Reconhecimento Robusto de Fala: a Experiência do Projecto SUNSTAR

Carlos J. Teixeira, Isabel M. Trancoso, Carlos M. Ribeiro,
Carlos A. Martins, António J. Serralheiro, e Moisés S. Piedade

INESC

RESUMO

Este artigo descreve sumariamente o projecto SUNSTAR, nele enquadrando a participação nacional, a qual teve como tema o aumento da robustez de um sistema de reconhecimento de fala. Mais concretamente, pretendeu-se aumentar a independência do sistema face às condições ambientais e em relação à utilização de palavras estranhas ao vocabulário por parte de utilizadores não treinados, factores que são susceptíveis de degradar significativamente o desempenho de um sistema de reconhecimento.

1. INTRODUÇÃO

O projecto ESPRIT SUNSTAR (Integration and Design of Speech Understanding Interfaces), que chegou recentemente ao fim dos seus três anos de duração, reuniu oito empresas, universidades e institutos de investigação europeus, com o objectivo de construir um sistema de processamento de fala baseado numa arquitectura modular e flexível (figura 1), com controlo integrado do diálogo homem-máquina. O interface com o utilizador foi um dos factores cruciais deste projecto, tendo sido dado um ênfase especial ao desenho do diálogo e à avaliação do sistema a vários níveis.

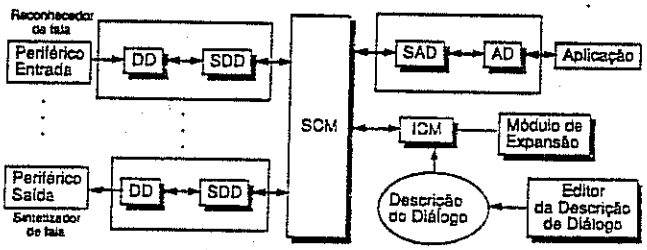

Figura 1: Diagrama de blocos da arquitectura SUNSTAR

Este desenho é feito numa estação de trabalho correndo o sistema operativo UNIX, com base numa ferramenta gráfica bastante potente que cobre várias fases desde a especificação do diálogo, passando pela implementação até à sua verificação e teste do sistema. A descrição do diálogo é feita a vários níveis (gráfico, de trama e textual). A esta representação, segue-se a compilação para uma linguagem de descrição de diálogos cujo resultado é usado pelo módulo de interpretação e controlo (ICM - Interpretation and Control Module), durante a execução. O ICM comunica directamente com o gestor de comunicações (SCM - SUNSTAR Communication Manager), o qual, por sua vez, centraliza a comunicação com os vários "device drivers" (DD) correspondentes por exemplo, ao reconhecedor, sintetizador, teclado, rato, écran, comutador, etc.. Um destes "drivers" é específico da aplicação pretendida (AD - Application Driver). A ligação dos "drivers" ao SCM é garantida pelos "SUNSTAR drivers" correspondentes (SDDs e SADs).

O progresso do projecto foi avaliado através da construção de demonstradores de aplicações de entrada e saída de fala em duas áreas representativas: o ambiente de escritório (que designaremos de profissional) e o ambiente da rede telefónica (público). Como exemplos dos primeiros, foram seleccionados dois demonstradores: uma ferramenta de CAD comandada por fala e uma agenda electrónica. No primeiro, cujo objectivo é a conversão de dados de CAD em informação sobre contornos, a fala é usada como meio de entrada de dados adicional, em combinação com o rato e o teclado, libertando as mãos do utilizador para outros fins. No segundo, pretende-se demonstrar as vantagens de utilizar fala como meio de entrada e saída adicional para processamento de documentos e endereços.

Como exemplo de aplicações no ambiente público, temos três outros demonstradores: serviço noticioso, marcação abreviada / despertar automático e controlo oral operações num PABX. O primeiro é uma aplicação típica de audiotexto que, em vez de ser comandada pelo teclado do telefone, é comandada oralmente através de palavras que permitem ao utilizador navegar num menu hierárquico, com estrutura de árvore. A cada passo, as opções disponíveis são faladas pelo sistema. Esta aplicação pressupõe a actualização diária das notícias. O segundo demonstrador tem como objectivo extender as facilidades típicas de despertar automático e marcação abreviada a assinantes com marcador decádico, mesmo quando não estão ligados a centrais digitais. Esta aplicação permite, através de comandos orais, verificar, activar e desactivar o despertar, e editar, verificar e utilizar a marcação abreviada. O terceiro demonstrador permite efectuar operações típicas de um PABX através de comandos vocais como, por exemplo, "ligue para", "transfira para", "repita a marcação", etc.

