Reconhecimento de Dígitos e Números Naturais

Frederico Rodrigues e Isabel Trancoso

INESC, Instituto de Engenharia de Sistemas e Computadores

IST, Instituto Superior Técnico

INESC, Rua Alves Redol, Nº 9, 1000 Lisboa

E-mail: {fspr, Isabel.Trancoso@inesc.pt}

Resumo

Este artigo descreve experiências de reconhecimento de dígitos e números naturais para o Português Europeu utilizando o *corpus* SPEECHDAT. É caracterizado o problema do reconhecimento de fala, em particular em condições adversas, e é analisada a influência destas condições no sinal de fala. Apresentamos resultados de um sistema de base, para reconhecimento de dígitos isolados e ligados, caracterizado por modelos de palavra e que serviu para aferir a qualidade do referido *corpus*. Posteriormente, são definidas extensões que se centraram em grande parte numa arquitectura alternativa para os modelos de ruído utilizados. Foi conseguido um decréscimo da ordem dos 40% na taxa de erro para a tarefa de reconhecimento de dígitos ligados e melhorias significativas para os dígitos ligados com as extensões ao sistema de base. No que diz respeito aos números naturais, a taxa de erro, de cerca de 5%, é francamente encorajadora na medida em que ainda são possíveis diversas optimizações.

Reconhecimento de Dígitos e Números Naturais

Resumo

Este artigo descreve experiências de reconhecimento de dígitos e números naturais para o Português Europeu utilizando o *corpus* SPEECHDAT. É caracterizado o problema do reconhecimento de fala, em particular em condições adversas, e é analisada a influência destas condições no sinal de fala. Apresentamos resultados de um sistema de base, para reconhecimento de dígitos isolados e ligados, caracterizado por modelos de palavra e que serviu para aferir a qualidade do referido *corpus*. Posteriormente, são definidas extensões que se centraram em grande parte numa arquitectura alternativa para os modelos de ruído utilizados. Foi conseguido um decréscimo da ordem dos 40% na taxa de erro para a tarefa de reconhecimento de dígitos ligados e melhorias significativas para os dígitos ligados com as extensões ao sistema de base. No que diz respeito aos números naturais, a taxa de erro, de cerca de 5%, é francamente encorajadora na medida em que ainda são possíveis diversas optimizações.

1 Introdução

Durante a última década, a pesquisa na área do reconhecimento automático da fala tem vindo a produzir resultados positivos e cada vez mais significativos. Multiplicam-se as tarefas e as línguas em que os níveis de desempenho são elevados.

Em muitas aplicações de reconhecimento de fala, nomeadamente nas que envolvem a rede telefónica, fixa ou móvel, o reconhecimento de dígitos e números naturais terá um papel importante. Muitas empresas (banca, seguros, energia, telecomunicações) oferecem serviços de apoio ao cliente baseados na comunicação através do teclado telefónico (multi-frequência ou decádico). Estes interfaces têm limitações bem conhecidas que tornam a comunicação pouco atractiva e natural. A migração destes interfaces para outros mais evoluídos em que a comunicação com a máquina se baseie essencialmente em linguagem falada, terá como consequência imediata uma maior aceitação por parte dos utilizadores.

Estudos efectuados mostram que, neste tipo de sistemas, o reconhecimento se centra em grande parte nas tarefas de dígitos e números naturais [Johnston97]. Contudo, a aplicabilidade com sucesso desta tecnologia depende de elevados níveis de desempenho.

1.1 Caracterização do Problema

É claramente reconhecido que um meio privilegiado para o interface Homem-máquina é, pela sua universalidade, o canal telefónico. Grande parte dos problemas levantados neste contexto estão directamente ligados a características intrínsecas deste canal de comunicação e do próprio micro-telefone, acrescidos, como veremos de

seguida, por outros problemas que se prendem por exemplo com diferenças naturais entre oradores, com a existência de dialectos e pronúncias diversos, etc.

