Word Alignment in Digital Talking Books Using WFSTs

António Serralheiro, Diamantino Caseiro, Hugo Meinedo, Isabel Trancoso

L^2F Spoken Language Systems Lab.
INESC-ID/IST
Rua Alves Redol 9, 1000-029 Lisbon, Portugal

Abstract. This paper describes the motivation and the method that we used for aligning digital spoken books, and the results obtained both at a word level and at a phone level. This alignment will allow specific access interfaces for persons with special needs, and also tools for easily detecting and indexing units (words, sentences, topics) in the spoken books. The tool was implemented in a Weighted Finite State Transducer framework, which provides an efficient way to combine different types of knowledge sources, such as alternative pronunciation rules. With this tool, a 2-hour long spoken book was aligned in a single step in much less than real time.

1 Introduction

The framework of this paper is a national project known as IPSOM, whose main goal is to improve the access to digitally stored spoken books by the visually impaired community. Spoken books have been mainly provided by the National Library (BN, *Biblioteca Nacional*) in analogue format (cassette) and have lately been under a gradual conversion process to digital format (CDROM). To improve the usability of these spoken books, the IPSOM project aims to provide both specific access interfaces for persons with special needs, and also tools for easily detecting and indexing *units* (words, sentences, topics) either written or spoken. Therefore, a good word-by-word synchronization between the text and its audio recording is mandatory for unit access and thus spoken book alignment is a major task of the IPSOM project. This time alignment can be further complicated by the co-articulation and the vowel reduction problems that occur in natural speech. Therefore, different pronunciations of each word should be taken into account by using either an enlarged lexicon or phonological rules. We have chosen the latter approach, which was implemented in a *WFST* (Weighted Finite State Transducer) framework. *WFSTs* have been successfully used in many written and spoken language applications, providing an efficient and elegant way of combining different types of knowledge sources, which makes them good candidates for alignment purposes.

From the point of view of research in the area of speech processing, one of the most interesting aspects of the IPSOM project is the fact that indexed spoken

books provide an invaluable resource for data-driven prosodic modeling and unit selection in the context of text-to-speech synthesis. This is a good motivation to perform the alignment not only on the basis of words but also of sub-word units. Simultaneously, the project also aims to broaden the usage of multimedia spoken books (for instance in didactic applications, etc.), by providing multimedia interfaces for access and retrieval.

Throughout this paper, we preferred avoiding the standard designation of Digital Talking Books (*DTB*), as the spoken books available from BN do not yet have the associated text or navigation structure and, as such, can only be regarded as a simplified form of a type *1-DTB* [1] [2]. *DTBs* may provide the "talking" capability by means of a text-to-speech synthesizer, allowing a direct access to each text word within the book. Our automatic aligner easily provides this same word synchronized access for books read by a human voice, with all the naturalness and emotions that current synthesizers are still unable to convey, which causes them to invariably induce some fatigue to the listeners.

2 Pilot Corpus

Existing spoken books at BN have been recorded by volunteers (non-professional readers) and stored in analogue tapes, that by their sequential access mode, results in an extremely slow (and error-prone) information retrieval process. This handicap could be easily overcome through their conversion to CDROM, if other problems had not been found, namely: low audio quality (multiple copies and damaged masters), and audible differences of quality through the same book (manual spectral equalization, and uncalibrated multiple recording sessions). These problems, together with the non-systematic reading of tables, figures, chapter numbers, footnotes, preface, etc., made the current material not suitable for automatic text-to-speech alignment. Consequently, it was decided to record a new spoken book - "O Senhor Ventura" by Miguel Torga, to serve as a *pilot corpus* for the new recording and alignment procedure. This fiction book was read by a professional speaker in a sound-proof booth. It was recorded directly to DAT and later down-sampled to 16kHz. The digital audio file was then manually edited to remove some reading errors and extraneous noises (although breathing sounds were kept to enhance naturalness), resulting in 2h 15m of audio. The pilot corpus text, amounting to 137,944 words, was pre-processed to deal with abbreviations, numbers and special symbols, resulting in a lexicon with around 5k different forms. Although very intelligible, as expected from a professional speaker, the speaking rate was relatively high, averaging more than 174 words per minute. At this stage, we decided to make a plain text-to-speech alignment without dealing with the textual structure (punctuation marks, paragraphs, sections, chapters, etc.).

3 Alignment System

Although the purpose of our alignment is directed to spoken books, figure 1 shows a diagram of a generic alignment system without the navigation structure associated to *DTBs*. The feature extraction block maps the input samples of the audio signal into a lower-rate time sequence of acoustic parameters, as described in the next subsection. The forced-alignment block is further detailed in another subsection.