Na realização de todos estes demonstradores procurou-se utilizar tanto quanto possível módulos "state-of-the-art", tirando partido do facto de que a arquitectura adoptada permite a integração eficiente de novos dispositivos e técnicas de entrada / saída de fala. Esta integração possibilitou, durante a última parte do projecto, o desenvolvimento em paralelo de um reconhecedor mais avançado que, em vez de lidar com palavras isoladas como o dos demonstradores anteriores, é capaz de reconhecer fala contínua , de acordo com uma gramática de estados finitos. Para efeitos de demonstração das potencialidades deste reconhecedor, foi seleccionada uma aplicação artificial de CAD, caracterizada por um vocabulário de nomes de objectos geométricos que podem sofrer deslocações, mudanças de dimensão, cor, etc.. Embora o seu desenvolvimento em protótipo não tenha sido o objectivo deste projecto, o trabalho com este fim tem sido prosseguido, já fora do âmbito do projecto, pela universidade dinamarquesa que iniciou o seu desenvolvimento.

Com excepção desta aplicação, a qual foi demonstrada numa estação de trabalho, e da aplicação do PABX, a qual tem como base uma placa de processamento de sinal específica do fabricante do mesmo, todas as restantes aplicações são baseadas num computador pessoal correndo o sistema operativo VENIX, equipado com placas de processamento de sinal DSP32C da Loughborough Sound & Images. Numa destas placas está o software do reconhecedor de fala e os codecs para reprodução da informação falada pelo sistema, a qual deve ser previamente gravada. A outra placa está reservada para pré-processamento de sinal, incluindo também, no caso do ambiente público, o software de cancelamento de eco, necessário para que o utilizador possa interromper as mensagens do sistema através da própria fala.

O trabalho do INESC dentro do projecto teve como tema o reconhecimento robusto de fala. Pretendeu-se aumentar a independência do sistema face às condições ambientais e em relação à utilização de palavras estranhas ao vocabulário por parte de utilizadores não treinados. Neste artigo, descrever-se-ão sumariamente as várias facetas deste trabalho. Previamente, porém, introduzir-se-á o reconhecedor de base desenvolvido pelos parceiros dinamarqueses que foi adoptado para quase todas as aplicações (Secção 2). Na Secção 3, descreveremos as bases de dados de fala e ruído que serviram de base ao treino e teste do sistema e por cuja especificação, análise e verificação fomos em grande parte responsáveis. A Secção 4 trata do problema da robustez face ao ruído, particularizando para o ambiente público e profissional. A Secção 5 descreve o trabalho realizado na área de rejeição de palavras fora do vocabulário e identificação de palavras-chave (word rejection / spotting), sendo as conclusões finalmente apresentadas na Secção 6.

2. RECONHECIMENTO DE FALA

O reconhecedor de palavras isoladas adoptado nos quatro primeiros demonstradores é baseado em modelos de Markov contínuos não-observáveis (Continuous Hidden Markov Model - CHMM). O detector de extremos de palavra segue de perto o algoritmo baseado em limiares de energia e duração. Adoptou-se uma topologia de 10 estados para cada modelo de palavra, sem saltos sobre estados. As probabilidades de saída são descritas por uma única componente gaussiana, com matriz de covariância diagonal. O sinal de fala é amostrado a 8 kHz. Cada observação é calculada sobre uma janela de Hamming de 20 ms, com uma sobreposição de 50% entre janelas. Os vectores de obervação incluem os coeficientes de cepstro e delta-cepstro. O cepstro é baseado numa análise de predição linear de 8ª ordem e o delta-cepstro utiliza um atraso de 4 tramas. A fase de treino é inicializada por meio de um alinhamento de Viterbi. Aplica-se então iterativamente o algoritmo de Baum-Welsh, até que a variação percentual verosimilhança total seja inferior a um dado limiar ou até que se exceda um número máximo de iterações. Este reconhecedor foi compilado de C para o Assembly da placa de DSP32C, incluindo também este software rotinas para gravação e reprodução de mensagens (3R - Recognise, Record and Replay).

