Automatic alignment of audiobooks in Afrikaans

Charl J. van Heerden
Multilingual Speech Technologies
North-West University
Vanderbijlpark, South Africa
Email: cvheerden@gmail.com

Febe de Wet[1,2]
[1]Human Language Technology
Competency Area
CSIR Meraka Institute
[2]Department of Electrical and
Electronic Engineering
Stellenbosch University, South Africa
Email: fdwet@csir.co.za

Marelie H. Davel
Multilingual Speech Technologies
North-West University
Vanderbijlpark, South Africa
Email: marelie.davel@gmail.com

***Abstract*—This paper reports on the automatic alignment of audiobooks in Afrikaans. An existing Afrikaans pronunciation dictionary and corpus of Afrikaans speech data are used to generate baseline acoustic models. The baseline system achieves an average duration independent overlap rate of 0.977 on the first three chapters of an audio version of *"Ruiter in die Nag"*, an Afrikaans book by Mikro. The average duration independent overlap rate increases to 0.990 when the speech data from the audiobook is used to perform Maximum A Posteriori adaptation on the baseline models. The corresponding value for models trained on the audiobook data is 0.996. An automatic measure of alignment accuracy is also introduced and compared to accuracies measured relative to a gold standard.**

I. Introduction

Audiobooks are available in many languages. Before the advent of the digital era, books were made available in analogue format. More recently new books are created in digital format and older books that were published on cassettes are gradually being converted to digital format.

Some digital formats facilitate audiobook access and navigation by people who have challenges using regular printed media. DAISY is an internationally established standard for creating digital audiobooks for use by print-disabled people [1]. DAISY books exist in a variety of formats. For some books, both the audio and text are available and the audio and text are aligned at word level. However, many DAISY books are published with limited alignment between audio and text (typically at the chapter level) or with no text at all.

Automatic speech recognition (ASR) technology can enhance audiobook publication in two ways. Firstly, for books that are published as audio only, ASR can be used to generate the text corresponding to existing audio. Secondly, ASR can be used to enhance the level of mark-up for books that are currently only aligned at chapter level. Finer grained alignments between audio and text enable word level search in audiobooks as well as synchronised reading, i.e. the text corresponding to the audio is highlighted during playback.

In this paper we will focus on using ASR technology to align large audio files at word level. The process will specifically be investigated for an under-resourced language for which, until fairly recently, only limited text and speech resources were available, namely Afrikaans. The ultimate aim of the work reported here is to improve the level of mark-up for existing books in any language by automatically converting the recognition output into DAISY *.smil* files. Section II provides some background on previous research on audiobook alignment. The pronunciation dictionary and acoustic data that were used during the study are described in Section III. Section IV describes the ASR systems that were used to perform alignment and Section V introduces a measure to verify alignment accuracy automatically. Results are presented in Section VI and conclusions in Section VII.

II. Background

Word and phone-level alignments between the audio and text versions of audiobooks are used either to enhance the level of accessibility of the books [2], [3] or to develop resources for text-to-speech (TTS) development [4], [5], [6].

A large project was undertaken in Portugal to improve the access to digital audiobooks by print-disabled readers [2]. Amongst other things, an ASR system was developed to automatically align the audio and text at phone level. The authors reported challenges such as bad audio quality of the original analogue recordings, differences of quality within the same book, inconsistent reading of tables, figures, chapter numbers, etc. A pilot corpus was therefore compiled for the development of their alignment system which used a hybrid of Hidden Markov Models (HMMs) and a Multi-Layer Perceptron (MLP) to perform acoustic modelling and a Weighted Finite State Transducer (WFST) framework for pronunciation modelling. The system achieved phone level alignment accuracies of more than 90%. Speaker adaptation as well as pronunciation variation modelling were found to enhance system performance substantially [2]. Pronunciation variation seems especially beneficial to capture phenomena like vowel reduction that are often observed in read speech [2]. In addition to an automatic alignment system, a *Digital Talking Book* player incorporating TTS playback and ASR-enabled navigation were also developed during the same project [3].

From a TTS point of view, aligned audiobooks constitute rich speech databases for more natural acoustic modelling because they capture broader prosodic contexts such as discourse, information structure and affect that are expressed

beyond sentence level. However, many books are published as large, unsegmented audio files and traditional alignment strategies may fail because of the huge memory requirements associated with the alignment of big audio files. In [4] and [5] the authors propose modifications to the Viterbi algorithm that enable the automatic segmentation of large, multi-paragraph speech databases. The proposed technique is independent of the duration of the target audio file.

