CORPORA DE FALA EM P.E.
CONSTITUIÇÃO, SEGMENTAÇÃO E ETIQUETAGEM

M. Céu Viana
Isabel Mascarenhas
CLUL
Isabel Trancoso
Luís C. Oliveira
INESC/IST
Carlos M. Ribeiro
INESC/ISEL

Introdução

A utilização de corpora de fala de grandes dimensões tornou-se um procedimento comum tanto para o treino e teste de sistemas de síntese e reconhecimento de fala como, em termos gerais, para os trabalhos de investigação fundamental que têm por objectivo o estudo da relação entre os diferentes níveis da representação linguística e as propriedades físicas dos sons Todos estes trabalhos requerem a formulação de hipóteses sobre conjuntos de dados empíricos relativamente vastos e o teste dessas hipóteses sobre novos conjuntos de dados. Para qualquer destes efeitos não basta, contudo, possuir gravações em banda magnética, nem sequer um número significativo de ficheiros de fala acessíveis em computador. Para que os materiais de fala possam ser eficazmente utilizados para fins de investigação, é necessário que estejam anotados e documentados e que a sua qualidade seja controlada.

A constatação desta necessidade tem conduzido a esforços consideráveis realizados para várias línguas, tanto no âmbito de programas nacionais como europeus. Apenas a título de exemplo, e para dar uma ideia do tipo de recursos que estão envolvidos, estão actualmente em curso trabalhos para o inglês americano que envolvem a recolha e tratamento de cerca de 30000 frases, 10000 das quais são produzidas por um conjunto de falantes não inferior a 60. As restantes são ditas por conjuntos mais pequenos de falantes e da sua análise pretende-se extrair informações sobre a variação intra e inter-locutor.

Para o Português Europeu (P. E.), existem apenas algumas recolhas de materiais de fala quer para efeitos de estudo de diferentes aspectos parcelares da sua fonética e fonologia, quer para o teste de ferramentas de processamento automático. Devido às limitações materiais em que a maior parte desses trabalhos foram realizados, as condições de recolha não são comparáveis entre si, como também o não são os critérios utilizados na anotação dos dados, sendo difícil, se não mesmo impossível, uma acumulação gradual de conhecimentos sobre o P.E. Este facto tem vindo a constituir naturalmente um sério entrave ao bom andamento dos trabalhos de investigação em curso na área do processamento automático de fala, tanto de um ponto de vista da linguística como da engenharia.

No planeamento das actividades a realizar no âmbito do convénio entre o CLUL e o INESC tem-se procurado preencher esta lacuna, participando em diferentes projectos que têm por objectivo a recolha e tratamento de corpora de fala e garantindo a coerência das condições de recolha e dos critérios de anotação. Dada a morosidade do trabalho que é preciso realizar e a escassez dos recursos materiais e humanos disponíveis para o efeito, têm vindo a ser utilizadas ferramentas de processamento automático na realização de algumas tarefas, procurando reduzir, sempre que possível, o processamento manual à verificação e correcção de erros.

Depois de uma breve apresentação dos corpora existentes e em preparação e das condições de recolha que têm vindo a ser utilizadas, esta comunicação centrar-se-á no tratamento dado a um pequeno conjunto de entrevistas e excertos de noticiários televisivos (a integrar no corpus BDFALA) e a uma parte dos materiais recolhidos no âmbito do projecto europeu SAM_A [15].

A anotação sistemática de corpora de fala encontra-se ainda numa fase incipiente e estes materiais têm sido utilizados sobretudo como base para uma proposta da metodologia geral a seguir, isto é, como meio para definir os principais critérios de anotação e desenvolver e testar um conjunto de ferramentas de tratamento automático ou semi-automático, de modo a reduzir o trabalho de constituição, recolha e anotação de corpora.

Como consideramos que um dos aspectos mais interessantes da investigação futura nesta área será, certamente, o da comparação e confrontação de diferentes descrições feitas sobre os mesmos materiais sonoros, pensámos que seria interessante pôr desde já em discussão os níveis de representação e os principais critérios seguidos no alinhamento das representações simbólicas com o sinal de fala (segmentação e etiquetagem).

1. Corpora de fala existentes e em preparação para o P.E.

São fundamentalmente três os corpora de fala actualmente recolhidos ou em preparação para o Português Europeu, no âmbito convénio CLUL-INESC.

(1) – EUROM.1 (projecto europeu ESPRIT III SAM_A- já concluído)
– BDFALA (projecto JNICT/PLUS – em curso)
– SPEECHDAT (projecto europeu TELEMATICS – em curso)

Nem todos os materiais têm vindo a ser recolhidos em condições idênticas: as recolhas para o corpus SPEECHDAT são feitas via linha telefónica e uma parte do corpus BDFALA é constituída por gravações de noticiários, entrevistas e debates que tiveram lugar em meios de comunicação social. À excepção destes casos particulares, todos os outros materiais têm vindo a ser gravados em condições idênticas e são comparáveis entre si. Sempre que possível, têm também vindo a ser asseguradas as mesmas condições de compatibilidade para todos os materiais parcelares que têm vindo a ser recolhidos tanto no âmbito do CLUL como do INESC, em projectos conjuntos ou não. De entre estes materiais, destacam-se os que constituem o corpus RED-93, recolhido no âmbito do projecto JNICT "Estudo Experimental de Processos de

Lenição Vocálica em Português Europeu" liderado por A. Andrade e os conducentes ao estudo de diferentes aspectos da prosódia desta língua, no âmbito do projecto JNICT "Sistema de Apoio Vocal para Pessoas com Deficiência Motora e de Fala", nomeadamente os utilizados em [8] e [20]. Esta uniformização é fundamental para a constituição de uma base de dados de fala para o Português Europeu que possa vir a ser progressivamente alargada.

1.1 O Corpus EUROM.1 para o P.E.

A recolha sistemática de materiais de fala para o P.E. foi iniciada no âmbito do projecto europeu ESPRIT III SAM_A "Speech Technology Assessment in Multilingual Applications". Este projecto teve por objectivo a construção de corpora de fala para o Português, Castelhano e Grego. Trata-se de uma extensão de outro projecto, designado por SAM, em que foram aferidos métodos de entrada e saída de fala, estabelecidas metodologias para a aquisição de dados e construídas para diferentes línguas europeias bases de dados estruturalmente idênticas, comparáveis entre si e capazes de possibilitar uma aferição adequada dos métodos de reconhecimento e de síntese de fala desenvolvidos nos vários países. Na aquisição da base de dados EUROM.1 para o Português Europeu, foram rigorosamente seguidas todas as recomendações emanadas da fase anterior do projecto, tanto no que diz respeito à definição dos conteúdos como às condições de gravação e aos equipamentos utilizados.

1.1.1 Conteúdo do Corpus

No que diz respeito ao conteúdo, e à semelhança do que tinha sido feito para outras línguas, o corpus para o Português Europeu é constituído pelos seguintes subconjuntos de materiais de fala (cf. [15]):

(2) 1 – 100 números (entre 0 e 9999);
2 – 121 palavras (ou logátomos) com estrutura CVC, CVCV, CCVC organizados em sete blocos diferentes que, em conjunto, permitem analisar alguns aspectos da variação contextual;
3 – 40 passagens de 5 frases cada uma, das quais 20 são semelhantes às recolhidas para as outras línguas (tradução livre ou adaptação) e as outras 20 correspondem a frases construídas de raiz ou a excertos de artigos de jornais ou de livros;
4 – Um conjunto de frases de compensação desenhadas para, em conjunto com as passagens, assegurarem a cobertura total dos difones e, não contando com o peso das palavras funcionais, garantirem o equilíbrio fonético do corpus em termos da frequência relativa dos segmentos fonéticos.

Para o levantamento do conjunto total de difones e para o estabelecimento das frequências relativas de ocorrência de fones foi utilizado o corpus de pronúncia PF_FONE, constituído a partir do conjunto de entrevistas realizadas para o Português Fundamental [19].

1.1.2 Locutores

Os materiais acima referidos foram lidos por 60 locutores, 30 do sexo feminino e 30 do sexo masculino, com idades compreendidas entre os 10 e os 65 anos de idade, com especial incidência no grupo etário entre os 20 e os 40 anos, como o quadro em (3) ilustra.