O reconhecedor mais avançado adoptado no último dos demonstradores é também baseado em CHMMs, pelo que o pacote de treino é comum a ambos. Dado que as unidades de fala a reconhecer são configuráveis pelo utilizador, o reconhecedor pode aceitar modelos de palavras ou de unidades sub-palavra. A busca é baseada no algoritmo de Viterbi com passagem de marca (token-passing), constrangido por regras sintácticas armazenadas numa tabela de estados finitos. A topologia adoptada para cada estado é idêntica à do reconhecedor anterior. O software de reconhecimento correspondente foi designado de SUNCAR (SUNSTAR Continuous Advanced Recogniser).

3. BASES DE DADOS DE FALA E RUÍDO

O facto do consórcio SUNSTAR incluir 4 parceiros industriais de países diferentes, cada um dos quais responsável pela construção de um (ou dois) demonstrador(es), levou à necessidade de recolher corpora de fala em línguas diferentes: alemão (para os demonstradores do ambiente profissional), espanhol (para o serviço noticioso), dinamarquês (para os serviços de despertar automático e marcação abreviada) e italiano (para o demonstrador do comando oral do PABX). Dado que cada uma das aplicações deveria também ser demonstrada em inglês, foi decidido utilizar oradores nativos e não-nativos para a gravação deste último corpus, de modo a obter uma base de dados representativa em termos de pronúncia, num ambiente multi-nacional. Assim, o vocabulário de cada uma das 5 aplicações foi gravado na própria língua por 100 oradores nativos e em inglês por um conjunto de 100 oradores constituído por 20 oradores de cada um dos 4 países e por 20 oradores ingleses.

Este corpus multi-lingue foi gravado de acordo com as recomendações do projecto europeu SAM, em condições ambientais muito favoráveis (câmara surda ou anecóica), utilizando a estação de trabalho SESAM (computador pessoal equipado com placa OROS) e o software EUROPEC, desenvolvidos neste último projecto. A frequência de amostragem recomendada foi de 20 kHz. Para o treino e teste do SUNCAR, desenvolvido pelos

parceiros dinamarqueses, foi apenas recolhida uma base de dados de âmbito mais restrito na língua respectiva.

As condições ambientais típicas dos dois ambientes de funcionamento são bastante diferentes. No ambiente profissional, são frequentes ruídos altamente não estacionários oriundos, por exemplo, dos teclados, da abertura e fecho de portas, janelas e gavetas, do folhear de documentos, das impressoras, etc. Há também outros ruídos mais estacionários provocados por ventoínhas de computadores e aparelhos de ar condicionado. No ambiente público, há que contar não só com o ruído ambiente (de casas, escritórios, lojas, restaurantes, estações, ruas e um sem número de locais onde possa estar instalado um telefone público ou privado), mas também com o ruído introduzido pelo canal telefónico não ideal. Foram excluídas à partida condições adversas do tipo interior de fábricas, carros, aviões, etc., as quais são objecto de estudo por parte de um outro projecto europeu.

A recolha de ruído teve lugar nos 4 países mencionados, conseguindo-se um total de 99 ficheiros de 1 minuto de duração cada. Destes, cerca de metade corresponde a ruído de linhas telefónicas (13 internas e 29 externas) e a outra metade a ruído ambiental (40 ficheiros de ruído gravado em escritórios e 17 de outros locais). O primeiro subconjunto foi amostrado a 8 kHz e o segundo a 20 kHz, também por razões de compatibilidade com as recomendações do projecto SAM.

Após a recolha dos vários ficheiros, efectuou-se uma análise de componentes principais, de modo a permitir seleccionar exemplos representativos da base recolhida. Esta análise teve como base uma sequência de 50 espectros de 1/3 de oitava de cada um dos ficheiros de ruído e conduziu à selecção de 66 ficheiros dos 99 originais (4 de linhas telefónicas internas, 17 de externas, 30 de ruído de ambiente de escritório e 15 de outros ambientes). Tanto estes como os outros menos representativos foram incluídos juntamente com os resultados da análise de componentes principais num CD-ROM designado de SUN-ROM 1 que constitui um dos produtos comercializados pelo consórcio.

