboa.

Silva, F. H. C.-R. M. (1989). Técnicas de redução de ruído em sinais de fala. Relatório técnico, INESC — Instituto de Engenharia de Sistemas e Computadores, Lisboa. Relatório Interno.

SIRtrain (1991). *SIRtrain Training Software — User Guide, Vers.2.1.* SUNSTAR Esprit Project 2094, Aarhus.

Steeneken, H. J. M. e Houtgast, T. (1991). On the mutual dependency of octave-band-specific contributions to speech intelligibility. Em *Proc. of the European Conf. on Speech Comm. and Tech.*, volume 3, págs. 1133–1136, Génova.

Steinberg, J. C. (1929). Effects of distortion upon the recognition of speech sounds. *Journal of the Acoustical Society of America*, 1(1):121–137.

Stork, D. G. e Hennecke, M. E., editores (1996). *Speechreading by Humans and Machines – Models, Systems, and Applications*, volume 150 da série *F: Computer and Systems Sciences*. Springer-Verlag, Nova Iorque.

Strange, W. (1995a). Phonetics of second-language acquisition: past, present, future. Em *ICPHS*, volume 4, págs. 76–83, Estocolmo.

Strange, W., editor (1995b). *Speech Perception and Linguistic Experience: Issues in Cross-Language Research*. York Press, Timonium, Maryland.

Strik, H. e Cucchiarini, C. (1998). Modeling pronunciation variation for ASR: Overview and comparison of methods. Em *Proceedings of the Workshop Modeling Pronunciation Variation for Automatic Speech Recognition*, págs. 137–144, Rolduc.

Sundström, A. (1998). Automatic prosody modification as a means for foreign language pronunciation training. Em *ESCA Workshop on Speech Techonology in Language Learning (STiLL 98)*, págs. 49–52, Estocolmo.

Tavares, L. V. e Correia, F. N. (1986). *Optimização linear e não linear — Conceitos, Métodos e Algoritmos*. Fundação Calouste Gulbenkian, Lisboa.

Teixeira, A. e Vaz, F. (1997). A software tool to study Portuguese vowels. Em *Proc. of the European Conf. on Speech Comm. and Tech.*, Rodes – Grécia.

Teixeira, C. (1989). Identificação de um modelo de fonte composta no electroencefalograma. Tese de Mestrado, Instituto Superior Técnico, Lisboa.

Teixeira, C. (1992a). Guidelines for including word rejection capabilities in the SUNSTAR speech recognition applications. Relatório técnico, SUNSTAR Esprit Project 2094, Aalborg.

Teixeira, C. (1992b). Recommendations for the distribution of models and cepstrum files. Relatório técnico, SUNSTAR Esprit Project 2094, Aalborg.

Teixeira, C. e Lindberg, B. (1992). Word rejection experiments on the SUNSTAR multilanguage speech database. Em *Conferência da Associação de Reconhecimento de Padrões*, págs. 77–83, Coimbra.

Teixeira, C. e Trancoso, I. (1990). Functional description of noise reduction module (public). Relatório técnico, SUNSTAR Esprit Project 2094, Lisboa.

Teixeira, C. e Trancoso, I. (1991a). Redução de ruído de sinais de fala por subtracção espectral. Em *Conferência da Associação de Reconhecimento de Padrões*, Aveiro.

Teixeira, C. e Trancoso, I. (1991b). Spectral subtraction for front-end noise reduction in a speech recognizer. Em *Proc. of the European Conf. on Speech Comm. and Tech.*, volume 2, págs. 499–502, Génova.

Teixeira, C. e Trancoso, I. (1992). Word rejection using multiple sink models. Em *Proc. Int. Conf. on Spoken Language Processing*, volume 2, págs. 1443–1446, Banff – Canadá.

Teixeira, C. e Trancoso, I. (1993). Continuous and semi-continuous HMM for recognising non-native pronunciations. Em *Proc. IEEE Workshop ASR*, págs. 26,27, Utah.

Teixeira, C., Trancoso, I., Serralheiro, A. (1992). Single vs. multiple sink models for isolated and connected word recognition. Em *Proc. Speech Processing in Adverse Conditions*, págs. 179–182, Cannes.

Teixeira, C., Trancoso, I., Serralheiro, A. (1993a). On the performance of CHMM and SCHMM for isolated word recognition and rejection. Em *New Advances and Trends in Speech — Recognition and Coding*. NATO ASI, Granada.

Teixeira, C., Trancoso, I., Serralheiro, A. (1996). Accent identification. Em *Proc. Int. Conf. on Spoken Language Processing*, Filadélfia.

Teixeira, C., Trancoso, I., Serralheiro, A. (1997). Recognition of non–native accents. Em *Proc. of the European Conf. on Speech Comm. and Tech.*, volume 5, págs. 2375–2378, Rodes – Grécia.

Teixeira, C. J., Trancoso, I. M., Ribeiro, C. M., Martins, C. A., Serralheiro, A. J., Piedade, M. S. (1993b). Reconhecimento robusto de fala: a experiência do projecto SUNSTAR. Em *Actas do Encontro de Processamento da Língua Portuguesa*, págs. 75–80, Lisboa.

Tomlinson, M. J. (1990). SAM-RSRE-15. guide to database generation — recording protocol. Relatório técnico, RSRE, Malvern.

Townshend, B., Bernstein, J., Todic, O., Warren, E. (1998). Estimation of spoken language proficiency. Em *ESCA Workshop on Speech Techonology in Language Learning (STiLL 98)*, págs. 179–182, Estocolmo.