À semelhança também do que foi feito para outras línguas, os locutores foram divididos em 3 sub-grupos, designados como *Muitos*, *Poucos* e *Muito poucos*. Os materiais gravados por cada um destes grupos são apresentados no quadro em (4).

(3)

IDADE	SEXO FEMININO	SEXO MASCULINO
10-12	1	1
18-20	1	1
20-29	12	13
30-39	9	11
40-49	5	2
> 50	2	2
TOTAL	30	30

(4)

SUB-GRUPOS	LOCUTORES	MATERIAIS
MUITOS	30 homens 30 mulheres	100 números 3 passagens de 5 frases cada uma 5 frases de compensação
POUCOS	5 homens 5 mulheres	100 números isolados com 5 repetições cada um; 15 passagens de 5 frases cada uma; 25 frases de compensação
MUITO POUCOS	1 homem 1 mulher	121 palavras ou logátomos, em 5 frases fixas diferentes; 10 palavras isoladas que são utilizadas nas frases fixas, repetidas 5 vezes

Repare-se que os grupos de locutores não se excluem: Os *muito poucos* fazem parte dos *poucos* e ambos se incluem no grupo dos *muitos*. Assim, por exemplo, tanto os *poucos* como os *muitos* leram cerca de 600 números: os 100 que também foram lidos pelos *muitos* e as cinco repetições de cada número em frases fixas diferentes.

1.1.3 Condições de gravação e equipamento

Todas as gravações foram realizadas em câmara anecóica e cuidadosamente monitorizadas. Os locutores encontravam-se comodamente instalados, falando a uma distância de 30 cm do microfone, com uma inclinação de 15º. Os materiais de leitura foram organizados em blocos, apresentados com uma cadência constante em écran vídeo e gravados simultaneamente de dois modos diferentes: em gravador DAT SONY DTC-57ES, com uma frequência de amostragem de 44 kHz, e directamente para o computador utilizando uma placa ORUS AU22 e uma frequência de amostragem de 20 kHz (pacote EUROPEC desenvolvido durante o projecto SAM).

Um aspecto considerado durante o processo de gravação foi o da linearidade da fase, tendo como objectivo uma análise de predição linear síncrona com o período da frequência fundamental para estudo dos efeitos da excitação das cordas vocais e do filtro constituído pelos tractos vocal e nasal. Essa linearidade é garantida pela utilização de um microfone de fase linear (B&K 2230 de 1/2 polegada) e pela aquisição do sinal em câmara anecóica directamente para o gravador DAT, cuja frequência de amostragem característica é 44 kHz. Deste gravador, o

sinal de fala pode ser extraído para disco a qualquer frequência de amostragem inferior por *down sampling* digital, dispensando filtragem analógica de fase não linear. Este tipo de gravação é essencial não só de um ponto de vista da investigação fundamental mas também para posterior utilização em várias áreas do processamento de fala, nomeadamente para o desenvolvimento de sintetizadores articulatórios.

Para o controle dos níveis de gravação foi utilizado um sonímetro B&K 2230 e para o armazenamento e salvaguarda dos dados, uma unidade de *back-up* PYTHON DAT.

1.2. O Corpus BDFALA

Este corpus – que foi definido e está a ser recolhido no âmbito do projecto BDFALA – surge como uma extensão natural do projecto anterior. Pretende-se assegurar a continuação das recolhas de materiais de fala para o Português Europeu, contemplando novos conjuntos de dados que se consideram imprescindíveis para a análise de diferentes aspectos da fonética desta língua, assim como para o treino e teste de algoritmos de processamento automático de fala. Pretende-se ainda, por outro lado, criar condições para um melhor aproveitamento dos recursos materiais e humanos disponíveis, definindo critérios uniformes de anotação e desenvolvendo um conjunto de ferramentas que permitam facilitar o trabalho de constituição e tratamento de corpora, reduzindo, tanto quanto possível, a intervenção manual à correcção de erros. O conjunto de materiais que nos propusemos recolher no âmbito deste projecto não é ainda suficiente para o desenvolvimento de muitas aplicações. Os materiais seleccionados são, no entanto, suficientemente diversificados para o desenvolvimento e teste de ferramentas auxiliares para a constituição, recolha e tratamento de materiais de fala que consideramos essenciais para um rápido alargamento do corpus a um maior número de locutores e de estilos de fala.

1.2.1 Conteúdo do Corpus

O corpus BDFALA contempla 5 subconjuntos de materiais, parte dos quais ainda se encontram subdivididos, como (5) mostra.

(5) A *Lista de 100 palavras* lidas em modo contínuo e em modo de silabação.

B 990 *logátomos* para extracção de difones a utilizar na síntese por concatenação de unidades pré-gravadas. É assegurada uma cobertura total dos difones que podem ocorrer em posição inicial, média e final de palavra, contemplando já (embora parcialmente) encontros consonânticos resultantes da elisão de vogais átonas.

C 1 – 500 *enunciados* especialmente desenhados para permitirem um estudo mais aprofundado dos padrões prosódicos e da sua relação com os fenómenos de *sandhi* externo e interno.

2 – 200 *provérbios* – seleccionados como complemento do subcorpus anterior. Dada a naturalidade com que são ditos os provérbios relativamente conhecidos supõe-se que serão menos afectados pela entoação artificial que os locutores não profissionais frequentemente adoptam ao ler outro tipo de enunciados.

3 – 30 *textos de jornais*. Este corpus, que deverá vir a ser alargado, é constituído actualmente por 10500 palavras que ocorrem em cerca de 250 parágrafos.

4 – 60 *enunciados* seleccionados a partir do subcorpus E (fala espontânea). De momento pretende-se apenas constituir uma pequena amostra para um estudo piloto comparativo dos agrupamentos prosódicos que ocorrem na fala lida e na fala espontânea e dos seus reflexos na qualidade fonética dos segmentos.

D 1 – 3500 *Palavras isoladas e em frase fixa* – para controle dos efeitos da estrutura silábica, da posição do acento lexical e dos ecos do acento a nível da palavra.

2 – 500 *sequências de 2 e 3 palavras* que podem constituir unidades entonacionais mas que não formam necessariamente frases – para análise da resolução de encontros vocálicos e consonânticos em fronteira de palavra.

3 – 18 *enunciados foneticamente completos*, seleccionados automaticamente de textos de jornais utilizando o algoritmo *greedy* [7] [17]. Cada enunciado contém, em média, cerca de 40 palavras e pode ser utilizado para diversos fins, nomeadamente para testes de avaliação ou para selecção de locutores a utilizar na síntese por concatenação.

Pode também ser utilizado como semente de outros algoritmos *greedy* para completar o corpus a nível de difones, trifones, etc.

E 1 – 4 *entrevistas curtas* de cerca de 10-12 minutos cada.

2 – 1 *Debates televisivos* em que intervêm 4 pessoas para além do moderador.

3 – 1 *Excerto de Noticiário* que inclui várias intervenções em directo

1.2. 2 Locutores e condições de gravação

A maior parte dos materiais estão a ser gravados por 10 locutores (5 homens e 5 mulheres) que já tinham participado nas recolhas feitas para o projecto SAM_A. Exceptuando naturalmente o subcorpus E e o subcorpus B que está a ser gravado apenas por 1 homem e uma mulher cujas vozes foram seleccionadas para a síntese por concatenação. A quantidade de materiais incluídos nos diferentes subconjuntos não é fixa e tem vindo a ser aumentada progressivamente, sendo imprescindível, também, alargar o número de locutores. De momento, pretendeu-se apenas criar um conjunto inicial que assegurasse uma cobertura razoável dos principais fenómenos sonoros da língua, e permitisse iniciar o treino e teste de algoritmos de processamento do Português Europeu falado.

Embora obedecendo ao conjunto de requisitos básicos do projecto SAM, foram introduzidas algumas alterações nas condições de recolha de dados. As gravações e a digitalização foram feitas utilizando apenas o gravador DAT e uma unidade de controle DATLink, em vez do pacote EUROPEC. Por uma questão de facilidade de utilização dos equipamentos e para criar condições mais agradáveis e naturais para os locutores, as gravações passaram a ser feitas em câmara surda, sendo o seu ritmo controlado pelos locutores e os materiais a ler apresentados quer num terminal alfanumérico quer em fichas ou folhas A4.