Another technique that was proposed in the TTS domain is Lightly Supervised alignment [6]. The book under investigation was first segmented into small audio chunks of about 30 seconds each. The resulting audio files were submitted to a two-pass recognition strategy. During the first pass the files were processed by a large-vocabulary, speaker independent system for general segmentation and during the second pass the alignments were improved by using Maximum Likelihood Linear Regression (MLLR) to adapt the models to the speaker specific characteristics of the reader. In addition, the acoustic models are supported by a language model that consists of an interpolation between a general background language model and one trained on the text of the audiobook. The authors show that the proposed approach is able to extract the majority of correctly read sentences without any manual intervention [6].

In this study, automatic alignment was first performed with acoustic models trained on out-of-domain but channel-matched data. Alignment was subsequently repeated using acoustic models that were either adapted using Maximum A Posteriori (MAP) estimation or trained with in-domain data, and the effectiveness of the various approaches compared.

III. Pronunciation dictionary & Speech data

A. Pronunciation dictionary

An existing Afrikaans pronunciation dictionary containing around 24 000 entries [7] was used during system development. Grapheme-to-phoneme (g2p) rules [8] were extracted from the dictionary to generate pronunciations for words in the text that are not in the dictionary.

B. Speech data

In 2010 the National Centre for Human Language Technology (NCHLT) launched a number of projects to support HLT resource development for all 11 official languages of South Africa. During one of these projects broadband (16 kHz) speech corpora were collected for each language. The corpora all contain in the order of 80 to 90 hours of speech data. In this study, the Afrikaans NCHLT speech corpus was used to *train* the baseline acoustic models.

The *test* data constitutes an audio version of *"Ruiter in die Nag"*, an Afrikaans book by Mikro that was published in 1936. The audiobook was originally recorded on analogue tapes in 1960 and was recently converted to digital format. "Ruiter in die Nag" (loosely translated as *"The Rider in the Night"*) was chosen because we had access to both an audio and a text version and because the copyright on it has already expired, so the data can be made available freely for research purposes. The book consists of 17 chapters, each with an average duration of about 12 minutes. In total, it yielded 3.25 hours of read speech produced by a single speaker.

IV. ASR systems

Three different ASR systems were developed in order to evaluate the effect of different acoustic modelling approaches on alignment accuracy. The systems all had the same basic system architecture and were implemented using HTK [9], a well-known Hidden Markov Model Toolkit.

A. Feature extraction

Standard 39-dimensional (13 static, 13 delta and 13 delta-delta) MFCC features were extracted from the data. Cepstral mean and variance normalisation was applied.

B. Acoustic models

All the acoustic models were standard 3-state, left-to-right context dependent triphone HMMs with decision tree clustering and semi-tied transforms, corresponding to the Afrikaans phone set. Three different sets of acoustic models were used to perform alignment: baseline, MAP-adapted and audiobook models.

1) Baseline models: The baseline acoustic models were trained on approximately 90 hours of broadband (16 kHz) Afrikaans speech data from the Afrikaans NCHLT corpus.

2) Maximum A Posteriori (MAP) adapted models: A second set of acoustic models was created by using the speech data from the audiobook to perform MAP adaptation on the baseline models.

3) audiobook models: The third set of acoustic models was trained on the audiobook itself.

V. Automatic alignment verification

Once the audiobook has been aligned, it would be ideal to have a clear measure of the accuracy of the alignment without requiring manual verification. As an automatic measure of alignment accuracy, we compare the difference in the final aligned starting position of each word, with an estimate of the starting position obtained using phoneme recognition.

Specifically, we decode each chapter using a flat phone grammar, creating a single string of phonemes. We also generate a target phoneme string per chapter, using the aligned text and dictionary as input. Forced alignment is used to select the best among competing pronunciation variants. Once these two phone strings have been obtained, we use dynamic programming to find the corresponding phones (and therefore words) in the two strings. As each phone is associated with timing information (either from the alignment, or from the decoding process) we now have two estimates of the word starting position. If there is a discrepancy in starting position estimates, we flag this as a potential alignment error.