4. ROBUSTEZ FACE A CONDIÇÕES AMBIENTAIS

Do ponto de vista de reconhecimento de fala, são vários os métodos normalmente adoptados para fazer face a condições ambientais não ideais: treino dos modelos com base em fala produzida nessas condições, pré-processamento do sinal de entrada de modo a retirar-lhe tanto quanto possível a componente de ruído, adopção de parâmetros e de medidas de distância entre parâmetros menos sensíveis à presença de ruído e modelamento integrado de fala e ruído. O primeiro e o último métodos foram excluídos à partida devido, respectivamente, à variabilidade das condições ambientais previstas e à reduzida complexidade exigida para implementação em tempo real. A comparação de parâmetros e medidas de distância do ponto de vista de robustez ao ruído foi da responsabilidade do parceiro italiano.

Cada um dos ambientes é caracterizado por condicionantes diferentes: no ambiente público, o pré-processador dispõe apenas do sinal captado pelo microfone do telefone, o qual, além de estar corrompido pelo próprio ruído ambiente, é posteriormente distorcido pelo canal telefónico; no ambiente profissional, torna-se viável a instalação de dois microfones (pelo menos): um destinado primordialmente a capturar o sinal de fala e outro à captura do ruído ambiente. Em termos de implementação em tempo real, as condicionantes diferiam também, dado que, no ambiente público, a placa de DSP32C disponível para pré-processamento é partilhada com o algoritmo de cancelamento de eco que ocupa o processador cerca de 50% do tempo.

No ambiente público, o módulo de pré-processamento é uma variante do método de subtracção espectral clássico (Boll, 1979). A ideia básica consiste em estimar a amplitude espectral do ruído durante segmentos que não contenham fala e subtrair esta estimativa ao sinal contaminado. Este método pressupõe a contaminação do sinal de fala ($\hat{s}_i$) com ruído aditivo do tipo estacionário (isto é, o sinal de entrada é dado por $x_i = s_i + n_i$). O espectro do sinal de saída (X_i')é obtido subtraindo da amplitude espectral do sinal de entrada (X_i) a estimativa da amplitude espectral do ruído (N_i') e mantendo a fase do sinal de entrada. O diagrama de blocos (figura 2) inclui os módulos de FFT directa e inversa, o detector fala / não-fala, o estimador da amplitude espectral do ruído, o subtractor propriamente dito e os módulos de aplicação de janelas de análise e sobreposição de janelas de síntese (50%).

No ambiente profissional, o pré-processamento é feito através de um cancelador de ruído acústico (ANC - Acoustic Noise Canceller) que recebe o sinal através de dois microfones, primário e de referência. A ideia básica consiste em filtrar adaptativamente o sinal de referência para obter uma estimativa do ruído, e subtraí-lo do sinal do microfone primário de modo a conseguir um sinal de fala menos contaminado. A adaptação dos coeficientes do filtro é feita através do algoritmo LMS normalizado, modificado de modo a conseguir uma elevada rapidez de convergência compatível com um bom nível de cancelamento do ruído (Martins, 1990). Este pré-processador pressupõe também a quasi-estacionaridade do ruído além duma boa correlação do mesmo nas duas entradas. A figura 3 mostra o diagrama de blocos deste cancelador.

Foram realizados três tipos de testes para avaliação do desempenho dos pré-processadores. Os primeiros testes foram do tipo auditivo, e envolveram um número muito pequeno de ouvintes, sendo portanto pouco significativos. Os segundos consistiram em medidas comparativas de relações sinal-ruído (SNR - signal-to-noise ratio) dos sinais contaminados com e sem pré-processamento. Os terceiros testes consistiram na comparação dos resultados de reconhecimento obtidos com os mesmos sinais e são obviamente os mais relevantes para o fim em questão. A base de dados de fala utilizada para teste é um subconjunto do corpus gravado em inglês (40 palavras correspondentes ao vocabulário de uma das aplicações, faladas por 5 oradores masculinos e 5 femininos, fora do universo utilizado para treino do reconhecedor).

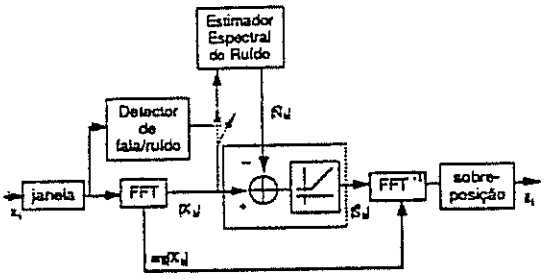

Figura 2: Diagrama de blocos da subtracção espectral.