Existem vários tipos de ruído que podem intervir a diferentes níveis do processo de produção/reconhecimento da fala. Admitindo um esquema usual, temos presentes, para além do orador e do sistema de reconhecimento, um microfone (ou um telefone), um canal de ligação (ou linha telefónica) do microfone ao sistema de reconhecimento e o ambiente que circunda o utilizador. Todos estes elementos introduzem no sinal uma forma qualquer de ruído ou de distorção. Em ambientes laboratoriais, estes aspectos são minimizados recorrendo a microfones e canais de alta qualidade e a ambientes relativamente silenciosos. Este procedimento não se coaduna com as aplicações reais e tem-se vindo a verificar que, mesmo em sistemas em que as condições de treino tentam aproximar as condições reais ou de teste, existem défices significativos de desempenho.

É comum admitir-se que algum tipo de ruído pode ser modelado como uma perturbação estacionária adicionada ao sinal e não correlacionada com este. Nesta categoria de ruído inclui-se o ruído de fundo de carácter não intermitente, com uma estrutura diversa e que passa por exemplo por ruídos de um automóvel (motor, ventos, pneus, estrada), ruídos de fundo de locais públicos, música, conversações humanas (*cocktail party effect*), etc. No entanto, existe uma grande variedade de eventos não linguísticos, e de carácter não estacionário, que podem ocorrer durante o reconhecimento e que não podem ser modelados de forma tão simplificada. Entre estes incluem-se ruídos intermitentes como o toque de um telefone, o bater ou a campainha de uma porta, mas também ruídos e pausas preenchidas pelo orador. Foram desenvolvidas técnicas ([Wilpon90]) que consistem em criar modelos de Markov não observáveis (HMM – Hidden Markov Models) para estas classes de

eventos que são reconhecidos da forma tradicional e que são ignorados em fases subsequentes do reconhecimento. Para a implementação destas técnicas são necessárias bases de dados com anotação destes eventos. O sinal de fala sofre ainda várias distorções que podem afectar a sua estrutura de forma não linear. Podem ser, por exemplo, distorções devidas às propriedades acústicas da sala, e.g. reverberação, ao tipo de microfone e sua localização, ou ao próprio canal de transmissão.

Por outro lado, existe uma grande variabilidade entre oradores devido a características anatómicas associadas, por exemplo, à dimensão do tracto vocal ou a especificidades das cordas vocais (e.g. comprimento). Essas diferenças têm um peso significativo nas taxas de reconhecimento e não é por acaso que se torna já comum o uso de modelos separados em função do género. Contudo, as alterações na produção de fala não se limitam apenas a características físicas do orador, passando também pelo débito silábico e pelo estilo do orador. O estilo pode depender de características próprias deste ou do ambiente em que se encontra, que pode induzir alterações de comportamento como em casos de *stress* ou o chamado efeito de *Lombard*.

O *stress* induzido pelo ambiente pode ser devido a ruído, factores mecânicos como aceleração e vibração, emoções (e.g. medo), ou quaisquer outras formas de agentes físicos como o calor ou a pressão [Junqua95]. O efeito de *Lombard* caracteriza-se pela alteração do mecanismo de produção de fala do orador, e é difícil de modelar dado que, para além de exibir uma grande variabilidade de orador para orador, muitos dos factores que o induzem ainda são desconhecidos. Em [Junqua95], o autor enumera as principais diferenças registadas a nível acústico em função da presença ou não deste efeito, que passam pela alteração da localização das formantes, aumento da duração das vogais e da amplitude, entre outras.

1.2 Objectivos

O trabalho descrito é pois motivado pela necessidade que existe em ultrapassar os problemas referidos anteriormente por forma a obter um sistema com um grau de fiabilidade aceitável. Não é possível, contudo, tentar resolvê-los a todos os níveis do processo, dada a quantidade de variáveis em causa. Nesse sentido, optámos por fixar alguns dos parâmetros, para poder experimentar e comparar técnicas em etapas subsequentes do reconhecimento.

