ALINHAMENTO DE LIVROS FALADOS

António Serralheiro*, Hugo Meinedo**, Diamantino Caseiro*, Isabel Trancoso*

* INESC-ID/IST, ** INESC-ID

http://www.l2f.inesc-id.pt

0. RESUMO

O presente trabalho foi desenvolvido no âmbito do Projecto IPSOM - Indexação, Integração e Pesquisa em Documentos Multimédia, financiado pela FCT e em que participam as seguintes entidades: Laboratório de Sistemas de Língua Falada do INESC ID Lisboa (L2F), Grupo Multimédia da Faculdade de Ciências da Universidade de Lisboa (LASIGE) e Biblioteca Nacional (BN). O objectivo deste projecto é melhorar o acesso aos livros falados pela comunidade invisual, através de ferramentas que neles possibilitem a fácil detecção e indexação de unidades (palavras, frases, tópicos). Neste trabalho descreve-se o processo de alinhamento de livros falados, desde a selecção do corpus piloto até aos resultados (ainda que preliminares) obtidos pelo alinhamento automático. Apresentam-se ainda conclusões e trabalho a realizar futuramente.

1. INTRODUÇÃO

Este artigo pretende ser uma introdução ao projecto IPSOM - Indexação, Integração e Pesquisa em Documentos Multimédia. Trata-se de um projecto nacional, subsidiado pela Fundação para a Ciência e Tecnologia (FCT), iniciado em Novembro de 2000 e com uma duração de 3 anos. O consórcio envolve membros do Laboratório de Sistemas de Língua Falada (L^2F) do INESC ID Lisboa, do grupo Multimédia (LASIGE) da Faculdade de Ciências da Universidade de Lisboa e da Biblioteca Nacional (BN) que actua como fornecedor dos recursos linguísticos tratados no projecto – livros falados. Em Portugal, os livros falados têm vindo a ser usados principalmente pela comunidade de pessoas portadoras de deficiência visual. Na sua maioria, estes livros são gravados por voluntários em cassetes analógicas que são, a pedido, distribuídas aos interessados pela BN. O formato actualmente em uso nestes livros, torna a pesquisa de informação muito morosa e sujeita a erros, pelo que a primeira prioridade é a sua conversão para o formato digital.

O principal objectivo do projecto IPSOM é melhorar o acesso aos livros falados pela comunidade invisual, através de ferramentas que neles possibilitem a fácil detecção e indexação de unidades (palavras, frases, tópicos). Simultaneamente, pretende-se alargar a utilização de livros falados (p.ex., para aplicações didácticas), através de interfaces multimédia para acesso e pesquisa, obtendo-se assim um livro falado multimédia. Descreveremos neste trabalho, a escolha e obtenção do corpus piloto (escrito e falado), e o processo automático do alinhamento utilizado. As ferramentas de reconhecimento automático de fala desenvolvidas para minimizar a intervenção humana neste processo de alinhamento e que são baseadas em modelos híbridos serão também apresentadas. Incluem-se ainda resultados preliminares e as conclusões e linhas de trabalho futuro.

2. ALINHAMENTO DE LIVROS FALADOS

2.1 *Problemas do Alinhamento*

Do vasto acervo de livros falados actualmente existente na BN (cerca de 2.000 obras), foram inicialmente seleccionados dois textos «A Noite e a Madrugada» de Fernando Namora e «A Viúva do Enforcado» de Camilo Castelo Branco. Estes dois livros falados foram analisados e confrontados com as respectivas versões escritas por forma a se detectarem eventuais inconsistências. De facto, verificou-se, entre outros problemas, uma não sistematização na leitura de prefácio, da contracapa, dos nomes e números dos capítulos e secções (por exemplo, no mesmo livro, o informante tanto leu *primeiro capítulo* como leu *capítulo dois*), das notas de rodapé, etc. Outros problemas detectados foram desde a falta de trechos vocais (inclusivé de parágrafos inteiros), má qualidade sonora (consequência talvez de cópias sucessivas em suportes magnéticos de qualidade inferior), até uma equalização espectral não uniforme, em resultado de um processo de equalização manual anterior à gravação e que varia de sessão para sessão,

do tipo de microfone utilizado, etc. Estas deficiências[1] são, contudo, determinantes para o desempenho do sistema de alinhamento automático que pretendemos desenvolver.