Para os testes no ambiente público, cada um destes ficheiros de fala foi artificialmente contaminado com ficheiros representativos seleccionados do SUNROM 1: 6 do tipo ruído de linha telefónica e 6 do tipo ruído ambiental. Foram adoptados dois níveis de contaminação (SNR de entrada = 6 e 12 dB). Para os ruídos do tipo mais estacionário (i.e., telefónico), obteve-se, qualitativamente, um sinal pré-processado mais claro, à custa de uma certa distorção tonal; em termos de SNR, conseguiu-se uma melhoria média de cerca de 8 dB; finalmente, em termos de resultados de reconhecimento, obteve-se uma melhoria de cerca de 20 pontos percentuais (nos casos em que o ruído é desprezável, o módulo de pré-processamento é curto-circuitado). Para os ruídos de tipo menos estacionário (tipicamente, os de um ambiente de escritório), a violação das hipóteses subjacentes à subtracção espectral conduz a um desempenho medíocre, isto é, o pré-processador não reduz a contaminação do sinal de entrada.

No que diz respeito ao ambiente profissional, foram adoptados níveis de contaminação variando dos 4 aos 18 dB. Para medir o efeito do cancelador em termos de SNR, os sinais de fala foram artificalmente contaminados por ruídos do tipo laboratorial (onda sinusoidal ou quadrada), tendo-se obtido cerca de 17 dB de melhoria para uma contaminação correspondente a 11 dB de SNR. Para os testes auditivos e, principalmente, para os testes com o reconhecedor, utilizaram-se, para além destes ruídos do tipo laboratorial, ruídos reais (aparelho de ar condicionado, teclado, impressora de agulhas, ruído de fundo produzido por um grupo de pessoas a falar simultaneamente). Embora os resultados obtidos sejam bons para ruídos laboratoriais (cerca de 30% de melhoria em reconhecimento), o mesmo não foi conseguido para os ruídos reais menos estacionários (melhorias pouco significativas) onde as hipóteses em que o cancelador assenta são obviamente violadas. Estes resultados demonstram mais uma vez as limitações de métodos do tipo pré-processamento na redução de ruído.

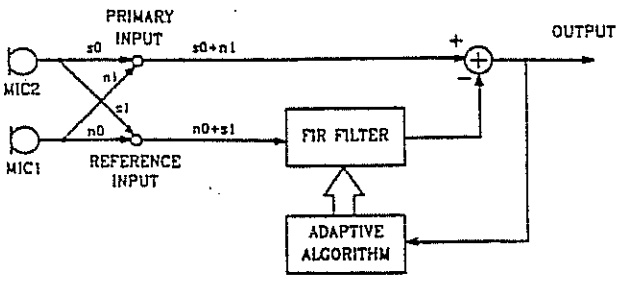

Figura 3: Diagrama de blocos do cancelador de ruído.

5. REJEIÇÃO / DETECÇÃO DE PALAVRAS

A utilização de palavras estranhas ao vocabulário para o qual o reconhecedor é treinado é o segundo factor relativamente ao qual se tentou aumentar a robustez do sistema. Esta utilização é relativamente frequente. Numa aplicação que tenha, por exemplo, a palavra-chave *chamada*, é de prever sequências de entrada do tipo: "gostaria de fazer uma *chamada*, por favor". De facto, se o sistema não for capaz de identificar essas palavras, ocorrerão erros de substituição que comprometem significativamente o seu desempenho.

5.1. Método

O método mais conhecido para impedir que palavras não pertencentes ao vocabulário da aplicação sejam reconhecidas incorrectamente, consiste na utilização de modelos adicionais, normalmente designados por *garbage*, *sink* ou *filler models* - modelos de escoamento (ME). Estes modelos, estruturalmente idênticos aos modelos de palavras-chave (MP) diferem destes quer na forma como são criados, quer na sua utilização. Enquanto cada modelo de palavras-chave é treinado exclusivamente com diversas fonações da palavra correspondente, um modelo de escoamento deverá ser treinado com fonações de diversas palavras. Na impossibilidade de se obterem e processarem modelos para cada palavra possível na fala, o objectivo deste procedimento é o da obtenção de um número reduzido de modelos que permitam representar de forma indiscriminada todas as palavras que não pertençam ao vocabulário da aplicação.