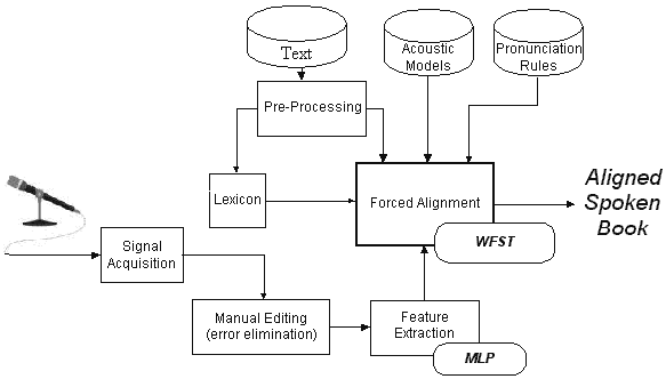

Fig. 1. Alignment System Diagram.

3.1 Acoustic Modeling

The hybrid acoustic models used in the alignment of spoken books were originally developed for a dictation task [3], in an effort to combine the temporal modeling capabilities of *HMMs* (Hidden Markov Models) with the pattern classification capabilities of *MLPs* (Multi-Layer Perceptrons). The models have a topology where context-independent phone posterior probabilities are estimated by three *MLPs* given the acoustic parameters at each frame. The streams of probabilities are then combined using an appropriate algorithm [4]. The processing stages are represented in Figure 2. The *MLPs* use the same basic structure and were trained with different feature extraction methods: *PLP* (Perceptual Linear Prediction) [5], *Log-RASTA* (log-RelAtive SpecTrAl) [5] and *MSG* (Modulation SpectroGram) [6]. *PLP* modeling is based on some specific human hearing characteristics, namely: non-linear frequency resolution, asymmetry of auditory filters, unequal hearing sensitivity at different frequencies and intensity-loudness non-linear relation. *RASTA* modeling attempts to model the sensitivity of human speech perception to preceding context and also to model the apparent insensitivity to absolute spectral shape. The *MSG* is a technique that models

the slow modulations in speech signals across time and frequency, emphasizing amplitude modulations in critical bands at rates of 0 to 8Hz. For the first two processes, the features are log-energy and *PLP/Log-RASTA* 12^{th} order coefficients and their first temporal derivatives summing up to 26 parameters. The *MSG* method uses 28 coefficients. Each *MLP* classifier incorporates local acoustic context via a multi-frame input window of 7 frames. The resulting network has a single hidden layer with 500 units and 39 output units (38 phones for European Portuguese plus silence).

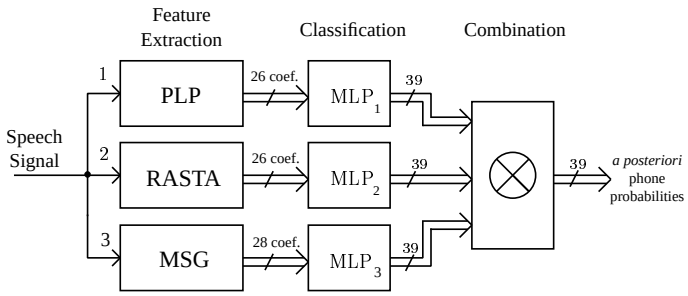

Fig. 2. Acoustic modeling combining several MLPs.

3.2 Alignment

An aligner is just a decoder that keeps track of the time boundaries between words or phones. Our decoder is based on *WFSTs* [7] in the sense that its search space is defined by a distribution-to-word transducer that is built outside the decoder. That search space is usually constructed as $H \circ L \circ G$, where H is the *HMM* or phone topology, L is the lexicon and G is the language model. For alignment, G is just the sequence of words that constitute the orthographic transcription of the utterance. The main advantage is that no restrictions are placed on the construction of the search space, which means that it can easily integrate other sources of knowledge, and the network can be optimized and replaced by an optimal equivalent one. This last advantage is a disadvantage from the perspective of alignment, as there are no warranties that the output and input labels are synchronized. To solve this problem, the decoder was extended to deal with special labels, on the input side, that are internally treated as epsilon labels, but are used to mark time transitions or boundaries. Whenever such end-of-segment labels are crossed, the time is stored in the current hypothesis. The user may choose to place those labels at the end of each phone *WFST* or at the end of each word *WFST*.

Phonological Rules Instead of building a lexicon with multiple pronunciations per word, our goal is to develop phonological rules that can be used with a lexicon of canonical forms, in order to account for alternative pronunciations. These rules are specified using a finite-state grammar whose syntax is similar to the Backus-Naur-form augmented with regular expressions. Each rule is represented by a regular expression, and to the usual set of operators we added the operator $\rightarrow$, simple transduction, such that $(a \rightarrow b)$ means that the terminal symbol a is transformed into the terminal symbol b. The language allows the definition of non-terminal symbols (e.g. *$vowel*). All rules are optional, and are compiled into *WFSTs*.

```
$Vocalic = $Vowel | $NasalVowel | $Glide | $NasalGlide;
DEF_RULE SANDHI_ch_z, ( $Vocalic (ch -> z) WORD_BREAK  $Vocalic)
```

Fig. 3. Example of a rule specified using the *rule specification language*.