2.2 *Metodologia*

2.2.1 *Aquisição do sinal de fala*

As condições dos livros falados estavam, portanto, longe das ideais pelo que foi gravado um novo livro em que tais problemas fossem eliminados ou, no pior dos casos, minimizados. Seleccionou-se um novo trecho uma vez que um dos exemplares anteriores (A Noite e a Madrugada) continha muitos vocábulos estrangeiros[2] e o outro não era tão actual como se pretendia. Assim, a escolha recaiu sobre «o Senhor Ventura» de Miguel Torga que constitui o nosso corpus piloto:

- dimensão do texto: 137.944 palavras;
- dimensão do léxico: 5.226 vocábulos, incluindo formas verbais;
- duração: 2 horas e 15 minutos.

A versão electrónica do *Senhor Ventura* foi manualmente editada por forma a retirar todos os trechos associados (notas editoriais várias, prefácio, etc.) bem como a numeração das páginas no intuito de facilitar o mais possível a leitura. Esta decorreu numa câmara insonorizada existente no INESC-ID ao longo de duas semanas em sessões que não ultrapassaram os 90 minutos no intuito de evitar ao máximo o cansaço do leitor. Para garantir uma maior uniformidade na qualidade sonora ao longo das diferentes sessões, foi utilizado um microfone facial (*headset microphone*) colocado a

---

1 Importa referir que, apesar dos problemas mencionados, a BN tem prestado um importantíssimo e insubstituível serviço à comunidade de portadores de deficiência visual. Este serviço tem sido possível graças à enorme boa-vontade e grande empenho pessoal de voluntários que, infelizmente, não têm formação técnica adequada.

2 Os livros falados servirão, entre outras aplicações, como uma valiosa fonte de informação para trabalhos de investigação em síntese de fala usando concatenação de entidades sub-palavra. Assim, nesta primeira fase, importa obter trechos com baixo número de vocábulos estrangeiros, uma vez que, para a sua leitura, não há regras perfeitamente estabelecidas.

curta distância dos lábios do orador. O sinal assim captado, foi gravado directamente e sem interrupções em suporte digital - DAT - com uma frequência de amostragem de 48kHz, para posterior transferência para computador. Esta transferência foi efectuada sempre em formato digital tendo sido obtida, por decimação, uma versão *raw*[3] de 16 bits e a 44.100 amostras/segundo para posterior gravação em formato CD-Áudio.

2.2.2 *Alinhamento automático: pré-processamento*

O alinhamento automático consiste, nesta aplicação, na localização das fronteiras de palavras ao longo do sinal de fala, conforme se resume na figura 1. Como se depreende, há dois processos paralelos: um diz respeito ao tratamento da informação escrita e o outro ao tratamento da informação sonora.

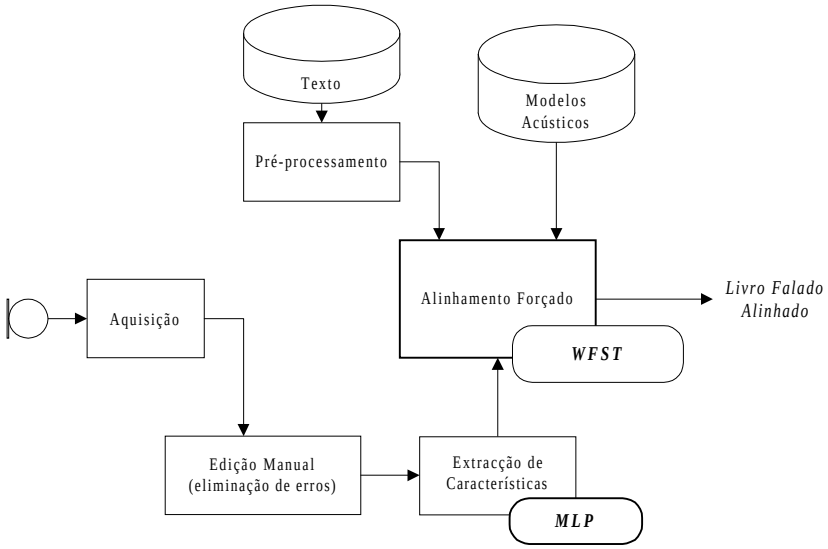

Figura 1 *- Processo de Alinhamento de Livros Falados.*

[3] Este formato consiste em digitalizar linearmente cada amostra em 16 bits que são armazenados em ficheiros em formato binário sem quaisquer cabeçalhos.