1.3 O corpus SPEECHDAT

A recolha de materiais de fala via linha telefónica tem como principal objectivo o treino de reconhecedores de fala independentes do locutor, para aplicação em serviços telefónicos. Dado que, neste treino, é necessário contemplar não só a variabilidade a nível de locutores

(características físicas, estilos de fala, dialectos), mas também a nível de microtelefones, canais telefónicos e ruídos ambientais, a recolha tem de incluir vários milhares de chamadas telefónicas. O projecto europeu SPEECHDAT [22], no qual a presente equipa de investigação participa subcontratada pela Portugal Telecom, tem justamente por objectivo a recolha de 5000 chamadas efectuadas por locutores diferentes. A primeira fase, quase concluída, consiste numa recolha piloto de apenas 1000 chamadas por língua, feita por instituições de 8 países. A segunda fase está já em definição, contando com uma participação alargada a quase uma vintena de línguas.

Cada chamada contém uma parte de respostas espontâneas e outra de respostas lidas. A primeira tem como objectivo a recolha de informação acerca do locutor (p. e. sexo, data do nascimento, local onde passou a maior parte da infância) e das condições de gravação (p.e. telefone sem fios). Simultaneamente, pretende-se com estas respostas recolher um vocabulário essencial a um grande número de aplicações: as palavras *sim* e *não* e a forma espontânea de dizer datas e horas. São 9 as respostas espontâneas incluídas na primeira fase da nossa recolha.

A segunda parte consiste num conjunto de itens lidos. Para esse efeito, foram elaboradas para a primeira fase 1000 fichas de leitura diferentes, cujo esqueleto é o apresentado em (6), onde @ delimita itens que são automaticamente preenchidos com os materiais seleccionados para este corpus, utilizando um *script* especialmente concebido para esse efeito.

(6) @indice@

Pedir-lhe-emos agora que leia a coluna da direita da seguinte lista:

1.	Leia a frase:	@frase1@
2.	Leia o número por extenso:	@numero56@
3.	Leia a frase:	@ws1@
4.	Soletre a palavra letra a letras:	@pal-soletr1@
5.	Leia a frase:	@frase6@
6.	Leia as horas:	@hora-a1@
7.	Leia a palavra	@palavra1@
8.	Leia a quantia em dinheiro:	@quant-elev@
9.	Leia a frase:	@frase2@
10.	Leia o número do cartão de crédito:	@cartao-cred@
11.	Leia a frase:	@ws2@

12.	Leia o número por extenso:	@numero3@
13.	Leia a frase:	@frase8@
14.	Leia a data:	@data1@
15.	Leia a palavra:	@palavra2@
16.	Soletre a palavra letra a letra:	@pal-soletr2@
17.	Leia a frase:	@frase1@
18.	Leia o número de telefone:	@num-telef@
19.	Leia a frase:	@ws3@
20.	Leia a palavra:	@palavra3@
21.	Leia a frase:	@frase7@
22.	Leia a data:	@data2@
23.	Leia a palavra:	@palavra1@
24.	Leia a quantia em dinheiro:	@quant-peq@
25.	Leia a frase:	@frase3@
26.	Leia a palavra	@palavra5@
27.	Leia o número	@digito-isol@
28.	Soletrte a palavra letra a letra	@pal-soletr3@
29.	Leia a frase	@frase5@
30.	Leia o número por extenso	@numero1@
31.	Leia a frase	@frase9@
32.	Leia a palavra	@palavra6@

Muito obrigada pela sua participação.
Esperamos que seja um dos premiados no sorteio. Até à próxima

Cada ficha contém 32 itens diferentes, para além de um número de identificação: 9 frases foneticamente ricas, 4 números naturais, 6 palavras de comando, 3 palavras soletradas (nomes próprios, no nosso caso), 2 quantias em dinheiro, 2 datas, 1 hora, 1 número de cartão de crédito, 1 número de telefone e 3 frases em que se inclui uma palavra de comando para efeitos de teste de sistemas de detecção de palavras-chave. A grande maioria das frases foneticamente ricas foi extraída (e/ou adaptada) de textos jornalísticos, impondo um limite máximo ao número de palavras de cada frase. O algoritmo de selecção foi elaborado de modo a garantir pelo menos dois exemplos de cada fone no conjunto de 9 frases de cada folha, procurando um número de difones diferentes o maior possível (ver secção sobre ferramentas).

As gravações via rede telefónica são efectuadas utilizando um pacote de gravação desenvolvido pelo INESCTEL, sob consultadoria

da equipa de processamento digital de fala do INESC e uma plataforma de *hardware* muito comum neste tipo de aplicações que é constituída por um computador pessoal equipado com uma placa D/81A da firma DIALOGIC.Estas placas de interface telefónica realizam a codificação e descodificação da fala a vários ritmos (métodos PCM e ADPCM), a geração de tons multi-frequência, o controle automático de ganho e a detecção de silêncios para um conjunto de canais telefónicos. As mensagens do programa terminam com um sinal sonoro (BIP) que indicam ao utilizador que o sistema está pronto a gravar. O tempo de gravação é controlado pelo detector de silêncios e por um temporizador.

2. Ferramentas auxiliares para a constituição e recolha de corpora

O equilíbrio fonético e fonotáctico é um critério fundamental que, à semelhança do que tem vindo a ser feito em iniciativas congéneres para outras línguas, se procurou observar em todos os projectos de recolha de corpora acima mencionados. Garantir este critério não é uma tarefa simples: os enunciados a gravar são em geral seleccionados por tentativa e erro e, por vezes, a saída mais viável para garantir esse equilíbrio consiste na inclusão de conjuntos de frases de compensação (ex. corpus EUROM.1), nem sempre fáceis de ler. Dado que a selecção dos materiais a gravar é uma questão recorrente, importante na fase inicial de desenho de um corpus e em todas as tentativas posteriores de alargamento do mesmo, foi desenvolvido um pacote com 3 programas que pode ser utilizado sempre que se pretenda garantir um critério de cobertura fonética ou fonotáctica ou uma distribuição uniforme de um dado vocabulário:

(7) 1. *Guloso* ficheiro.entrada_1 ficheiro.entrada_2

Implementa o algoritmo *greedy* (cf. [7], [17]). A ideia fundamental por detrás deste algoritmo é aumentar o primeiro ficheiro de entrada com as linhas do segundo ficheiro de entrada que mais enriqueçam o primeiro de um ponto de vista da cobertura de fones, difones ou trifones distintos. Foram implementadas várias versões deste algoritmo correspondendo a critérios diferentes.

Pode garantir-se, por exemplo, um mínimo de duas ocorrências de cada fone no ficheiro aumentado e, uma vez obtida essa garantia, tentar enriquecê-lo em termos de difones ou trifones. A saída deste algoritmo também varia conforme a versão sendo seleccionadas, no caso mais simples, as linhas do segundo ficheiro de entrada com um maior número de trifones adicionais. (Note-se que *linha*, neste sentido, não é entendida como uma linha do terminal propriamente dita, mas como qualquer número de linhas que, no editor de texto, não se encontram separadas por parágrafos, o que garante que não são nunca truncadas frases ou enunciados que se pretendem coesos).

2. *verif_fones* ficheiro.entrada

   Verifica se a transcrição incluída no ficheiro de entrada é foneticamente completa, isto é, se esta inclui pelo menos um exemplar de cada fone. No caso afirmativo a saída é nula e, no caso negativo, é fornecida uma lista dos fones inexistentes no ficheiro.