Figure 3 presents an example of the specification of a rule; that specification is first transformed into a transducer T, and then compiled into $R_T = \Sigma^*(T\,\Sigma^*)^*$[1]. That transducer, when composed with the canonical phone transducer S will produce $S_T = \pi_2(S \circ R_T)$ that allows new pronunciation alternatives.

We do not apply the rules one by one on a cascade of compositions, but rather build their union $R = R_{T_1} \cup R_{T_2} \cup \cdots \cup R_{T_n}$. R is applied 3 times $(S_R = \pi_2(S \circ R \circ R \circ R))$, to allow the application of one rule to the results of another. By performing the union of the rules we avoid the exaggerated growth of the resulting transducer, which can be exponential with the length of the composition cascade.

The main phonological aspects that the rules are intended to cover are vowel reduction and word co-articula-tion phenomena. Vowel reduction is specially important for European Portuguese, being one of the features that distinguishes it from Brazilian Portuguese and that makes it more difficult to learn for a foreign speaker. In our experiments, we used 37 such rules.

4 Experimental Results

The tests described in this section involve both word level and phone level experiments. The pilot corpus allows us to do alignment tests at a word level, but not at a phone level, as required for text-to-speech research. In order to evaluate the quality of the phone level transcriptions obtained using the pronunciation rules, we used a fragment of the EUROM.1 corpus [8], for which we have manual phone level alignment. In addition, we also performed recognition experiments with spoken books.

[1] Σ is the identity transducer, that converts each input symbol into itself.

4.1 Alignment experiments with spoken books

A major advantage of our approach is that it allowed us to align the full audio version of the book in a single step. This is specially important if we take into account that the memory limitations of our previous alignment tool imposed a maximum of 3-minute audio segments. We thus avoid the very tedious task of manually breaking-up the audio into smaller segments with their associated text.

The word segmentation of the book ran in 0.024 real-time (RT), requiring 200MB of RAM. The phone level alignment of the book ran at 0.027 xRT when using the canonical pronunciations of the lexicon, and 0.030 xRT when using also the pronunciation rules.

An informal evaluation of the alignment procedure at word level was done using the publicly available Transcriber tool[2], which allowed us to subjectively access the good quality, by simultaneously listening and seeing on a word-by-word basis. Figure 4 illustrates the use of this tool.

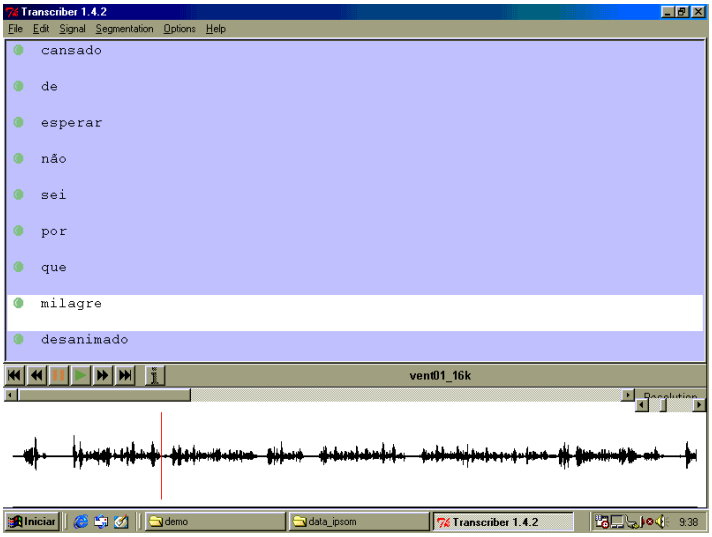

Fig. 4. Illustration of the word-level alignment of spoken books.

4.2 Recognition experiments with spoken books

The edition of recordings to remove reading errors and extraneous noises produced by the speaker is also a very labor intensive task. As a first step to automate this procedure, we tried to match text recognized using a dedicated

[2] http://www.etca.fr/CTA/gip/Projets/Transcriber/

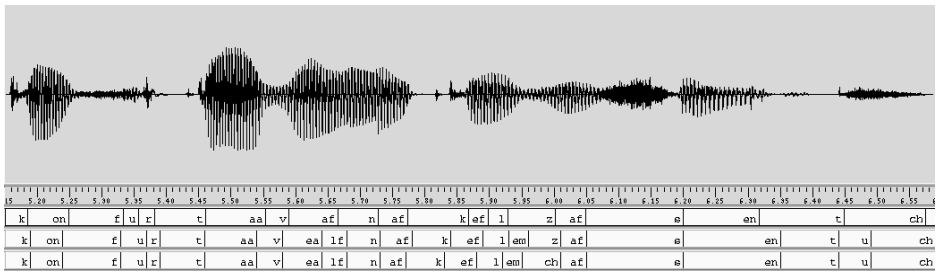

Fig. 5. Illustration of the phone level alignment of the EUROM.1 corpus.

recognizer with the original text in order to detect incorrect audio portions. The dedicated recognizer uses a lexicon and an n-gram language model estimated from all the book's text and achieved a word error rate (*WER*) of 17.2%.