O processamento da informação escrita é efectuado de um modo automático e destina-se a obter uma versão do texto original sem quaisquer formatações, nomeadamente:

- ausência de pontuação;
- conversão de abreviaturas, ex: *Dr* = doutor, *Sr^a* = Senhora, etc.;
- conversão por extenso de numerais/ordinais, de datas, ex: *1* = primeiro, *21* = vinte e um, etc.;
- redução de caracteres maiúsculos a minúsculos.

Esta versão simplificada do texto lido é utilizada como entrada do processo de alinhamento forçado (figura 1). Obteve-se ainda, após o pré-processamento do texto, o correspondente léxico de pronúncia (SAMPA).

O sinal sonoro gravado inicialmente teve de ser editado manualmente de modo a se eliminarem todos os enganos e repetições ocorridas durante a leitura. É um tratamento moroso, estimando-se a sua duração entre 3 a 5 vezes tempo-real. Numa primeira fase, e devido a limitações do sistema de alinhamento automático, houve que partir as gravações de áudio em ficheiros de «curta» duração: cerca de 3 minutos. Esta partição, efectuada automaticamente, implicou ainda que houvesse uma sobreposição do sinal entre ficheiros consecutivos (15% no início e 15% no final, ou seja, 30% no total) para que o processo de edição manual fosse simplificado, resultando num total de 82 ficheiros de áudio. Em seguida, de cada um destes ficheiros foi obtida uma versão decimada ao ritmo de 16.000 amostras/segundo para efeitos de processamento com vista ao posterior alinhamento.

2.2.3 *Alinhamento Automático*

Extracção de Características

O modelo acústico usado no alinhamento foi originalmente desenvolvido para uma tarefa de ditado, tendo sido treinado com fala proveniente de um grande número de oradores de ambos os géneros que leram textos seleccionados de um jornal nacional diário [Neto, 1998]. O modelo acústico usa uma topologia onde as probabilidades *a posteriori* de fones independentes do contexto são estimadas a partir de 3 MLP (*multi-*

layer perceptrons, perceptrões multi-camada) operando simultaneamente sobre os dados acústicos. As sequências de probabilidades obtidas nas saídas dos 3 MLPs são posteriormente combinadas [Meinedo, 2000]. As três redes MLP usadas têm a mesma estrutura e são treinadas com métodos de extracção de parâmetros distintos: PLP e RASTA [Hermansky, 1992] e MSG [Kingsbury, 1998]. Os dois primeiros processos usam vectores de 26 coeficientes PLP e Log-RASTA ao passo que o último usa vectores de 28 coeficientes por cada segmento temporal. Cada MLP incorpora localmente informação contextual através de uma janela alargada a sete segmentos do sinal de fala: 3 segmentos à esquerda e outros 3 segmentos à direita do segmento central. A rede utilizada tem uma única camada escondida de 500 unidades com 39 saídas correspondentes a 38 classes de fonemas do Português europeu e ao silêncio.

Alinhador

Um alinhador não é mais que um descodificador que tem em conta as fronteiras temporais entre palavras ou fones. O nosso descodificador é baseado em transdutores ponderados de estados finitos (*Weighted Finite State Transducers*, WFST) no sentido em que o espaço de busca é definido por um transdutor de distribuições-para-palavra que é construído fora do descodificador. O espaço de busca é tipicamente construído como $H \circ L \circ G$, em que H representa a topologia do HMM ou fone, L representa o léxico e G o modelo de língua. No caso da tarefa de alinhamento, G é apenas a sequência de palavras que constitui a transcrição ortográfica da locução. A principal vantagem da utilização de WFSTs é a de não se colocar quaisquer restrições na construção do espaço de busca, o que significa que se pode facilmente integrar outras fontes de conhecimento, e a rede pode ser optimizada e substituída por uma rede óptima equivalente. Do ponto de vista de alinhamento, esta última vantagem é na realidade uma desvantagem, uma vez que essas optimizações não garantem que as etiquetas de entrada e saída estejam sincronizadas. De modo a resolver este problema, o descodificador foi preparado para lidar com etiquetas especiais, no lado da entrada, que são internamente tratadas como etiquetas *epsilon*, e que são usadas para marcar