Os primeiros trabalhos surgidos neste contexto utilizaram um único ME com sucesso e pareciam indicar a inutilidade da utilização de MEs adicionais que só demorariam o processo de reconhecimento. Contudo, quando se utilizam modelos de observações contínuas modeladas por uma única componente ou se pretendem contemplar fonações de grande variabilidade, pode questionar-se a utilidade dos MEs múltiplos. Um dos problemas da utilização de múltiplos MEs é a divisão do corpus de fala para o treino de cada um dos modelos. A utilização de técnicas de agrupamento tradicionais tais como o método das k-médias para o treino de HMMs não revelaram melhorias significativas em relação a outras divisões mais ou menos arbitrárias. Como tal, utilizou-se como critério a ordenação alfabética das palavras disponíveis para o treino, colocando a palavra m no conjunto de treino $m \bmod n$ em que n é o número de conjuntos de treino (= número de MEs) a obter. Na figura 4 é possível verificar os melhoramentos conseguidos na taxa de rejeição com a adição de múltiplos MEs. Estes resultados foram obtidos para o demonstrador em inglês do despertar automático / marcação abreviada (40 palavras-chave faladas por oradores nativos). O material usado para o treino dos MEs correspondeu ao vocabulário de um dos demonstradores do ambiente profissional falado pelos mesmos oradores. Para o teste utilizaram-se oradores nativos diferentes e o vocabulário do outro demonstrador.

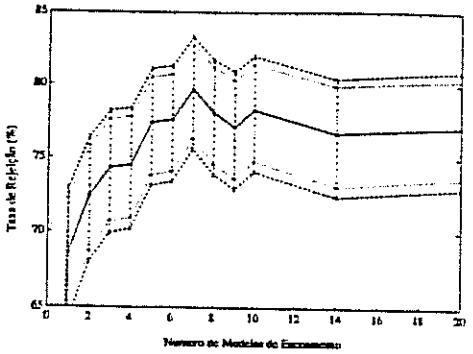

Figura 4: Taxa de rejeição versus número de modelos de escoamento. Os resultados experimentais encontram-se assinalados por pequenos círculos e as linhas tracejadas e ponteadas representam respectivamente os intervalos de confiança a 95 e a 90 %.

5.2. Efeito da pronúncia não nativa do inglês

Pretendendo construir sistemas de reconhecimento para operarem num contexto pan-europeu (em inglês, por exemplo), há que estudar os efeitos da pronúncia de uma dada língua por oradores não nativos. Para tal, dispõe-se, na base de dados do SUNSTAR, de vários sotaques do inglês (nativo e não nativos). De entre estes, seleccionaram-se os subconjuntos pronunciados respectivamente por ingleses e por dinamarqueses, e utilizou-se um vocabulário para treino e teste de MPs e MEs idêntico ao descrito em 5.1.

Realizaram-se experiências com 1, 2 e 5 MEs. Os resultados obtidos estão resumidos nos quadros seguintes, nos quais I representa um conjunto de treino/teste falado por ingleses e D o conjunto falado por dinamarqueses. Na Tabela 1, as duas primeiras colunas de resultados apresentam as taxas de reconhecimento de palavras-chave, sem MEs. É notória a quebra de desempenho devida à utilização de diferentes pronúncias no treino e no teste relativamente à utilização da mesma pronúncia em ambos. É também assinalável a assimetria entre o par *treino = I, teste = D* e o par *treino = D, teste = I*, a qual poderá ser explicada por uma maior uniformidade de pronúncias entre os falantes nativos. O par de colunas seguinte apresenta as taxas de rejeição obtidas com este conjunto de MPs acrescido de um ME isolado e o último par apresenta resultados correspondentes para 5 MEs. Estes últimos são apreciavelmente superiores aos obtidos com um único ME.

Tabela 1

Treino/Teste	reconhec.		rej. c/1 ME		rej. c/5 MEs	
	I	D	I	D	I	D
I	96.9	79.7	62.3	59.0	68.4	69.8
D	86.9	96.9	71.5	64.9	76.4	71.0

Na Tabela 2, repetem-se algumas das experiências anteriores mas tendo sido os MP treinados simultâneamente com ingleses e dinamarqueses. Verifica-se que as taxas de reconhecimento atingem agora valores comparáveis aos obtidos nos conjuntos treino/teste com a mesma pronúncia. Em contrapartida, as taxas de rejeição desceram significativamente e podem agora ser comentadas da mesma forma que para as taxas de reconhecimento na Tabela 1.