Significant improvements were obtained by using spea-ker adaptation. That was done by retraining the acoustic models on 80% of the available audio, using the time stamps provided by the automatic aligner. The adapted models obtained with the alignment made using the canonical pronunciation lexicon achieved a *WER* of 7.8%, and the one using the phonological rules yielded 7.1%.

4.3 Phone level alignment evaluation

Our experiments with the EUROM.1 corpus showed us that the phone level alignment using the phonological rules is closer to the manual transcriptions than the canonical one (95.62% vs. 93.65% phone correction, respectively). The same conclusion was drawn when we analyzed the time deviation of the alignments: 38.6% (vs. 37.4%) of the deviations are less than 10ms and the maximum deviation obtained for 90% of the segments was 44ms (vs. 52ms).

We also compared the *WFSTs* generated by the rules with the manual transcriptions, in order to obtain the oracle performance of the rules (i.e. the performance of a perfect decoder, using all the possible paths in the phone lattices allowed by the rules): 97.73% correctness and 82.11% accuracy. Most of the errors are due to deletions observed in what the speakers said, that are neither allowed by the canonical lexicon nor by the rules. Figure 5 illustrates the phone level labels obtained with and without rules (middle and bottom layers), which can be compared with the manually assigned labels (top layer).

5 Conclusions and future work

The paper described our work on spoken books in the framework of the IPSOM project, emphasizing the problems of the actual repository and the alignment tools that were developed. We verified that, with proper recording procedures, the alignment task can be fully automated in a very fast single-step procedure, even for a 2-hour long recording. This is specially important if we take into

account that the memory limitations of our previous alignment tool imposed a maximum of 3-minute audio segments. With this new tool, we avoid the very tedious process of partitioning audio and text into corresponding segments. In addition, the use of a dedicated recognizer can also contribute to speeding up the manual process of removing reading errors.

The word boundaries computed using the *WFST*-based alignment will allow for the development of more sophisticated browsing tools for spoken books, which is one of our next tasks in the IPSOM project. Such browsing and indexing tools can be specially important for non-fiction, technical books, for which there is a great request from the visually impaired community.

The use of phonological rules seems to provide reasonably good alternative pronunciations, specially accounting for vowel reduction and inter-word coarticulation phenomena. However, a more exhaustive comparison with manual labeling still needs to be conducted in order to improve these rules. The better phone level alignment of spoken books achieved with these rules will also be crucial for our research in text-to-speech synthesis, namely for prosodic modeling and unit selection, using data-driven approaches.

6 Acknowledgments

This work was partially funded by FCT projects POSI/ 3452/ PLP/2000 and POSI/33846/PLP/2000. INESC-ID Lisboa had support from the POSI Program of the "Quadro Comunitário de Apoio III". The authors would like to thank Isabel Bahia and João Lopes Raimundo for their kind cooperation in reading the book and our colleagues from CLUL, Céu Viana and Isabel Mascarenhas, for their help with the phonological rules and manual labeling.

References

[1] ANSI/NISO Z39.86 - 2002 Specifications for the Digital Talking Book, *http://www.niso.org/standards/index.html*

[2] DAISY 2.02 Specification, Formal Recommendation, Feb. 28, 2001. *http://www.daisy.org/products/menupps.htm*

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

[4] H. Meinedo and J. Neto, "Combination of acoustic models in continuous speech recognition hybrid systems", In Proc. ICSLP 2000, Beijing, China, 2000.

[5] H. Hermansky, N. Morgan, A. Baya and P. Kohn, "RASTA-PLP Speech Analysis Technique", In Proc. ICASSP 92, San Francisco, USA, 1992.

[6] B. E. Kingsbury, N. Morgan, and S. Greenberg, "Robust speech recognition using the modulation spectrogram", Speech Communication, 25:117–132, 1998.

[7] M. Mohri, M. Riley, D. Hindle, A. Ljolje, F. Pereira, "Full Expansion of Context-Dependent Networks in Large Vocabulary Speech Recognition", In Proc. ICASSP 98, Seattle, Washington, 1998.

[8] C. Ribeiro, I. Trancoso and M. Viana, *EUROM.1 Portuguese Database*, Report of ESPRIT Project 6819 SAM-A, 1993.