SPOKEN BOOK ALIGNMENT USING WFSTS

Diamantino Caseiro, Hugo Meinedo, António Serralheiro, Isabel Trancoso, João Neto

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

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, used primarily by the visually impaired community, by providing tools for easily detecting and indexing units (words, sentences, topics). 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. Hence, spoken book alignment is a major task.

From the point of view of research, 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 modelling and unit selection in the context of text-to-speech synthesis. This motivated doing the alignment not only on the basis of words, but rather sub-word units and also automatically generating multiple pronunciations by applying phonological rules in a WFST (Weighted Finite State Transducer) framework.

2. ACOUSTIC MODELLING

The hybrid acoustic model used in the alignment of spoken books was originally developed for a dictation task [1]. The model uses a topology where context-independent phone posterior probabilities are estimated by three MLPs (Multi-Layer Perceptrons) given the acoustic data at each frame. The streams of probabilities are then combined using an appropriate algorithm [2]. The processing stages are represented in Figure 1. The MLPs use the same basic structure and are trained with different feature extraction methods: PLP [3], Log-RASTA [3] and MSG [4]. 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).

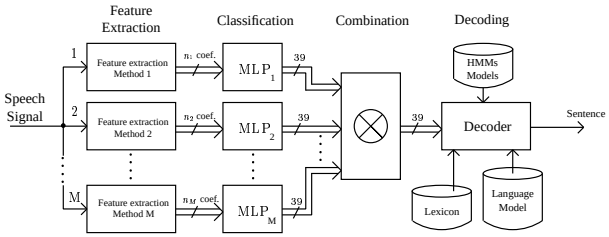

Fig. 1. *Acoustic model combining several MLPs trained on different feature sets.*

3. ALIGNMENT

An aligner is just a decoder, that keeps track of the time boundaries between words or phones. Our decoder is based on WFSTs [5] 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 optimised 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 synchronised. 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.

3.1. Phonological Rules

Instead of building a lexicon with multiple pronunciations per word, our goal is to develop phonological rules that can

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

Fig. 2. Example of a rule specified using the *rule specification language*.

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.

Figure 2 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 ($S_R = \pi_2(R \circ (R \circ (R \circ S))))$, 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 coarticulation 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

A small pilot corpus (*O Senhor Ventura*, by Miguel Torga) was chosen as a test bed for spoken book alignment. The high-quality DAT recordings were manually edited to remove reading errors and extraneous noises, amounting to a total of 2h15m (around 138k words, corresponding to 5k different forms). Although very intelligible, as expected from a professional speaker, the speaking rate was relatively high - 174 words per minute.

This spoken book allows us to do alignment tests at a word level, but not at a phone level, as required for TTS 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 [6], for which we have manual alignment.

[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 tedious task of manually breaking-up the audio into smaller segments with their associated text.

The word segmentation of the book took 197.5 seconds in a 600MHz Pentium III computer (0.024 xRT), and required 200MB of RAM. The phone level alignment of the book ran at 0.027 xRT when using the canonical pronunciations of the lexicon, and 0.03 xRT when using also the pronunciation rules.

4.2. Recognition experiments with spoken books

The edition of recordings to remove reading errors and extraneous noises produced by the speaker is a very labour intensive task. As a first step to automate this procedure, we tried to match text recognized using a dedicated recogniser with the original text in order to detect incorrect audio portions. The dedicated recogniser 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 the automatic alignment labels for speaker adaptation of the acoustic model, using only 80% of the available audio. The adapted model created using the alignment made with the canonical pronunciation lexicon achieved 7.8%, and the one using the phonological rules obtained 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 analysed 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.

5. CONCLUSIONS

The paper described our efforts towards aligning spoken books. We verified that the alignment task can be fully automated in a very fast single-step procedure, even for a 2-hour long recording. The use of phonological rules seems to provide reasonably good alternative pronunciations. However, a more exhaustive comparison with manual labelling still needs to be conducted in order to improve these rules.

6. REFERENCES

[1] 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.

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

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

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

[5] 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.

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