3. *gerador_unif*

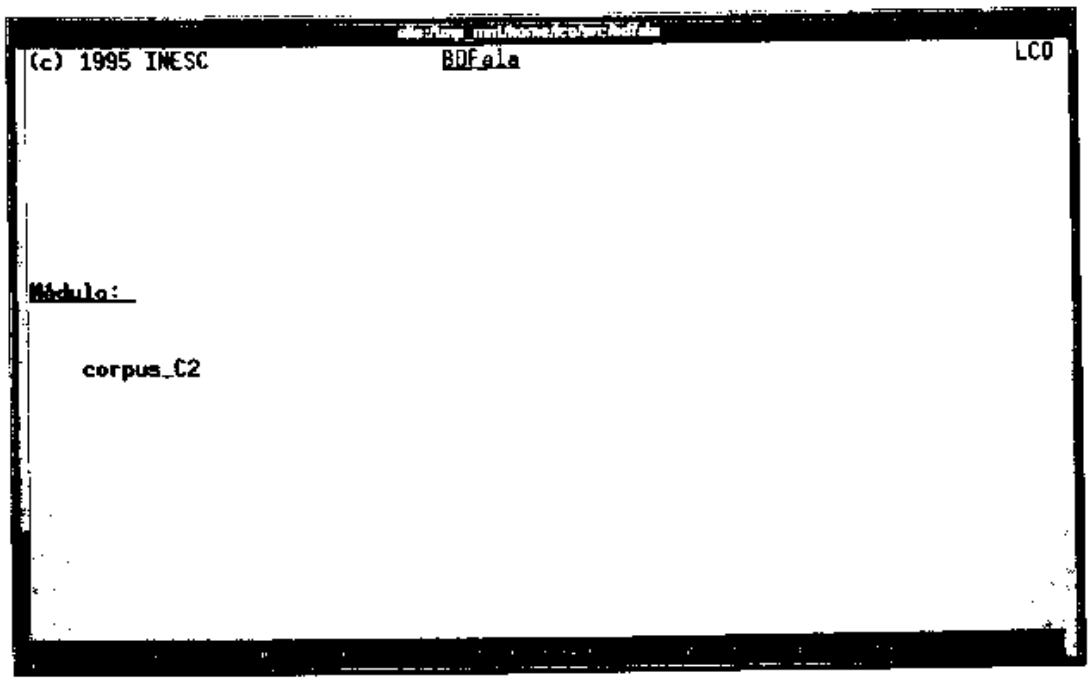
   Gera números, palavras ou frases com uma distribuição o mais uniforme possível de um determinado vocabulário (dígitos, por exemplo, no caso dos números).

A maior parte das gravações já efectuadas na câmara contêm grandes intervalos em que apenas se ouvem os ruídos do virar das fichas, do aclarar da voz, ou da mudança de posição na cadeira e correspondem a ficheiros de fala demasiado grandes e difíceis de manipular. A primeira tarefa consiste, por conseguinte, na eliminação desses intervalos e no armazenamento das produções dos falantes em ficheiros menores, de acesso mais rápido e mais eficiente.

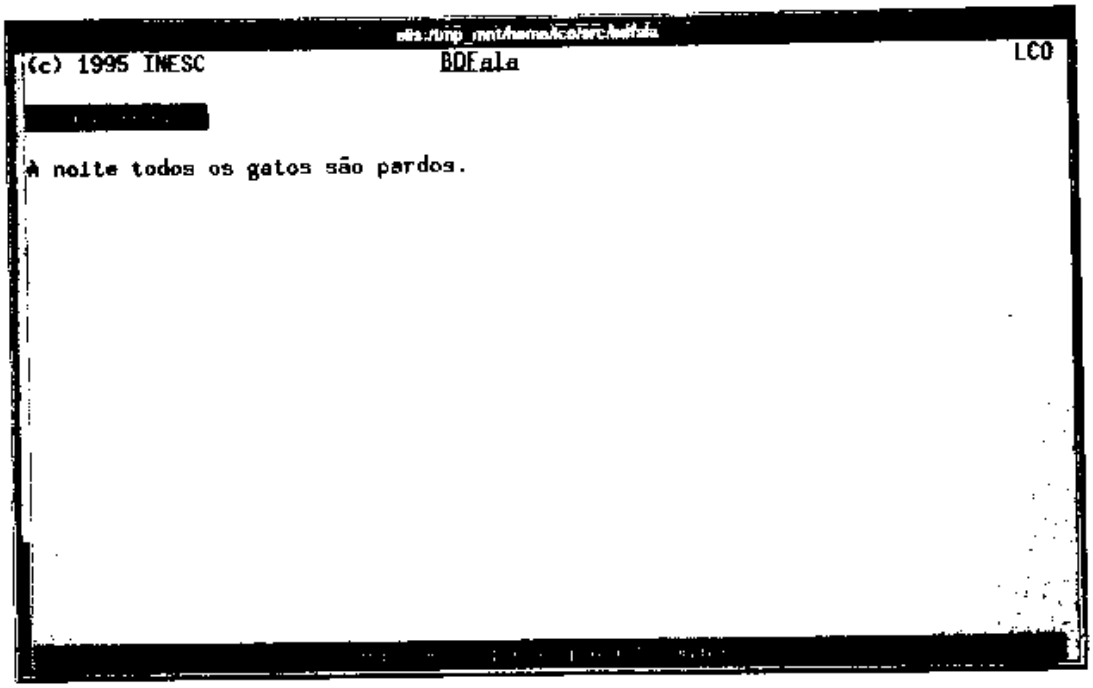
Apesar de a maior parte dos programas de edição de fala que se encontram actualmente disponíveis no mercado permitirem realizar esta tarefa com relativa facilidade, o tempo dispendido na segmentação, nomeação e verificação de ficheiros de fala é ainda considerável. Para automatizar esta tarefa está a ser desenvolvido um pacote de programas especificamente para gravação em câmara surda, já em fase de teste. Este pacote foi desenhado de forma a permitir a gravação o mais auto-suficiente possível dos materiais de leitura recolhidos para cada locutor. Os materiais são apresentados num terminal alfanumérico que se encontra no exterior para evitar ruído na gravação, mas que está posicionado diante de uma janela de vidro duplo, sendo claramente visível do interior da câmara.

Nesse terminal são apresentados, em instantes diferentes, o nome do corpus (8.a), o enunciado a ler (8.b) e o menu de opções (8.c). Um teclado, colocado no interior da câmara, permite ao locutor controlar o programa de gravação.

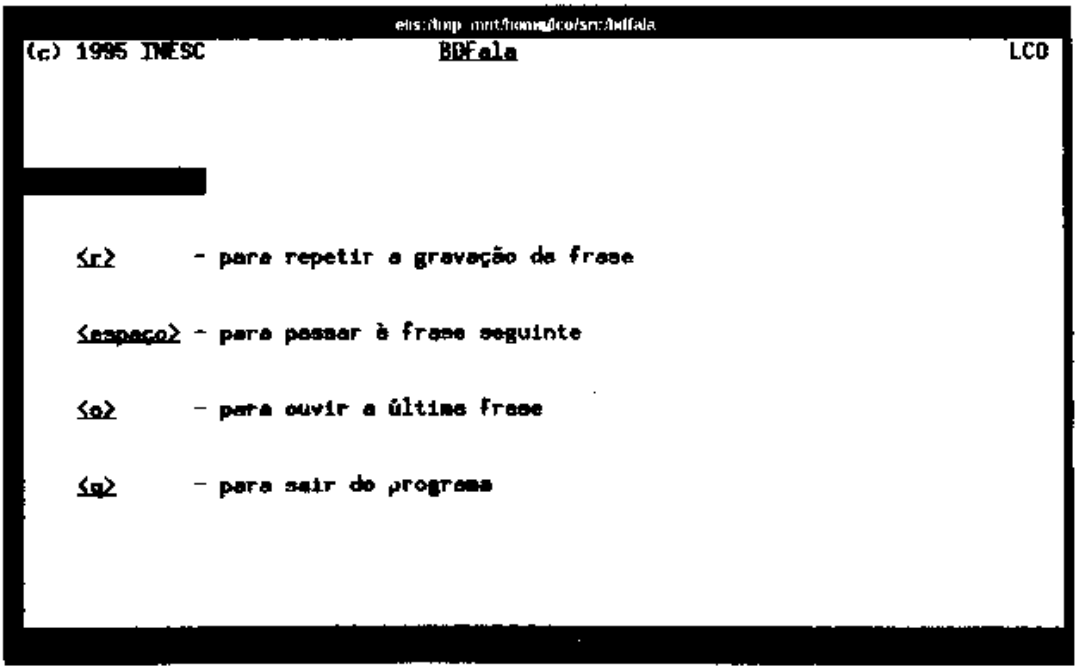
(8) a.