Tabela 2

Treino/Teste	reconhec.		rej. c/ 1 ME	
(MEs)	I	D	I	D
I	96.9	95.9	52.8	23.4
D			29.3	57.1

Similarmente ao que foi feito com MPs, treinaram-se MEs utilizando simultâneamente as pronúncias inglesas e dinamarquesas. O resultado obtido com um único ME (Tabela 3) foi substancialmente melhor do que os que foram obtidos com diferentes pronúncias dos falantes de treino e teste. É todavia inferior ao obtido com idênticas pronúncias no treino e no teste. Utilizando MEs múltiplos, consegue-se contudo melhorar a rejeição até valores consideravelmente superiores aos obtidos na Tabela 2.

Tabela 3

Teste/Nº MEs	1	2	5	10
I	48.6	54.3	59.7	64.6
D	47.9	55.2	62.5	67.9

Os resultados obtidos sublinham a importância de incluir todas as pronúncias no conjunto de treino no contexto particular de ambientes multi-nacionais, quer para os MEs quer para os MPs convencionais. Mais relevante ainda é a constatação das vantagens da utilização de MEs múltiplos nestas mesmas situações.

5.3. Aplicações em fala ligada

Designe-se por gram0 a gramática de estados finitos empregue no reconhecedor de palavras ligadas. Esta gramática inclui arcos associados a modelos de silêncio (ou ruído) nos estados inicial e final. A gram0 foram sucessivamente adicionados arcos extra associados aos MEs e aos modelos de silêncio, criando-se as gramáticas designadas de gram1, gram2 e gram3. Em gram1, estes arcos foram apenas adicionados no início e no final de cada frase, o que corresponde a prever palavras estranhas ao vocabulário apenas nos extremos da frase. Em gram2, a introdução de arcos foi feita na transição entre sub-gramáticas (por exemplo, entre o sujeito e o verbo, mas não entre o artigo e o objecto). Finalmente, na gram3 incluíram-se estes arcos adicionais entre todas as palavras da frase.

Criaram-se 3 MEs: um treinado exclusivamente com palavras isoladas, *pi*, outro treinado exclusivamente com frases completas, *fc*, e o terceiro utilizando as palavras e frases empregues no treino de *pi* e *fc* para treinar um único modelo, *pf*. O primeiro conjunto de experiências utilizou apenas frases de teste correctas de acordo com gram0. Os resultados obtidos (Tabela 4) mostram uma degradação no desempenho devido à introdução de MEs, particularmente quando esta é efectuada no meio da frase (gram2 e gram3).

Tabela 4

Gramática	Frases correctas	Palavras correctas
gram0	75.9	87.8
gram1	73.7	84.3
gram2	69.4	79.9
gram3	69.1	79.9

As frases de teste utilizadas no segundo conjunto de experiências incluem palavras adicionais no meio de frases correctas (de acordo com gram0). Na tabela 5, apresentam-se os resultados obtidos com a gram0 e utilizando os MEs *pi*, *fc*, *pf* e *pi+fc*, nas gramáticas gram1, gram2 e gram3. Nestes resultados, é patente a superioridade das gramáticas mais versáteis nomeadamente a gram3. O desempenho inferior da gram1 só pode ser explicado pela maior ocorrência das palavras adicionais no meio das frases de teste do que nos extremos destas frases. Não se detectou qualquer diferença entre o desempenho dos vários MEs quer isolados quer múltiplos (*pi+fc*). Este último facto põe em causa a vantagem da utilização de MEs múltiplos apontada pelos resultados apresentados na Figura 4.