transições ou fronteiras temporais. De cada vez que uma etiqueta deste tipo é atravessada, o instante de tempo correspondente é armazenado na hipótese corrente. Consoante o nível de alinhamento pretendido, estas etiquetas podem ser colocadas no final de cada WFST de fone ou no final de cada WFST de palavra.

Regras fonológicas

Em vez de construir um léxico com múltiplas pronúncias por palavra, o nosso objectivo é desenvolver regras fonológicas que possam ser usadas conjuntamente com um léxico contendo apenas as formas canónicas, de modo a ter em conta pronúncias alternativas. Estas regras são especificadas usando uma gramática de estados finitos cuja sintaxe é semelhante à forma Backus-Naur aumentada com expressões regulares. Cada regra é representada por uma expressão regular, e ao conjunto usual de operadores junta-se o operador $\rightarrow$, transdução simples, tal que $a \rightarrow b$ significa que o símbolo terminal *a* é transformado no símbolo terminal *b*. A linguagem permite a definição de símbolos não-terminais (por exemplo: *$vogal*). Todas as regras são opcionais e são compiladas para WFSTs. A figura 2 mostra um exemplo de especificação de uma regra. Essa especificação é primeiramente transformada num transdutor *T*, e posteriormente compilada para $R_T = \Sigma^* (T \Sigma^*)^*$ [4]. Esse transdutor, uma vez composto com o transdutor correspondente à pronúncia canónica *S*, produzirá $S_T = \pi_2 (S \circ R_T)$ que permite novas pronúncias alternativas:

```
$Vocálico = $Vogal | $VogalNasal | $Glide | $GlideNasal;

DEF_RULE SANDHI_S_z, ($Vocálico (S→ z) FRONTEIRA_PALAVRA $Vocálico)
```

Figura 2 - *Exemplo de uma regra especificada usando a* linguagem de especificação de regras.

[4] Σ é o transdutor identidade, que converte cada símbolo de entrada em si próprio.

As regras não são aplicadas uma-a-uma, em cascata de composições; em vez disso, aplica-se a sua união $R = R_{T_1} \cup R_{T_2} \cup \ldots R_{T_n}$. *R* é aplicada 3 vezes $S_R = \pi_2 (R \circ (R \circ (R \circ S)))$, de modo a permitir a aplicação de uma regra aos resultados da outra. Ao fazer a união das regras, evita-se o crescimento exagerado do transdutor resultante, o qual poderia ser exponencial com o comprimento da cascata de composições. Os principais aspectos fonológicos cobertos por estas regras são fenómenos de redução vocálica e de coarticulação entre palavras. Nas nossas experiências foram usadas 37 regras.

2.2.4 *Resultados experimentais*

O sistema de alinhamento inicialmente adoptado no projecto apresentava limitações de memória que impunham um limite máximo de 3 minutos para cada segmento a alinhar. Isso obrigava a uma partição do áudio e do texto correspondente, que requeria uma intervenção manual deveras morosa. A principal vantagem do alinhador baseado em WFSTs consiste em permitir o alinhamento de todo o áudio e texto numa só etapa. O alinhamento a nível de palavras de todo o livro adoptado como corpus piloto (2,25h) levou 197,5 segundos num computador Pentium III a 600MHz (0,024 vezes tempo-real), e precisou apenas de 200MB de RAM. O alinhamento a nível de fones correu a 0,027 vezes tempo real usando apenas as formas canónicas do léxico, e a 0,03 vezes tempo real usando também as regras de pronúncia alternativa.