Os objectivos que pretendemos alcançar foram, numa primeira fase e com um sistema de base, aferir a qualidade dos dados de que dispúnhamos para o treino do sistema. Numa segunda fase, procurámos ajustar o sistema à realidade da fala telefónica e a um aumento progressivo da complexidade da tarefa. Sempre presente esteve a intenção de que o sistema fosse perfeitamente funcional, isto é, para além de apresentar taxas de reconhecimento elevadas, não comprometesse a aceitabilidade em termos de tempo de resposta, aspecto crucial neste tipo de aplicações.

Este artigo está organizado em 4 partes. Na próxima secção é descrito o *corpus* utilizado no treino do sistema desenvolvido e nas experiências realizadas. Na secção 3 são descritos o sistema de base, respectivas extensões e resultados. Finalmente, apresentam-se as conclusões e são sugeridos desenvolvimentos futuros.

2 Corpus

O *corpus* foi constituído a partir de dois esforços de recolha distintos, ambos no âmbito de projectos europeus SPEECHDAT[1]. Numa primeira fase, que designaremos

[1] http://speechdat.phonetik.uni-muenchen.de/SpeechDat.html

de SPEECHDAT(M), foram recolhidas 1000 chamadas telefónicas e numa segunda fase, que designaremos de SPEECHDAT II, foram recolhidas mais 4000 chamadas.

2.1 Características do *Corpus* SPEECHDAT

O *corpus* SPEECHDAT é um *corpus* multilingue, recolhido via linha telefónica, que inclui o Português europeu e outras línguas europeias. Para além da recolha da fala propriamente dita, num total de 5000 chamadas de oradores diferentes, todas as locuções foram ortograficamente transcritas, foram anotados eventos não linguísticos e foram criados léxicos de pronúncia. Os falantes foram recrutados de forma a garantir uma boa representação de todas as pronúncias regionais, idade e sexo.

A definição do conteúdo de cada chamada sofreu alterações da 1ª para a 2ª fase, mas, de uma forma geral, cada chamada era constituída por 33 itens lidos e 7 itens espontâneos. Os itens espontâneos serviram dois propósitos: o primeiro e essencial – obter os dados de identificação do cliente (nome, telefone, idade, sexo, cidade onde passou a maior parte da infância); o segundo – obter fala espontânea. As questões foram desenhadas de forma a maximizar a utilidade da fala obtida. Por exemplo, em vez da idade foi pedida a data de nascimento que permite obter, para além da idade, uma colecção de datas ditas de forma espontânea, importantes para o treino de um sistema de reconhecimento de datas. Os itens lidos incluem dígitos isolados, dígitos ligados (números de telefone, códigos de identificação pessoal, números de cartão de crédito), números naturais, quantias monetárias, datas, horas, frases e palavras foneticamente ricas, palavras de comando, palavras soletradas e alguns itens direccionados para o treino de serviços informativos telefónicos.

Como foi referido, todas as locuções foram ortograficamente transcritas e foram marcados os eventos não linguísticos embora de forma diferente nas 1ª e 2ª fases do

projecto. Enquanto que na 1ª fase foi feita uma anotação muito detalhada destes eventos, na 2ª fase existiam apenas 4 categorias, nomeadamente:

- [sta]: Ruído estacionário. Por exemplo, ruído de automóvel, ruído de estrada, ruído de canal, ruído de fundo de locais públicos, etc.
- [spk]: Todos os tipos de ruídos produzidos pelo orador e não pertencentes ao texto a ler, e.g., tosse, ruídos de lábios ou garganta, respiração, sopros, risos, etc.
- [int]: Ruídos de natureza intermitente, nomeadamente, o bater de uma porta, o toque de um telefone, vozes, música, campainhas, etc.
- [fil]: Pausas preenchidas pelo orador, "eh", "ah", "mm", etc.