This is related to the validation technique used in [10], except that the dynamic programming scores are not used at all, and the difference in timing information is directly used as a confidence measure. As in [10] the dynamic programming process to match the two phone strings can be made more accurate by using a variable cost matrix or, if limited errors in the corpus, a flat scoring matrix can be used.

VI. RESULTS

Manually verified word-level segmentations of the first three chapters of the audiobook were created to serve as a gold standard. Specifically, the alignments obtained using the baseline models were manually verified by a language practitioner and word boundaries moved where these were not correctly aligned with the audio. This is illustrated in Fig. 1: four different alignments are displayed below the waveform and spectrogram. The language practitioner was provided with the first (top) alignment, and moved word boundaries where words were not correctly aligned. This resulted in the gold standard alignment shown fourth (at the bottom). In this example, the word 'oom' was wrongly aligned to the left of the silence portion, and corrected.

Note that, while this provides a trustworthy alignment when identifying word-level errors, the gold standard will at the millisecond-level be biased towards the models that were used to create the initial alignments. See for example the boundaries of the word 'renen'in Fig. 1; these are at identical positions for the gold standard and the first two alignments (baseline and MAP-adapted), but drawn in a slightly different position by the Audiobook models, which are the models that are most different from the initial baseline.

Before extracting final results, the gold standard itself was evaluated. All possible alignment errors of more than 100ms (obtained using the automated verification tools, which does not use the gold standard at all) were flagged for manual evaluation. All segments flagged by all three models were reviewed. This resulted in a subset of 'difficult-to-align' segments that were carefully reviewed for protocol errors, which were corrected if the observed error caused a discrepancy of more than 50ms. Two main protocol errors were observed: silence that was not inserted when needed and word starting points that were not correctly set if a silence preceded the word. 240 segments were reviewed and 24 segments corrected. (An additional random selection of 50 segments resulted in no addiontal corrections.)

The audiobook was already aligned at chapter level. Forced alignment was performed for each chapter individually using ASR systems based on the three sets of acoustic models described in Section IV-B. Alignment accuracy was evaluated by comparing the automatically generated word boundaries to the gold standard. The comparison was quantified in terms of *duration independent overlap rate* (DIOR), defined in [11] as:

$$DIOR = \frac{D_{com}}{D_{max}} = \frac{D_{com}}{D_{ref} + D_{auto} - D_{com}} \quad (1)$$

where D_{com}, D_{max}, D_{ref} and D_{auto} are the common, maximum, reference and automatic durations, respectively. This definition is not as directly applicable to audiobook alignment as to TTS; we therefore propose a modified measure where words are considered correct as long as their start times in the gold and automatic alignments respectively, are within ϵ of each other. At a value of $\epsilon = 100ms$ we obtain the DIOR results reported on in Table I. The values in the table represent the average value over the three chapters for which a gold standard was available.

Acoustic models	Average modified DIOR
Baseline	0.977
MAP-adapted	0.990
audiobook	0.996

TABLE I
AVERAGE MODIFIED DIOR FOR BASELINE, MAP-ADAPTED AND AUDIOBOOK MODELS

Table I shows that using the baseline acoustic models to perform forced alignment already result in an average DIOR of 0.977. This value increases to 0.990 for the MAP-adapted models and to 0.996 for the audiobook acoustic models.

Comparing the gold standard (manually corrected) alignments with the automatically obtained alignments, we find that fairly few errors occur. Table II lists the alignment errors found in the first three chapters of the audiobook, when using different error margins. (These errors represent individual words where the difference in starting time between the automated alignment and the manual alignment is more than the error margin ϵ.

Acoustic models	50ms	100ms	150ms	200ms
Baseline	484	270	182	131
MAP-adapted	334	114	72	46
audiobook	396	61	36	24

TABLE II
ALIGNMENT ERRORS FOR DIFFERENT ERROR MARGINS

If the 50ms margin is not considered, it is clear that the MAP-adapted models provide an accuracy improvement over the baseline, and that the audiobook models are again an improvement over the MAP-adapted models. At the 50ms margin, the superior performance of the MAP-adapted models (over the audiobook models) may be due to the bias of the gold standard, as described in Section VI.

Next, we evaluate our ability to flag possible alignment errors in the final aligned audiobook. Fig. 2 shows Detection Error Trade-off (DET) curves for the three acoustic models. Each curve plots the percentage of true errors flagged versus the percentage of correctly accepted alignments (where the number of true errors flagged depends on the error margin selected). The example illustrated in Fig. 2 corresponds to an error margin of 150ms. The difference in ms between aligned and decoded (estimated) word starting points is used as threshold when constructing the DET curves.