Presentemente, o corpus piloto - *O Senhor Ventura* de Miguel Torga, encontra-se completamente alinhado, resultando em 2 classes de ficheiros: ficheiros *wave* e ficheiros de transcrição *lab*, em formato texto. Os primeiros contêm a gravação editada (PCM, 16 bits por amostra a 16.000 amostras/segundo) da leitura do corpus piloto e os segundos uma lista ordenada correspondendo às ocorrências (distância temporal contada a partir do início de cada ficheiro) de cada palavra. Deste modo, é possível a visualização e audição do livro falado em aplicações de domínio público (http://www.etca.fr/CTA/gip/Projets/Transcriber/), como se exemplifica na figura 3.

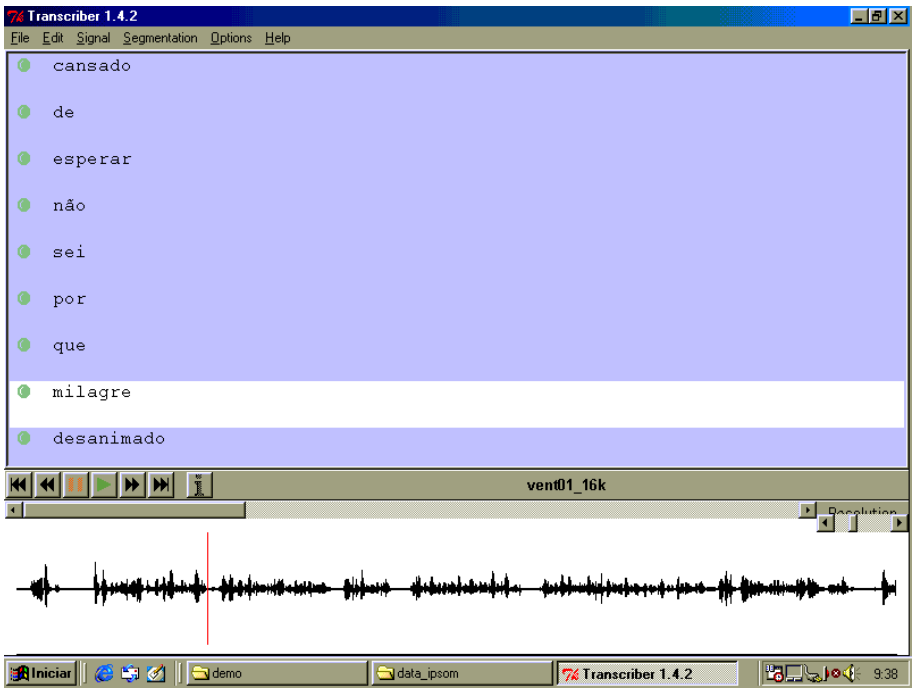

Figura 3 *- Exemplo de Alinhamento: O Senhor Ventura, 1ª Parte.*

O facto do corpus piloto não estar manualmente segmentado impede-nos de avaliar a precisão das fronteiras temporais detectadas automaticamente pelo alinhador quer a nível de fronteiras entre palavras, quer entre fones. O mesmo se pode dizer relativamente à avaliação da escolha entre pronúncias alternativas. Estes últimos aspectos são sobretudo importantes se pretendermos usar um corpus assim alinhado para investigação em síntese de fala. De modo a avaliar a qualidade do alinhador, este foi aplicado a um corpus manualmente segmentado e etiquetado de 15 frases ditas por 10 oradores (total de 150 frases), que constitui um subconjunto do corpus EUROM.1 [Ribeiro, 1993].

Os resultados incluídos na Tabela 1 mostram-nos que o alinhamento a nível de fones que se obtém usando as regras é superior ao que se obtém usando apenas as

formas canónicas, em termos de percentagem de fones correctamente detectados. O mesmo se pode dizer relativamente aos desvios temporais apresentados na Tabela 2.

Tabela 1 *- Alinhamento a nível de fones do subcorpus EUROM.1*

	Correcção	**Precisão**
canónica	93,65%	76,87%
regras	95,62%	79,74%

Tabela 2 *- Desvios temporais do alinhamento a nível de fones do subcorpus EUROM.1.*

	10ms	**90%**
canónica	37,4%	52 ms
regras	38,6%	44 ms

Nesta, a primeira coluna de resultados mostra a percentagem de desvios inferiores a 10 ms e a segunda coluna mostra o desvio máximo obtido para 90% dos segmentos. Foi também feita uma comparação dos WFSTs gerados pelas regras com as transcrições manuais, de modo a obter o desempenho de oráculo das regras (isto é, o desempenho de um descodificador perfeito, usando todos os caminhos possíveis nas redes de fones permitadas pelas regras). O valor obtido foi de 97,73% de correcção e de 82,11% de precisão. A maioria dos erros são devidos a apagamentos não permitidos nem pelo léxico, nem pelas regras.