2.2 Subconjunto Utilizado

Os elementos que constituem o subconjunto do SPEECHDAT utilizado para as experiências realizadas no contexto deste artigo são:

- Dígitos isolados – Locuções com apenas 1 dígito;
- Dígitos ligados - Locuções com sequências de 10 dígitos;
- Números naturais – Locuções constituídas por dígitos (unidades), números de 11 a 19, múltiplos de 10, múltiplos de 100, as palavras cento, mil, milhão, milhões, bilião, biliões e a palavra de ligação e;

Foram definidos, pelos parceiros do projecto, um conjunto de treino e correspondente conjunto de teste de 500 falantes para o SPEECHDAT II de modo a manter semelhante, em ambos os conjuntos, a distribuição dos falantes no que diz respeito à idade, região e género. Aplicámos o mesmo procedimento ao conjunto de treino remanescente de modo a dispormos de um conjunto de desenvolvimento com 300 falantes. Finalmente, de modo a podermos avaliar a 1ª fase de recolha, aplicámos ainda o procedimento a este conjunto, de onde seleccionámos 200 falantes para teste. Com este mecanismo de selecção atingimos um rácio global de 80% para treino e

20% para teste e desenvolvimento (tabela 2). Para os dígitos isolados, optámos por não utilizar o conjunto de desenvolvimento como tal dado que o número de dígitos no conjunto de teste era reduzido para aferir com rigor o desempenho.

	Treino	Teste	Desenvolvimento
I1	2954	768	-
B1	2905	491	277
N*	5059	467	260

Tabela 1 – Dimensão dos conjuntos associados a cada item.

Nem todos os ficheiros foram utilizados, devido a problemas de inteligibilidade ou à presença de palavras estranhas ao vocabulário.

3 Experiências realizadas e resultados obtidos

O sinal de fala foi codificado recorrendo a MFCC's (*Mel-Frequency Cepstral Coefficients*). Foram extraídos vectores com 30 parâmetros: 14 coeficientes cepstra, 14 delta-cepstra, um valor de energia e um de delta-energia. A banda do sinal foi limitada entre os 200 e os 3800 Hz e foi usada uma janela de *Hamming* de 25 ms cada 10 ms. Procedeu-se ainda à subtracção da média cepstral, um método simples mas eficiente de normalização do canal e do falante.

3.1 Sistema Base

Para modelar as unidades fonéticas foram usados HMM's contínuos, com estrutura esquerda-direita, sem saltos, no caso dos modelos de dígitos e com um salto do primeiro para o último estado emissor (e vice-versa) no caso dos modelos de ruído e de silêncio, como se pode observar na figura seguinte.

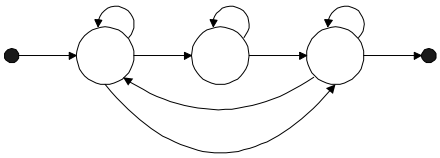

Figura 1 – Topologia dos modelos de ruído e de silêncio

Foram usados modelos de palavra e na tabela 3 é apresentada a atribuição do número de estados a cada dígito. Não são contabilizados o primeiro e o último estado que não emitem probabilidades. Cada modelo foi duplicado por forma a dispormos de modelos dependentes do sexo do orador.

Nº de Estados	Modelos
3	um, ruídos, silêncio
6	cinco, zero, nove, quatro, oito, três
7	Sete
8	dois, seis

Tabela 2 – Número de estados para cada modelo de palavra

O treino dos modelos acústicos iniciais foi efectuado tendo por base apenas o material de treino relativo aos dígitos isolados da 1ª fase do projecto. De entre eles, escolheram-se os que não tinham marcas de ruído (cerca de 1400) e reestimaram-se os parâmetros com recurso ao algoritmo de Baum-Welch embebido. A cada 2 ou 3 passos deste processo iterativo, o desempenho foi avaliado com o algoritmo de Viterbi com uma gramática de comprimento conhecido em que um dígito pode ser precedido ou seguido de qualquer número de ruídos ou silêncio. Quando não se verificaram alterações significativas nos resultados foi aumentado o número de misturas gaussianas por estado de uma para duas, e de duas para três numa iteração subsequente. O passo seguinte consistiu em introduzir os restantes ficheiros da 1ª fase, que continham marcas de ruído detalhadas, e aplicar de novo o processo iterativo

descrito anteriormente. Foram aumentadas as misturas gaussianas por estado de uma para três para os modelos de ruído e de silêncio.