Trancoso, I. e Moore, R., editores (1995). *Proceedings of the ESCA - NATO Tutorial and Research Workshop on Speech under Stress*. ESCA/ETWR INESC, Lisboa.

Trancoso, I., Teixeira, C., Martins, C., Piedade, M. (1990). General description of modules available from INESC. Relatório técnico, SUNSTAR Esprit Project 2094, Lisboa.

Trancoso, I. e Viana, M. C. (1997). On the pronunciation mode of acronyms in several European languages. Em *Proc. of the European Conf. on Speech Comm. and Tech.*, Rodes – Grécia.

Trancoso, I. M. M. (1987). *Codificação de Fala com Alta Qualidade a Médios e Baixos Ritmos*. Tese de Doutoramento, Instituto Superior Técnico, Lisboa.

Tseng, B. L., Soong, F. K., Rosenberg, A. E. (1992). Continuous probabilistic acoustic map for speaker recognition. Em *Proc. Int. Conf. on Acoustic Speech and Signal Processing*, São Francisco.

Varga, A. e Moore, R. (1990). Hidden Markov model decomposition of speech and noise. Em *Proc. Int. Conf. on Acoustic Speech and Signal Processing*, págs. 845–848, Albuquerque.

Varga, A. P. e Ponting, K. (1989). Control experiments on noise compensation in hidden Markov model based continuous word recognition. Em *Proc. of the European Conf. on Speech Comm. and Tech.*, págs. 167–170, Paris. ESCA.

Vaver, J. G. (1998). Experiments in confidence scoring using Spanish CallHome data. Em *Proc. Int. Conf. on Acoustic Speech and Signal Processing*, Seattle.

Vergin, R., Farhat, A., O'Shaughnessy, D. (1996). Robust gender-dependent acoustic-phonetic modelling in continuous speech recognition based on a new automatic male/female classification. Em *Proc. Int. Conf. on Spoken Language Processing*, Filadélfia.

Vonusa, R. S., Nelson, C. J. T., Smith, S. E., Parker, J. G. (1982). NATO AC/243 (PANEL III RSG10) Language Database. Em *Proc. of the Workshop on Standardization for Speech I/O Technology*, págs. 223–228, Gaithersburg, Maryland. National Bureau of Standards.

Vroomen, J., Collier, R., Mozziconacci, S. (1993). Duration and intonation in emotional speech. Em *Proc. of the European Conf. on Speech Comm. and Tech.*, volume 1, págs. 577–580, Berlim.

Widrow, B., Glover, J. R., McCool, J. M. (1975). Adaptive noise cancelling: Principles and applications. *Proceedings IEEE*, 63:1692–1716.

Wilpon, J. G., Lee, C.-H., Rabiner, L. R. (1989). Application of hidden Markov models for recognition of a limited set of words in unconstrained speech. Em *Proc. Int. Conf. on Acoustic Speech and Signal Processing*, págs. 254–257, Glásgua.

Wilpon, J. G., Miller, L. G., Modi, P. (1991). Improvements and applications for key word recognition using hidden Markov modelling techniques. Em *Proc. Int. Conf. on Acoustic Speech and Signal Processing*, págs. 309–312, Toronto.

Wilpon, J. G., Rabiner, L. R., Lee, C.-H., Goldman, E. R. (1990). Automatic recognition of keywords in unconstrained speech using hidden Markov models. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 38(11):1870–1878.

Witt, S. e Young, S. (1998). Performance measures for phone-level pronunciation teaching in CALL. Em *ESCA Workshop on Speech Techonology in Language Learning (STiLL 98)*, págs. 99–103, Estocolmo.

Young, S. J. e Bloothooft, G. (1997). *Corpus-based Methods in Language and Speech Processing*. Kluwer Academic Publishers.

Young, S. J., Jansen, J., Odell, J., Ollason, D., Woodland, P. (1996). *The HTK Book*. Cambridge University.

Young, S. J., Russel, N., Thornon, J. (1991). The use of syntax and multiple alternatives in the VODIS voice operated database inquiry system. *Computer Speech and Language*, 5:65–80.

Young, S. J. e Woodland, P. (1993). The use of state tying in continuous speech recognition. Em *Proc. of the European Conf. on Speech Comm. and Tech.*, volume 3, págs. 2203–6, Berlim.

Yu, H.-J. e Oh, Y.-H. (1997). A neural network for 500 vocabulary word spotting using acoustic sub-word units. Em *Proc. Int. Conf. on Acoustic Speech and Signal Processing*, págs. 3277–3380, Munique.

Zetterberg, L. H. (1969). Estimation of parameters for a linear difference equation with application to EEG analysis. *Mathematical Biosciences*, 5:227–275.

Zissman, M. (1993). Automatic language identification using gaussian mixture and hidden Markov models. Em *Proc. Int. Conf. on Acoustic Speech and Signal Processing*, volume 2, págs. 399–402, Mineápolis.

Zissman, M. (1995). Language identification using phoneme recognition and phonotactic language modeling. Em *Proc. Int. Conf. on Acoustic Speech and Signal Processing*, volume 5, págs. 3503–3506, Detroit.

Zue, V., Glass, J., Philips, M., Seneff, S. (1989). Acoustic segmentation and phonetic classification in the SUMMIT system. Em *Proc. Int. Conf. on Acoustic Speech and Signal Processing*, págs. 389–392, Glásgua.

Índice Remissivo