Os resultados preliminares obtidos nas experiências efectuadas permitiram-nos apreciar as vantagens da utilização de regras de pronúncia alternativa, mas levam-nos também à conclusão de que o seu potencial ainda não foi explorado, pelo que há ainda muito a investigar nesta área. Há também que explorar a possibilidade de adaptação ao orador e de usar o sistema de reconhecimento para ajudar a detectar erros de leitura no livro falado. Os primeiros testes que realizámos neste sentido foram

deveras prometedores e planeamos apresentar resultados mais exaustivos no próximo relatório.

3. CONCLUSÕES E TRABALHO FUTURO

Nesta fase inicial do projecto IPSOM fez-se um levantamento preliminar do acervo de livros falados da BN, tendo-se registado alguns problemas que, a não serem eliminados iriam condicionar negativamente os objectivos propostos. Uma vez que seria necessário efectuar novas gravações, decidiu-se escolher um novo texto para corpus piloto, cuja leitura foi efectuada por uma locutora profissional. Os ficheiros de áudio obtidos destas gravações foram primeiramente sujeitos a um processo de edição manual em computador. Seguidamente, efectuou-se o alinhamento automático da gravação editada com o texto correspondente. Assim, resultou um novo de ficheiro (ficheiro de alinhamento) que contém as marcas temporais das ocorrências das palavras. Poderemos resumir assim, após este primeiro ano de actividades do projecto, os seguintes resultados:

- obtenção de um primeiro livro falado;
- automatização parcial do alinhamento do corpus piloto;
- experiência adquirida com a aquisição do corpus piloto importante para recomendações de futuras gravações.

Futuramente, na continuação do projecto, desenvolveremos as seguintes actividades:

- automatizar ao máximo o processo de alinhamento do corpus;
- gravar outros livros falados;
- desenvolver interfaces multimédia para livros;

Do ponto de vista de investigação, um dos aspectos mais interessantes dos livros falados é o facto de fornecerem um recurso linguístico importantíssimo para o modelamento prosódico baseado em corpora e síntese por concatenação, aspecto esse que planeamos explorar muito em breve.

4. REFERÊNCIAS

[Hermansky, 1992] Hermansky, H., Morgan, N., Baya, A., Kohn, P., *RASTA-PLP Speech Analysis Technique*, Proc. ICASSP92, São Francisco, EUA, 1992.

[Kingsbury, 1998] Kingsbury, B., Morgan, N., Greenberg, S., *Robust Speech Recognition using the Modulation Spectrogram*, Speech Communication, 25, 117 a 132, 1998.

[Meinedo, 2000] Meinedo, H., Neto, J., *Combination of Acoustic Models in Contiuous Speech Recognition Hybrid Systems*, in Proc. ICSLP 2000, Pequim, China, 2000.

[Mohri, 1999] Mohri, M., Riley, M., *Untegrated Context-dependent Networks in Very Large Vocabulary Speech Recognition*, in Proc. Eurospeech 99, Budapeste, Hungria, 1999.

[Neto, 1998] Neto, J., Martins, C., Almeida, L., *A Large Vocabulary Continuous Speech Recognition Hybrid system for the Portuguese Language*, in Proc. ICSLP 98, Sydney, Austrália, 1998.

[Ribeiro, 1993] Ribeiro, C. , Trancoso, I., Viana, M., EUROM.1 Portuguese Database}, Relatório do Projecto ESPRIT 6819 SAM-A, 1993.

AGRADECIMENTOS

Os autores desejam expressar os seus agradecimentos à Drª Isabel Bahia e ao Eng. João Lopes Raimundo pela amável colaboração na leitura do texto seleccionado e ainda ao Prof. João Paulo Neto pela cedência e apoio prestado na utilização do MLP-HMM.