The effect of requiring stricter or more lenient error margins is illustrated in Fig. 3. We compare the DET curves for different error margins and the audiobook acoustic models. At one second, perfect error detection is achieved; at around 150 ms an equal error rate of 0.861 is obtained.

Further error analysis indicated that the main causes of alignment errors were (a) speaker errors resulting in hesitations, missing or repeated words, (b) rapid speech containing

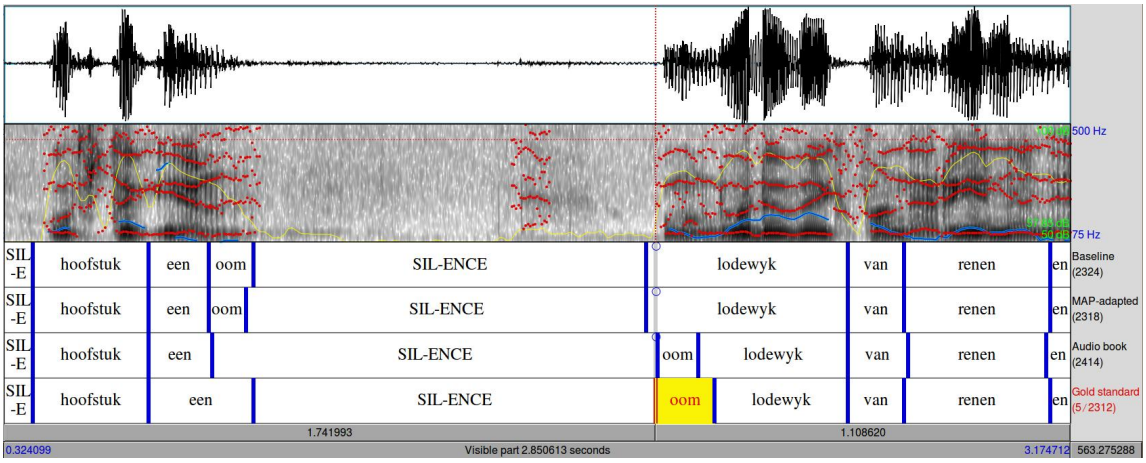

Fig. 1. Example of different alignments obtained for a sentence in the audiobook.

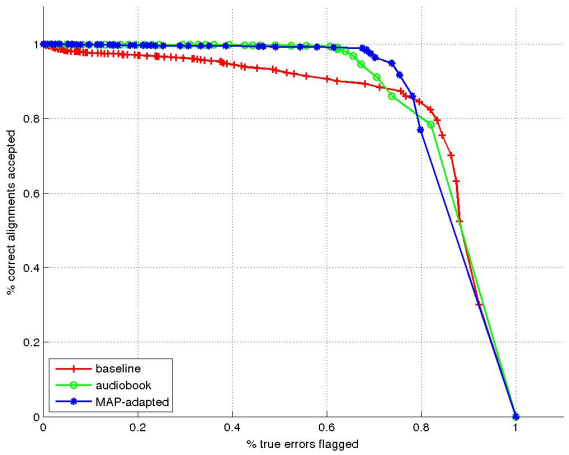

Fig. 2. DET curves for the three acoustic models at a 150ms error margin.

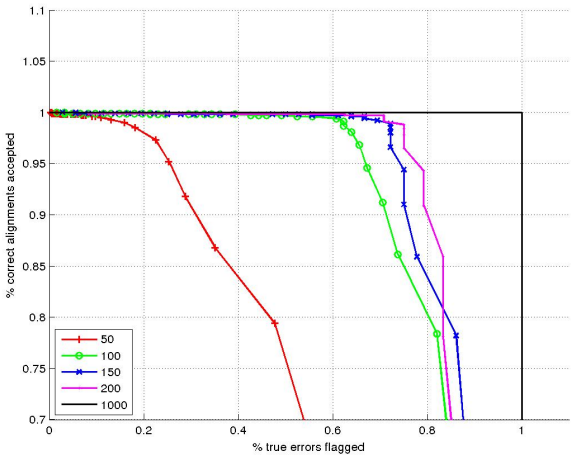

Fig. 3. DET curves for the audiobook acoustic model at different error margins.

contractions, (c) difficulty in identifying the starting position of very short (one- or two-phoneme words) and (d) a few text normalisation errors (for example, 'eenduisend negehonderd' for 'neentienhonderd').