(8) b.



(8) c.



A gravação inicia-se quando é pressionada a tecla *i* e termina quando é pressionada a barra de espaços. Esta última acção desencadeia a gravação em disco (através do sistema DAT_Link) de um sinal, do qual é extraída automaticamente a parte audível correspondente à pressão da barra de espaços

Os excertos de fala armazenados são precedidos e seguidos de um intervalo de silêncio de dimensões reduzidas. Embora a duração destes intervalos de silêncio seja um parâmetro do programa que pode ser alterado, nas nossas recolhas o seu valor foi fixado em 100 ms, valor que foi considerado suficiente para garantir uma separação visual clara entre as cercaduras das janelas utilizadas nos programas de análise e os instantes de início e fim dos excertos de fala seleccionados.

O nome do ficheiro contém as iniciais do locutor, a indicação do subcorpus e o número do item a que a gravação se refere. Uma vez armazenado o sinal de fala, o falante pode ouvir a gravação, repeti-la, se não a considerar natural ou adequada, ou passar ao item seguinte.

Este pacote não pode ser utilizado, naturalmente, na recolha de amostras de fala espontânea, em que os intervalos de silêncio, as sobreposições de voz, as hesitações, os risos, etc. são de fundamental

importância, tanto para o treino e teste de algoritmos de processamento de fala natural, como para a compreensão das estratégias utilizadas pelos falantes quando pretendem tomar a palavra, dá-la a outros ou simplesmente manifestarem a sua concordância ou discordância em relação ao que está a ser dito. Nestes casos, não existe material escrito prévio e é necessário proceder à transliteração manual de todas as gravações.

3. Níveis de anotação

A questão da anotação é, sem dúvida, uma questão central na constituição de corpora: é preciso decidir o que se anota e como se anota e definir um conjunto de critérios relativamente seguros que permitam manter um nível de consistência razoável de anotador para anotador.

No projecto SAM_A foram apenas considerados dois níveis de anotação: um primeiro nível que contém a representação ortográfica de todos os materiais lidos, e um outro em que é fornecida uma transcrição fonética larga correspondente à forma fonética das palavras quando pronunciadas isoladamente e em estilo formal (forma de citação). Nas transcrições, foi utilizado o alfabeto SAM_PA, alfabeto fonético para computador desenvolvido também no âmbito do projecto SAM. As formas fonéticas efectivamente produzidas pelos locutores na leitura do corpus não foram contempladas e o alinhamento das anotações com o sinal acústico é apenas assegurado em relação ao início e fim dos blocos de dados.

Os níveis de anotação contemplados no projecto SAM_A são considerados níveis mínimos que se asseguram para todos os corpora recolhidos no âmbito do convénio. Eles são, no entanto, manifestamente insuficientes para treino e teste da maior parte dos algoritmos de processamento de fala natural, tanto de um ponto de vista da engenharia como da linguística, que requerem um maior número de níveis de anotação e uma segmentação e etiquetagem muito mais finas do sinal de fala a cada um desses níveis.

A definição dos critérios de segmentação e etiquetagem, assim como o desenvolvimento de ferramentas que facilitem esse trabalho, é uma das tarefas do projecto BDFALA que tem vindo a ser realizada

sobre o corpus de passagens do Projecto SAM_A e sobre alguns excertos de entrevistas e debates televisivos. Para cada nível de anotação considerado é criado um ficheiro independente com o mesmo nome do ficheiro de fala correspondente mas com diferentes extensões.

3.1 Transcrição ortográfica

O primeiro passo consiste, naturalmente, na transcrição ortográfica dos materiais de fala recolhidos. A transcrição ortográfica é normativa, não sendo admitidos recursos ortográficos para marcar variações de pronúncia. A transcrição ortográfica em (9), por exemplo, é válida para qualquer das realizações em (10)

(9) Vou para Lisboa

(10) a. [v"o p6r6 liZb"o6]
b. [v"o pr6 liZb"o6]
c. [v"o p6 liZb"o6]

Para ter acesso a todas as realizações fonéticas da palavra *para* que se podem observar nos diferentes corpora é fundamental poder pesquisar todas as suas ocorrências de forma não ambígua. As indicações sobre o modo como foi efectivamente pronunciada podem então ser facilmente obtidas conjugando a transcrição ortográfica com as transcrições fonéticas correspondentes.

Relativamente à maior parte dos materiais lidos, esta tarefa consiste apenas a verificação de que os materiais de fala recolhidos coincidem com os textos fornecidos para leitura. Apesar do controle que é realizado pelo locutor durante a própria gravação, há inúmeros casos em que os dois tipos de materiais não coincidem. Nos diferentes excertos de fala contínua (textos, passagens, frases, etc.) há palavras que são substituídas sem que o locutor de aperceba: em vez de *Vou para o Porto amanhã* (texto inicial) pode muito bem ser lido *Vou ao Porto amanhã* ou, em vez do número *856934* podem ser lidos números como *865934* ou *856943*.

Em certo tipo de materiais, como por exemplo, o das recolhas telefónicas feitas para o SPEECHDAT ou os de fala espontânea, é necessário dar conta da ocorrência de determinados eventos: falsas partidas, truncações, hesitações, pausas, tosse, risos, sobreposições,

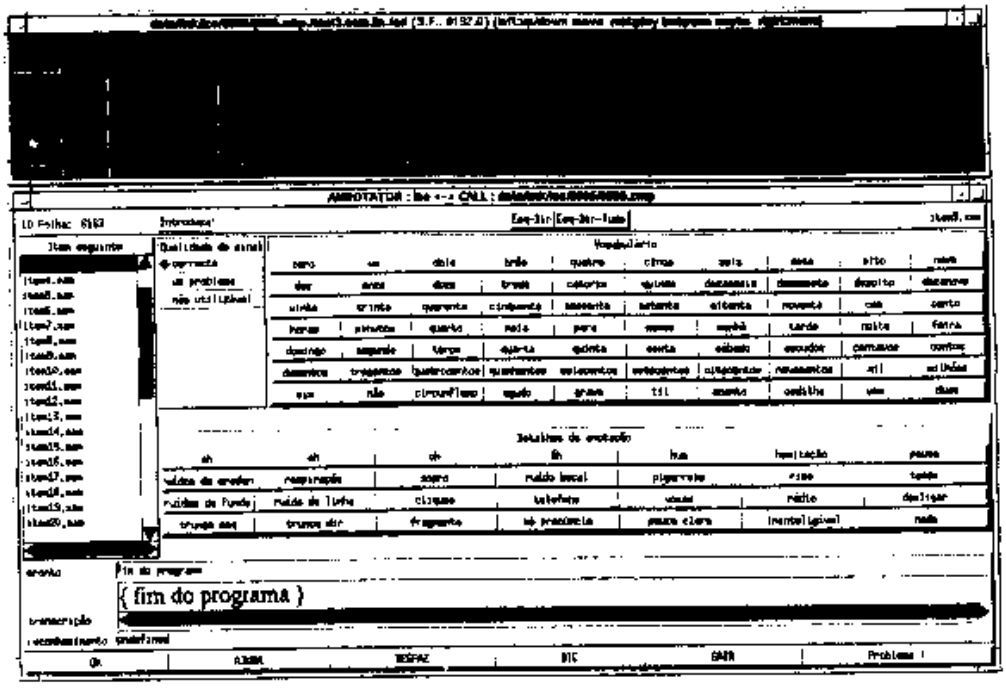
ruídos de fundo etc.. Estes eventos são marcados na transcrição ortográfica imediatamente antes da primeira palavra que é afectada e entre parêntesis rectos, de forma a poderem ser ignorados, sempre que tal se verifique necessário. Excepto para as palavras soletradas ou siglas, cuja transcrição é feita em maiúsculas, a transcrição ortográfica contém apenas minúsculas e não utiliza sinais de pontuação. Em (11) apresenta-se um exemplo de transcrição ortográfica de uma frase do corpus SPEECHDAT.