Os restantes ficheiros do SPEECHDAT II tinham, como foi referido, menos marcas de ruído do que as existentes no SPEECHDAT(M). Foi pois necessário realizar um alinhamento do conjunto de treino do SPEECHDAT II de modo a poder mapear as marcas deste conjunto nos 9 modelos treinados com as marcas do SPEECHDAT(M). Relativamente ao treino e após este mapeamento, repetiu-se o processo iterativo tendo-se estabilizado num valor de 99,6% de precisão (*accuracy*) para o conjunto de treino.

Relativamente aos dígitos ligados, utilizámos como modelos iniciais os melhores modelos de dígitos isolados obtidos. Foi repetido o processo iterativo e foi aumentado gradualmente o número de misturas, para os modelos de dígitos, de 3 para 5. A razão pela qual foi possível e frutífero este aumento de misturas prende-se com o aumento do material de treino disponível. Os melhores resultados são apresentados na tabela seguinte. Como seria de esperar, os resultados relativos aos dígitos ligados são substancialmente inferiores, em particular quando o comprimento da locução, i.e., o número de dígitos presentes não é conhecido.

	% Correcção	% Precisão
Dígitos Isolados	99,0	99,0
Dígitos Ligados (Gramática de Comprimento Fixo)	97,8	97,6
Dígitos Ligados (Gramática de Comprimento Variável)	97,2	95,1

Tabela 3 – Resultados de Reconhecimento do Sistema Base

3.2 Extensão ao Sistema Base

A extensão ao sistema base consiste essencialmente em encontrar uma nova forma de modelar o ruído e o silêncio. É largamente aceite [Rahim95], [Perdue97], que o desempenho de um reconhecedor de dígitos ligados se degrada significativamente quando em presença de ruído ou quando o comprimento da locução é desconhecido. A razão principal tem a ver com o facto deste tipo de locuções não obedecerem a uma estrutura gramatical definida, i.e., qualquer dígito pode aparecer por qualquer ordem e rodeado por um número arbitrário de eventos não linguísticos. Por outro lado, é muito difícil anotar explicitamente cada um destes eventos. No processo de anotação do *corpus* SPEECHDAT, alguns destes eventos, como o ruído estacionário ou ruído intermitente, foram anotados apenas uma vez na transcrição, quer no início da locução, caso afectassem a locução inteira, quer antes da primeira palavra que afectassem. A abordagem seguida no sistema base para a resolução deste problema consistiu em replicar esta marca de anotação (só para ruídos estacionários) entre cada dígito. Contudo, esta não é uma solução óptima, na medida em que, na tarefa de dígitos ligados, pode haver pequenas ou nenhumas pausas entre os dígitos e conduzir assim a um treino deficiente dos modelos de dígitos que são a nossa principal preocupação.

Nesta extensão efectuámos novamente os mesmo passos que no sistema base até treinarmos os 9 modelos detalhados de ruído e o modelo de silêncio. Utilizámos estes modelos para construir um único modelo de ruído/silêncio que consiste na colocação dos 10 modelos existentes em paralelo. A figura seguinte mostra um diagrama de blocos do modelo, onde cada caixa corresponde a um modelo de ruído com a topologia anteriormente referida. A transição do fim para o início do modelo é apenas

uma simplificação gráfica uma vez que, na realidade, existe uma transição do último estado de cada bloco para o primeiro estado de qualquer um deles.