Tabela 5

Gramát.	ME	Frases correctas	Palavras correctas
gram0	-	37.5	65.1
gram1	pi	26.9	56.5
gram2	pi	51.9	69.8
gram3	pi	54.4	74.8
gram1	fc	30.0	58.9
gram2	fc	51.2	69.8
gram3	fc	55.6	76.6
gram1	pf	28.8	57.9
gram2	pf	54.4	70.8
gram3	pf	56.2	75.8
gram1	pi+fc	28.1	55.8
gram2	pi+fc	51.9	69.6
gram3	pi+fc	55.0	76.0

Um factor que se revelou de maior importância é o da dimensão do vocabulário activo em cada instante de decisão do reconhecedor. De facto, o vocabulário empregue nos testes subjacentes à figura 4 era de 40 palavras, enquanto que o factor de ramificação médio da gramática gram0 é de 3.8 (com 3.3 de desvio padrão), a perplexidade é de 8.5 e a perplexidade do teste é de 6.64. A figura 5 mostra os resultados de um outro conjunto de experiências com o reconhecedor de palavras isoladas utilizando vocabulários com vários tamanhos e um número de MEs variável. A vantagem da utilização de MEs múltiplos é notória para vocabulários com mais de 10 palavras. Para vocabulários de dimensão inferior, essa vantagem diminui claramente com o número de palavras até que, com 5 palavras, não é relevante a melhoria obtida. Estes resultados confirmam assim os resultados obtidos na tabela 5 com MEs múltiplos.

Comparando agora os resultados da tabelas 4 e 5, verifica-se que os resultados obtidos na última com a gram3 estão longe de se igualarem aos obtidos na primeira com gram0 (sem utilizar MEs). Um cálculo simples permite determinar que, para que a utilização de MEs seja vantajosa, basta que mais de 30% das frases de teste incluam palavras estranhas à gramática pré-estabelecida. Em geral, é de esperar percentagens superiores num conjunto de teste que se pretenda próximo da realidade, facto que justifica o interesse da utilização de MEs no contexto da fala ligada.

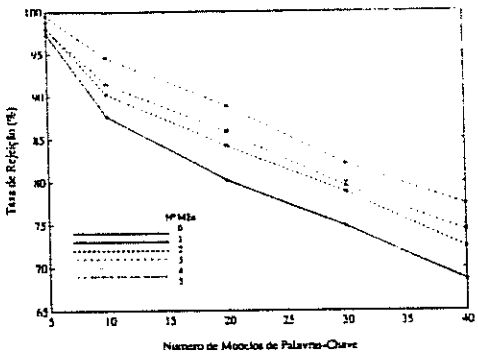

Figura 5: Taxa de rejeição em função do número de palavras chave para vários modelos de escoamento (0 a 5).

6. CONCLUSÕES

Como conclusão do nosso envolvimento no projecto SUNSTAR, há que apontar a experiência ganha na área do reconhecimento robusto ao ruído e, em particular, na área de reconhecimento de fala ligada, não esquecendo também a implementação em tempo real de algoritmos de processamento de fala.

O projecto culminou como uma série de ensaios de campo efectuados pelos parceiros industriais dos vários demonstradores do ambiente público. Embora não se tenha ainda atingido os níveis de performance ideais em termos de palavras correctamente reconhecidas, o sistema foi considerado aceitavelmente rapido, de fácil aprendizagem e bastante amigável para o utilizador.

REFERÊNCIAS

S. Boll, "Suppression of Acoustic Noise in Speech Using Spectral Subtraction", IEEE Trans. Acoustic, Speech and Signal Proc., Vol. ASSP-27, págs. 113-120, 1979..

B. Lindberg, B. Andersen, A. Bækgaard, T. Brøndsted, P. Daalsgaard, J. Kristensen, "An integrated Dialogue Design and continuous speech recognition environment", 1992 International Conference on Spoken Language Processing, ICSLP'92, Banff, Canada, págs. 1447-1450, Outubro de 1992. .

Carlos R. Martins, Teresa M. Almeida e Moisés S. Piedade, "An Adaptive Noise Canceller and Its Implementation on a DSP", Proceedings of the International Symposium on Circuits and Systems, New Orleans, págs. 1939 - 1942, Maio de 1990.

T. Renner, "Dialogue Design and System Architecture for Voice-Controlled Telecommunication Applications", IEEE Workshop on Interactive Voice Technology for Telecommunications Applications, Piscataway, NJ, Outubro de 1992.

C. Teixeira, I. Trancoso e A. Serralheiro, "Single vs. Multiple Sink Models for Isolated and Connected Word Rejection", ESCA Workshop on Speech Processing in Adverse Conditions, Cannes-Mandelieu, France, págs. 179-182, Novembro de 1992.