(11) o engenho descrito pelas autoridades como de grande potência foi lançado por desconhecidos de uma [hesitação] viatura circulando [ruído de fundo] a alta velocidade

Sempre que um corpus é segmental e prosodicamente anotado (EUROM.1 e BDFALA) a maior parte destas informações são retiradas da transcrição ortográfica e alinhadas com o sinal de fala num ficheiro independente (*ficheiro.mis)*, em que são indicados os instantes de princípio e fim de cada evento considerado.

Como meio auxiliar para a tarefa de transcrição ortográfica foi utilizado um outro pacote de programas que não foi construído de raiz mas inclui modificações bastante relevantes dos pacotes desenhados por investigadores do INDIAP (Suíça) e VOCALIS (Reino Unido). Como a figura (12) ilustra, o terminal são apresentadas duas janelas

(12).



principais: (a) uma janela superior onde aparece a forma de onda do sinal com marcadores temporais e (b) uma janela de anotação propriamente dita que inclui vários campos com funções especiais.

3.2 Etiquetagens fonéticas

Os textos ortográficos são em seguida foneticamente transcritos. Esta transcrição é feita automaticamente com Ler_PE [18], uma aplicação desenvolvida a partir do módulo fonético do sistema de síntese de fala para o Português Europeu: DIXI [12], assegurando-se uma transcrição fonética larga, ao nível da forma de citação, para todos os materiais de fala recolhidos. Esta transcrição, tradicionalmente designada por fonémica, é geralmente reconhecida como fundamental para o desenvolvimento de aplicações que envolvam síntese ou reconhecimento de fala. Trata-se de um nível intermédio entre o texto escrito e o sinal de fala que facilita o acesso ao léxico e que, além disso, constitui uma fonte de informação essencial para o estabelecimento de regras fonológicas (cf. [1], por exemplo). Para o treino e teste de sistemas de processamento de fala natural, é necessário, no entanto, que os diferentes níveis de anotação se encontrem temporalmente alinhados com o sinal de fala. Parte deste trabalho, que é extremamente moroso e sujeito a um grande número de erros de escrita, tem vindo a ser realizado para algumas línguas de modo semi-automático, com base em procedimentos de reconhecimento de fala. Numa fase inicial do projecto BDFALA, foram derivados modelos para os segmentos fonéticos do Português Europeu idênticos ou muito semelhantes aos do inglês, a partir do corpus TIMIT para o dialecto de Nova Iorque.

Para os restantes, foi utilizado um subconjunto das palavras e logátomos do corpus EUROM.1 que foram segmentados e etiquetados manualmente. O treino foi feito utilizando modelos H.M.M (do inglês *Hidden Markov Models*) restimados com o algoritmo Baum-Welsh com 3 estados e 3 misturas por estado. Cada vector tinha 26 coeficientes (12 cepstrum, 12 delta-cepstrum, energia e delta-energia). Os modelos iniciais foram utilizados para alinhar as transcrições fonéticas produzidas por Ler_PE com os materiais de fala correspondentes, utilizando o algoritmo de Viterbi. Para esse efeito, foram usadas as passagens do corpus EUROM.1 lidas pelos 5 homens e 5 mulheres que

constituem o grupo dos *Poucos* (ao todo, 150 passagens de 5 frases cada uma). Este subcorpus, automaticamente alinhado, foi então utilizado para retreinar 45 modelos-fonéticos (1 por cada fone contemplado nessas transcrições) e realinhar as transcrições, com base num algoritmo de *boot-strap* iterativo. Para além destes modelos, foram ainda treinados modelos de trifones. O treino e o teste foram feitos utilizando o pacote HTK (do inglês *'HMM Tool Kit*) comercializado pela Entropics.

Como é evidente, há problemas com o mapeamento das transcrições fonéticas largas produzidas por Ler_PE com o sinal de fala, uma vez que as produções dos locutores raramente coincidem com as das formas de citação: há segmentos fonéticos que muitas vezes não são realizados e outros que podem ser inseridos. Também não é evidente o mapeamento das oclusivas fricatizadas que não apresentam um intervalo de silêncio seguido de uma explosão. Embora Ler_PE também seja capaz de produzir transcrições fonéticas estreitas, não nos pareceu adequado utilizá-las como entrada para o sistema de alinhamento. Considerámos, antes, que a melhor maneira de evitar erros de alinhamento seria deixar o sistema decidir entre várias transcrições alternativas, escolhendo caso a caso a mais adequada. As transcrições fonéticas são, assim, modificadas automaticamente para fornecer as alternativas mais comuns. Uma palavra como *'abade'*, por exemplo, cuja transcrição larga realizada por Ler_PE é a apresentada em (13.a), é transformada em (13.b)

(13) a. 6b"ad@
b. 6{b,B}"a{d,D}{@, NULL}

em que as pronúncias alternativas são apresentadas entre chavetas, separadas por vírgulas, B e D representam realizações [+contínuas] de /b/ e /d/, respectivamente e 'NULL' indica que a vogal final pode ser suprimida.

O resultado do mapeamento é guardado num ficheiro independente (*ficheiro*.lab), em que são indicadas as fronteiras entre segmentos fonéticos consecutivos e fornecidas apenas as alternativas de transcrição que o sistema considerou mais prováveis. Estes ficheiros têm a estrutura em (14), onde as 9 primeiras linhas correspondem ao cabeça-

(14)
```
signal
type 0
color 121
comment
font -misc-*-bold-*-*-*-15-*-*-*-*-*-*-*
separator ;
nfields 1
#
0.2 121 p
0.23 121 u
0.27 121 d0
0.3 121 d
0.33 121 i
0.39 121 A
0.42 121 d0
0.45 121 d
0.48 121 a
0.57 121 r
0.64 121 m
0.68 121 j
0.71 121 u
0.74 121 m
0.78 121 A
0.81 121 l
0.89 121 i
0.95 121 S
1.03 121 t0
1.06 121 t
1.1 121 6
```

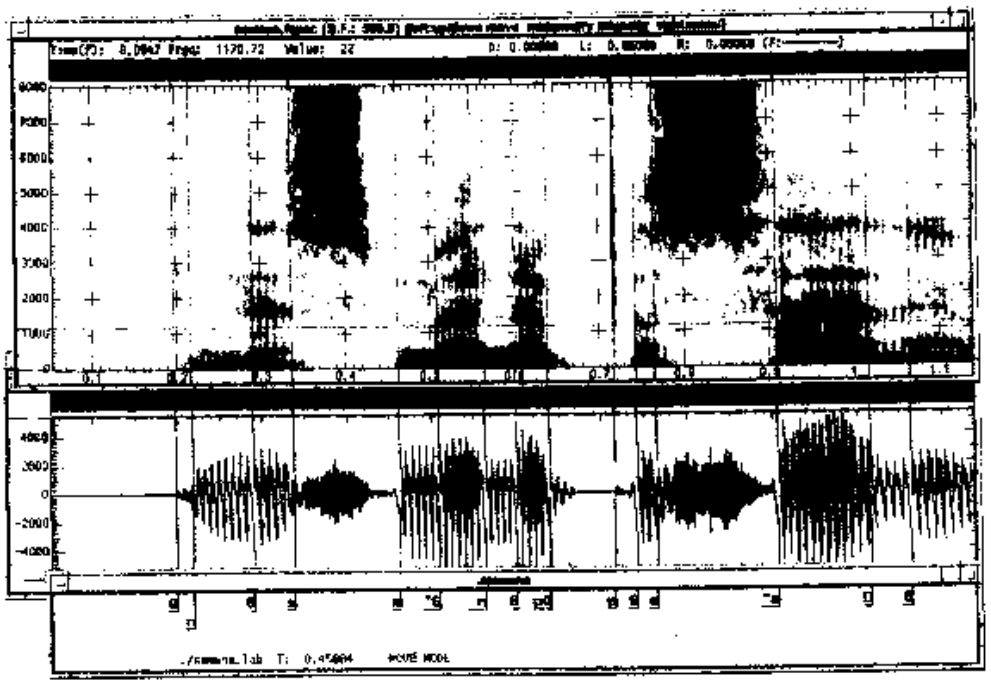
lho e as seguintes apresentam duas colunas. A primeira coluna contém a marca temporal do início do segmento fonético e a segunda a etiqueta simbólica correspondente. Como não existem por enquanto materiais de fala suficientes para treinar adequadamente o sistema de reconhecimento, os erros são ainda bastante frequentes e o trabalho manual de verificação e correcção de erros de segmentação e etiquetagem é ainda considerável. Para esse efeito, tem vindo a ser utilizado o pacote WAVES comercializado pela Entropics, em conjunto com alguns programas adicionais.