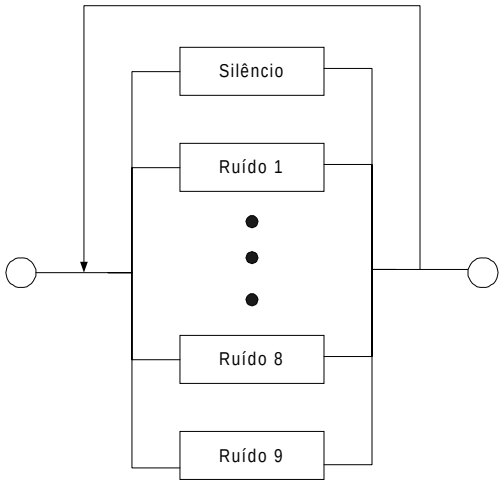

Figura 2 – Arquitectura do Novo Modelo de Ruído

O modelo foi construído a partir de um HMM com 32 estados dos quais o primeiro e o último não são emissores e os restantes correspondem aos 3 estados de cada um dos 10 modelos treinados. O primeiro estado (não emissor) foi ligado ao primeiro estado de cada um dos sub-modelos (estados 2, 5, 8, …, 29) com a mesma probabilidade de transição (0.1). O primeiro e segundo estados de cada sub-modelo manteve-se inalterado. As probabilidades de transição do último estado de cada bloco (estados 4, 7, 10, …, 31), quer para a saída quer novamente para o início, foram estimadas no conjunto de treino. A probabilidade de abandonar o modelo de ruído em cada bloco foi multiplicada por 60% e cada transição para o início correspondeu a 40% da probabilidade de sair do bloco a dividir pelo número de blocos. De notar que nesta fase ainda não foi feita nenhuma reestimação do modelo, pelo que o processo de construção apenas procurou criar um modelo inicial o mais próximo possível dos modelos que lhe serviram de base. Tal como no sistema base o número de misturas foi gradualmente incrementado de 1 para 3, para todos os modelos.

Os dígitos ligados foram novamente treinados escolhendo como modelos iniciais os melhores modelos de dígitos isolados obtidos, substituindo na transcrição ortográfica todas as marcas de ruído e silêncio existentes pelo novo modelo de ruído. Foi repetido o processo iterativo de reestimação/avaliação e são apresentados, na tabela 5, os resultados da aplicação desta extensão (EXT) ao sistema base (SB), correspondentes a modelos de dígitos e de ruído, respectivamente, com 5 e 3 misturas gaussianas por estado.

	% Correcção		Precisão	
	SB	EXT	SB	EXT
Dígitos Isolados	99,0	**99,4**	99,0	**99,4**
Dígitos Ligados (Gramática de Comprimento Fixo)	97,8	**98,1**	97,6	**98,0**
Dígitos Ligados (Gramática de Comprimento Variável)	97,2	**97,9**	95,1	**96,1**

Tabela 4 – Resultados Comparativos do Sistema Base e Respectiva Extensão

Podemos observar que houve um decréscimo da ordem dos 40% na taxa de erro para a tarefa de reconhecimento de dígitos isolados em relação ao sistema base. Foram também obtidas melhorias significativas no que diz respeito aos dígitos ligados, principalmente na sua vertente livre.

3.3 Números Naturais

A tarefa de reconhecimento de números naturais é composta por três sub-tarefas com graus de complexidade diferentes e crescentes: números estritamente naturais, quantias, combinação de números naturais e dígitos ligados. Esta última é frequente ao reconhecer, por exemplo, números de cartões de crédito constituídos por 16 dígitos organizados em grupos de quatro e que podem ser lidos de diversas formas,

frequentemente alternando entre dígitos ligados e números naturais. Os resultados e técnicas aqui descritos referem-se ainda apenas aos números estritamente naturais.

Uma vez que as dimensões do léxico são significativamente maiores, optámos por utilizar para esta tarefa modelos de fones. Por forma a dispor de modelos iniciais recorremos a modelos já treinados para uma tarefa de reconhecimento de topónimos pelo Engº Manuel João Silva. Usámos esses modelos para alinhar automaticamente o conjunto de treino por ele utilizado. Com base nesse alinhamento construímos os modelos, de acordo com a topologia descrita na figura 1 mas sem qualquer salto e de acordo com a nossa codificação (que é ligeiramente diferente). Seguiu-se então um processo iterativo, recorrendo ao algoritmo de Viterbi, para estimar as médias e variâncias de cada modelo HMM de fone. No que diz respeito aos modelos de dígitos de que dispúnhamos optámos por continuar a usá-los. Assim, na transcrição associada a cada dígito, em vez de aparecer a sequência de modelos de fones que o caracterizam, aparece o modelo de palavra treinado anteriormente. Dos testes efectuados e resultados obtidos verificou-se vantagem nesta abordagem. Contudo, à medida que o treino dos modelos de fones prossegue, essa vantagem tende a diluir-se. Este facto leva-nos a crer que pode ser útil começar o treino com este modelos e substitui-los gradualmente pelos modelos de fones. Foram realizadas uma série de reestimações que conduziram aos valores apresentados na tabela seguinte, para o conjunto de desenvolvimento. Num passo subsequente foi aumentado o número de misturas.