A final observation relates to the applicability of the pronunciation dictionary used. As the alignment verification process associates a decoded phone string with each word, this produces a set of alternative pronunciations that can be considered per word. By counting the number of times the same pronunciation is observed, frequently occurring pronunciations not found in the dictionary can be added and the system retrained. In the current work, initial pronunciations were of sufficient quality that this process was not necessary to improve alignment quality, but for audiobooks that contain large numbers of unknown words (such as expected from study guides or other technical material) this may be a useful addition to the process.

VII. Conclusions

The results obtained in this study indicate that the alignments obtained by a baseline system are already good enough for practical purposes, i.e. to provide word-level mark-up for DAISY books. They also show that alignment accuracy can be improved by performing MAP adaptation on the baseline models – a fast and efficient solution requiring minimal computation. The best results are obtained with acoustic models trained on the target audiobook.

We have also shown that dynamic programming can be used to align the freely decoded and forced aligned phone strings associated with each chapter to yield an automatic measure of alignment accuracy. Error margins are defined in terms of the difference between estimated starting positions of words in the two phone strings. For an error margin of 150 ms the technique is able to accept correct alignments and flag true errors with an accuracy of 86%. For a larger error margin (of 1 second), 100% accurate alignment accuracy is achieved: all true alignment errors are rejected, and all accurately aligned words are correctly accepted.

The process will be repeated for additional audiobooks in the near future. While the voice artist spoke very rapidly, the audiobook contained few speaker errors; it would be useful to understand the extent to which a larger percentage of errors can be tolerated (and identified during alignment verification). Follow-up research will also investigate the impact of using gender-dependent baseline models on the alignment accuracy of the final systems as well as the bias of the gold standard towards the initial alignments. The results will be used to design an automated process that can be used to align large volumes of audiobooks in a fully automated way.

ACKNOWLEDGEMENTS

We would like to thank Willem van der Walt for sparking our interest in audiobook alignment and for providing information on DAISY books.

REFERENCES

[1] "Daisy," 2012, http://www.daisy.org/, Accessed in October 2012.

[2] A. Serralheiro, D. Caseiro, H. Meinedo, and I. Trancoso, "Word alignment in digital talking books using WFSTs," *Research and Advanced Technology for Digital Libraries - Lecture Notes in Computer Science*, vol. 2458/2002, pp. 508–515, 2002.

[3] I. Trancoso, C. Duarte, A. Serralheiro, D. Caseiro, L. Carrico, and C. Viana, "Spoken language technologies applied to digital talking books," in *Proceedings of Interspeech*, 2006.

[4] K. Prahallad, A. R. Toth, and A. W. Black, "Automatic building of synthetic voices from large multi-paragraph speech databases," in *Proceedings of Interspeech*, 2007.

[5] K. Prahallad and A. W. Black, "Segmentation of monologues in audio books for building synthetic voices," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 5, pp. 1444–1449, July 2011.

[6] N. Braunschweiler, M. Gales, and S. Buchholz, "Lightly supervised recognition for automatic alignment of large coherent speech recordings," in *Proceedings of Interspeech*, 2010, pp. 2222–2225.

[7] M. Davel and F. de Wet, "Verifying pronunciation dictionaries using conflict analysis," in *Proceedings of Interspeech*, Tokyo, Japan, 2010, pp. 1898–1901.

[8] M. Davel and E. Barnard, "Pronunciation prediction with Default&Refine," *Computer Speech and Language*, vol. 22, pp. 374–393, 2008.

[9] S. J. Young, G. Evermann, M. J. F. Gales, D. Kershaw, G. Moore, J. J. Odell, D. G. Ollason, D. Povey, V. Valtchev, and P. C. Woodland, "The HTK book version 3.4." Cambridge, UK, 2006.

[10] M. Davel, C. J. van Heerden, and E. Barnard, "Validating smartphone-collected speech corpora," in *Proceedings of SLTU*, Cape Town, South Africa, May 2012, pp. 68–75.

[11] S. Paulo and L. C. Oliveira, "Automatic phonetic alignment and its confidence measures," *Advances in Natural Language Processing*, 2004.