A segmentação e etiquetagem são corrigidas sobre a onda sonora com base na audição de excertos do sinal de fala e na observação de

espectrogramas. Trata-se, na maior parte dos casos, de deslocar ligeiramente algumas marcas de fronteira e de alterar as etiquetas em que o segmento automaticamente identificado não coincide com o segmento que é percebido pelo transcritor. As fronteiras são marcadas sobre a onda sonora, sempre que possível numa passagem ascendente por zero, de acordo com os modelos de transição propostos em Allen et allii (1987, p. 117). Embora, em geral, sejam tidas em consideração as transições de F2 e F3, é F2 que é tomado como referência sempre que se observa um desfasamento temporal nas transições destes dois formantes.

No final destas diferentes etapas, a transcrição fonética obtida é estreita e temporalmente alinhada com o sinal de fala, como (15) ilus-

(15)



tra. Os ficheiros de segmentação e etiquetagem manualmente corrigidos são utilizados em seguida como base para o alinhamento automático da transcrição fonética larga, do próprio texto ortográfico ou de outros níveis de etiquetagem como o morfo-sintáctico, por exemplo.

Este nível de etiquetagem levanta alguns problemas, uma vez que na fala contínua não existem, como é bem sabido, fronteiras nítidas entre segmentos consecutivos. Os órgãos articulatórios estão em constante movimento e os gestos ou conjuntos de gestos conducentes à produção de uma determinado segmento podem ser co-produzidos com

outros que estão relacionados com a realização de segmentos adjacentes quer anteriores quer seguintes. Para evitar a colocação de fronteiras artificiais entre segmentos acústicos consecutivos, no âmbito do projecto SAM foram feitas para algumas línguas (ex. Italiano) experiências de alinhamento dos símbolos fonéticos com os centros das zonas acusticamente mais estáveis. Os sistemas de reconhecimento treinados com base em corpora segmentados e etiquetados deste modo apresentaram, no entanto, desempenhos inferiores aos tradicionais, em que as marcas de fronteira são colocadas nas zonas de transição (Daalsgard, comunicação pessoal). Uma explicação possível para este facto é a das zonas acusticamente mais estáveis não serem também fáceis de delimitar, prestando-se a uma maior inconsistência na colocação das marcas temporais que asseguram o alinhamento das categorias simbólicas com o sinal de fala.

Optou-se, assim, por marcar as fronteiras nas zonas de transição e, como a colocação de marcas de fronteira entre segmentos em determinados instantes precisos é arbitrária, foi elaborado um conjunto de normas a seguir pelos anotadores. A maioria destas normas são as habitualmente seguidas nos trabalhos de fonética experimental. São demasiado extensas para as expormos aqui, mas encontrar-se-ão brevemente disponíveis para consulta.

Para obviar a alguns dos inconvenientes deste tipo de mapeamento, é possível complementá-lo, introduzindo marcas temporais mais finas que descrevam com maior detalhe o desfasamento temporal de diferentes eventos articulatórios a que estão associadas propriedades acústicas precisas. Na realização de uma vogal nasal, por exemplo, a presença de nasalidade pode ser detectável quer antes do início da vogal quer depois, podendo prolongar-se também para além do fim da mesma. De igual modo, os índícios de presença / ausência de vozeamento também não coincidem com as fronteiras dos segmentos vozeados e não vozeados, respectivamente, havendo múltiplos casos de harmonização de vozeamento em segmentos consecutivos. Para uma descrição destes fenómenos, é conveniente uma segmentação mais fina do sinal acústico, devendo as diferentes propriedades ser consideradas em fiadas de representação independentes que permitam representar as sobreposições e os desfasamentos. Este nível de etiquetagem, embora utilizado sobre materiais parcelares para efeitos de investigação fundamental no âmbito de outros projectos, não é contemplado em nenhum

dos corpora. De facto, os níveis de etiquetagem destinam-se prioritariamente a permitir localizar facilmente na base de dados exemplos de determinados eventos ou conjuntos de eventos e estes, depois, poderão ou não vir a ser analisados com maior detalhe de acordo com as necessidades.

3.3 Transcrição prosódica

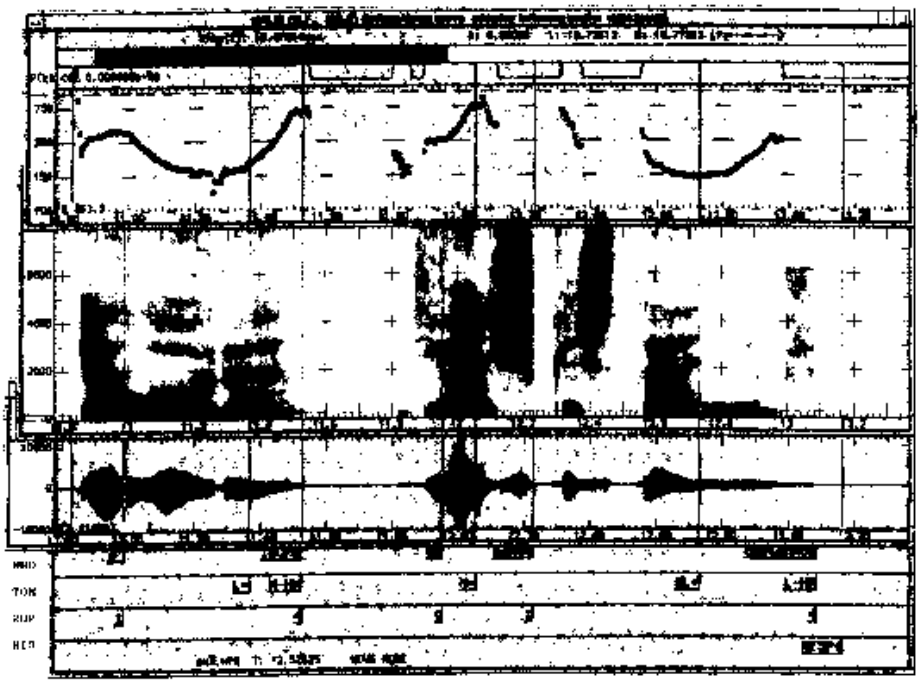
A inclusão de informações de ordem prosódica é também essencial para o treino de sistemas de processamento de fala natural por diferentes razões. Por um lado, porque a variação observada nas realizações fonéticas dos segmentos depende, em larga medida, do contexto prosódico em que estes ocorrem: a fala é hierarquicamente organizada em constituintes de diferentes dimensões e a forma como os segmentos são coarticulados pode ser explicada, em grande parte, em função da posição que estes ocupam nesses constituintes. Os coeficientes de reforço ou de redução podem ser automaticamente calculados se a transcrição fonética linear canónica for complementada com a informação prosódica necessária (cf. [6] e [11], entre outros). Por outro lado, as informações de ordem prosódica são também fundamentais para a compreensão da fala: sequências segmentais idênticas podem ser sintáctica e semanticamente interpretadas de maneira diferente, dependendo da forma como os enunciados são prosodicamente segmentados em constituintes e do tipo de proeminências que são associadas a esses constituintes. Para que os sistemas de processamento de fala natural possam extrair estas informações a partir do sinal acústico e utilizá-las para pôr hipóteses sobre as intenções dos falantes é necessário que estes possam ser treinados sobre corpora prosodicamente etiquetados ou que as relações entre os marcadores prosódicos utilizados e as intenções dos falantes sejam bem conhecidas e possam ser explicitadas em termos de regras. Note-se, no entanto, que o estabelecimento deste tipo de regras implica também, por sua vez, uma análise linguística cuidadosa de grandes corpora de fala prosodicamente anotados com base na audição e na observação do sinal acústico.