Nº de Misturas	% Correcção	Precisão
1	90,9	90,2
2	95,5	95,0

Tabela 5 – Resultados de Reconhecimento de Números Naturais

É importante realçar que, ao contrário do que foi feito relativamente aos dígitos isolados, as transcrições das locuções não foram ainda completadas, ao nível da anotação de eventos não–linguísticos, pelo que é natural que os valores venham a melhorar substancialmente. Por outro lado, a quantidade de dados para treinar os modelos de fones pode ser aumentada e é possível refinar a gramática de modo a que seja mais flexível sem baixar o nível de desempenho.

4 Conclusões e Trabalho Futuro

Apesar dos resultados para a tarefa de dígitos ligados estarem perto dos 98%, é nossa convicção de que é possível melhorar e atingir valores ao nível do estado da arte para outras línguas, [Rahim95], [Caminero97]. Por um lado, foi possível verificar que existe uma grande variabilidade nos dígitos, em particular nas zonas de fronteira, que poderão ser difíceis de modelar com modelos de palavra. Por outro lado, o subconjunto usado nesta tarefa não era completamente representativo do universo dos dígitos, dado serem pouco frequentes repetições do mesmo dígito consecutivamente. O trabalho futuro nesta tarefa centrar-se-á na incorporação de mais dados e na comparação do desempenho entre os modelos de palavra e os modelos de fones.

No que diz respeito aos números naturais, os resultados preliminares são muito encorajadores na medida em que dispomos ainda de um largo leque de técnicas para explorar que irão sem dúvida melhorar os resultados obtidos, por exemplo, através da introdução de informação de contexto ou aproveitando o facto de algumas palavras que do vocabulário terem elementos comuns (e.g. seis*centos*, …, nove*centos*), etc..

Finalmente, e de alguma forma dependentes do rumo escolhido para as tarefas referidas anteriormente, estão o reconhecimento de quantias monetárias (muito

próxima da tarefa dos números naturais) e de locuções que combinem o uso de números naturais e dígitos ligados.

Foi construída uma aplicação de interface com a linha telefónica que utiliza os modelos obtidos e que permite ao utilizador, utilizando apenas a fala, experimentar o sistema. O utilizador pode escolher o que pretende que seja reconhecido, dígitos ligados ou números naturais, e ouvir o resultado, sintetizado, pela linha telefónica. A síntese é efectuada por concatenação de dígitos ligados através do sistema SVIT[2] ou por regra através do sistema DIXI[3].

5 Agradecimentos

Este trabalho foi realizado no âmbito da Tese de Mestrado do Eng.º Frederico Rodrigues intitulada "Reconhecimento Robusto de Dígitos e Números Naturais", com uma bolsa (GGP XXI / BM / 3782 / 96) da Fundação para a Ciência e Tecnologia.

6 Bibliografia

[Caminero97] F.J. Caminero, L. Hernandez-Gomez, C. De La Torre, , C. Martin del Alamo, "Improving Utterance Verification Using Hierarchical Confidence Measures in Continuous Natural Numbers Recognition", Proc. ICASSP'97, Munique, Alemanha, 1997.

[Johnston97] D. Johnston et al., "Current and Experimental Applications of Speech Technology for Telecom Services in Europe", Speech Communication 23, pp. 5-16, 1997.

[Perdue97] R. J. Perdue, "The Way We Were: Speech Technology, Platforms and Applications in the 'Old' AT&T", Speech Communication 23, pp. 31-39, 1997.

[Rahim95] M. Rahim, C.H. Lee, B.H. Juang, "Robust Utterance Verification for Connected Digits Recognition", Proc. EUROSPEECH'95, Madrid, Espanha, 1995.

[2] http://www.speech.inesc.pt/~lco/svit
[3] http://www.speech.inesc.pt/~lco/dixi