Como muitos aspectos da prosódia do P. E. ainda não foram suficientemente estudados, tem vindo a ser seguida uma dupla abordagem na segmentação e etiquetagem prosódica dos corpora. Parte dos materiais foram segmentados e etiquetados de acordo com o sistema de

anotação ToBI [2] [16]) que, embora desenhado fundamentalmente para o inglês, têm vindo a ser utilizado com sucesso para descrever aspectos da prosódia de outras línguas como o japonês [3] [13], o Italiano [9] ou o Espanhol [14]. A figura em (16) é um exemplo de etiquetagem prosódica para o P.E. utilizando este sistema.

A segmentação em palavras é feita automaticamente a partir dos ficheiros de segmentação e etiquetagem segmental e todas as outras fiadas da representação prosódica são inseridas manualmente. Como forma de procurar ultrapassar algumas das dificuldades encontradas, os corpora já etiquetados a nível segmental têm vindo a ser automaticamente processados para detecção dos acentos tonais e para a pesquisa de correlatos acústicos dos índices de ruptura utilizados [21][4] [5]. Para esse efeito, foram inseridas outras informações de ordem prosódica que se consideram relevantes, nomeadamente no que diz respeito à estrutura silábica e a graus de acento.

(16)



Os agrupamentos das sílabas em pés, os graus de acento, a localização das fronteiras de sílaba e a descrição do modo como os segmentos se encontram organizados no interior das sílabas (ataque ou rima e, na rima, núcleo ou coda), por exemplo, são integralmente automáticas.

O facto de se poder automatizar integralmente um certo número de análises não significa que estas devam ser consideradas adequadas. Para além da economia de recursos que representa, a principal vanta-

gem da automatização de certos níveis de etiquetagem é a de permitir testar a adequação dos modelos linguísticos por confrontação com os dados e, se os resultados não forem satisfatórios, poder alterá-los sem que isso obrigue a refazer todo o trabalho de anotação, de cada vez que se pretende testar novos modelos.

Com o estudo acústico actualmente em curso sobre o fraseamento prosódico e os acentos tonais, pretende-se, justamente, reunir um conjunto de conhecimentos sobre a função dos diferentes marcadores prosódicos, para poder automatizar também a etiquetagem acima do nível da palavra.

4. Considerações e perspectivas futuras

Depois de uma breve introdução em que se refere a importância crescente de corpora de fala de grandes dimensões para a investigação actual em fala, foi sucintamente descrito o trabalho que tem vindo a ser realizado nesta área no âmbito do convénio entre o CLUL e o INESC. Como nem todas as equipas de trabalho utilizam os mesmos suportes informáticos e os mesmos programas de análise acústica, foi também desenvolvido um pacote de programas, não mencionado acima, que permite a conversão tanto dos ficheiros de fala, como dos ficheiros de segmentação e etiquetagem, para os formatos requeridos por esses programas.

O primeiro dos corpora recolhido foi o corpus EUROM.1 no âmbito do Projecto Europeu SAM_A. Este corpus serviu como embrião para as recolhas actualmente em curso no âmbito do projecto Nacional BDFALA. A adopção de formatos normalizados e de critérios de segmentação e etiquetagem comuns permite-nos ir alargando progressivamente cada um dos subcorpora definidos neste projecto, com o objectivo a médio prazo de se constituírem individualmente como corpora de dimensão comparável aos dos seus congéneres a nível mundial. Só este alargamento nos permitirá progredir na investigação em curso e estar a par com as recentes abordagens que, dada a sua importância crescente, tendem já a individualizar-se com o nome de "corpus-based methods".

Referências

[1] BARRY, W. J.. & J. FOURCIN (1992) – "Levels of labelling". *Computer Speech and Language*, 6:1-14.

[2] BECKMAN, M. & G. M. AYERS (1994) – *Guidelines for ToBI labeling*. Columbus, Ohio State University.

[3] BECKMAN, M. E. & J. B. PIERREHUMBERT (1986) – *Intonation structure in English and Japanese*. Phonology Yearbook, 3: 255-309.

[4] CAMPBELL, W. N. (1992) "Syllable-based segmental duration"., in C. Benoit & Savallis (Org.s)- *Talking Machines – Theories, Models and Design*, pp. 211-244

[5] CAMPBELL, W. N. (1993) – "Automatic detection of prosodic boundaries in speech". *Speech Communication*, 13: 343-354.

[6] COLEMAN, J. (1992) – "The phonetic interpretation of headed phonological structures containing overlapping constituents". *Phonology*, 9: 1-44.

[7] CORMEN, T. H., C. E. LEISERSON & R. L. RIVEST (1990) – *Introduction to Algorithms*. The M.I.T. Press, Cambridge, Mass.

[8] FALÉ, I. S. (1996) – "Fragmento da Prosódia do Português Europeu: as Estruturas Coordenadas". Dissertação de Mestrado, Faculdade de Letras da Universidade de Lisboa.

[9] GRICE, M. (1992) – "The intonation of interrogation in Palermo Italian: implications for intonation theory". Diss. PhD, University College, Londres (Publicada 1995, Tübingen, Niemeyer).

[10] GROSZ, B. & J. HIRSCHBERG (1992) – " Some intonational characteristics of discourse structure". *Proceedings ICSLP 92 – 1992 International Conference on Spoken Language Processing*, Banf, Alberta, Canada, pp. 429-432.

[11] KOHLER, K. J. (1994) – "Complementary Phonology. A theoretical frame for labelling an acoustic data base of dialogues". *Proceedings ICSLP 94 – 1994 International Conference on Spoken Language Processing*, Yokohama, Japão, pp. 427-430.

[12] OLIVEIRA, Luís Caldas, M. Céu VIANA & Isabel M. TRANCOSO (1992) – "A rule-based text-to-speech system for Portuguese". *Proceedings of the 1992 International Conference on Acoustics, Speech and Signal Processing*, San Francisco, Março, pp. 73-76.

[13] PIERREHUMBERT, J. & M. BECKMAN (1988) – *Japanese Tone Structure*. The M.I.T. Press, Cambridge, Mass.

[14] PRIETO, P., J. van Santen & J. Hirschberg (1994) – "Patterns of F0 peak placement in Mexican Spanish. *Proceedings of the ESCA Workshop on Speech Synthesis*, New Paltz, Nova Iorque, pp. 33-37.

[15] RIBEIRO, C., I. M. TRANCOSO & M. Céu VIANA (1993) – ESPRIT 6819 SAM_A – Speech Technology Assessment in Multilingual Applications, Relatório D6 – EUROM.1 Portuguese Database.

[16] SILVERMAN, Kim, M. BECKMAN, J. PITRELLI, M. OSTENDORF, COLIN Wightman, P. PRICE, J. PIERREHUMBERT & J. HIRSCHBERG(1992) – "ToBI: a standard for labeling English Prosody". *Proceedings ICSLP 92 – 1992 International Conference on Spoken Language Processing*, Banf, Alberta, Canada, pp. 867-870.

[17] VAN SANTEN, J.P.H. (1992) – Diagnostic perceptual experiments for text-to-speech system evaluation. *Proceedings ICSLP 92 – 1992 International Conference on Spoken Language Processing*, Banf, Alberta, Canada, pp. 555-558.

[18] VIANA, M. Céu, Ernesto D'ANDRADE, Luís C. Oliveira e Isabel M. Trancoso (1991) – "Ler_PE: um utensílio para o estudo da ortografia do português", *Actas do VII Encontro da Associação Portuguesa de Linguística*, Lisboa, pp. 474-489.

[19] VIANA, M. Céu e E. D'ANDRADE (1992) – "Uma questão de equilíbrio". *Actas do VIII Encontro da Associação Portuguesa de Linguística*, Lisboa, pp.

[20] VIGÁRIO, M. (1995) – "Aspectos da Prosódia do Português Europeu: Estruturas com Advérbios de Exclusão e Negação Frásica". Dissertação de Mestrado, Faculdade de Letras da Universidade de Lisboa.

[21] WIGHTMAN, C. W. & M. OSTENDORF (1994) – "Automatic labeling of prosodic patterns". *IEEE Transactions on Speech and Audio Processing*, 2(4): 469-480.

[22] WINSKI, R. F. SENIA, P. CONNER, R. HAB-UMBACH, A. CONSTANTINESCU, G. NIEDERMAIR, A. MORENO e I. TRANCOSO. (1995) LRE – 63314 Speechdat – Relatório D1.4-1 – Specification of Telephone Data Collection.