Lightly supervised recognition for automatic alignment of large coherent speech recordings

Norbert Braunschweiler, M.J.F. Gales, Sabine Buchholz

Toshiba Research Europe Ltd., Cambridge Research Laboratory, United Kingdom
{norbert.braunschweiler,mark.gales,sabine.buchholz}@crl.toshiba.co.uk

Abstract

Large quantities of audio data with associated text such as audiobooks are nowadays available. These data are attractive for a range of research areas as they include features that go beyond the level of single sentences. The proposed approach allows high quality transcriptions and associated alignments of this form of data to be automatically generated. It combines information from lightly supervised recognition and the original text to yield the final transcription. The scheme is fully automatic and has been successfully applied to a number of audiobooks. Performance measurements show low word/sentence error rates as well as high sentence boundary accuracy.

Index Terms: speech alignment, lightly supervised, large coherent speech corpora, speech synthesis.

1. Introduction

A number of research domains in the field of speech processing could benefit from the availability of training material based on large coherent speech corpora. An example of such a large coherent speech corpus is an audiobook where a narrator has read a single coherent text. This is in contrast to some speech corpora which contain unrelated sentences spoken in isolation, e.g. as typically found in a speech synthesis corpus. A coherent speech corpus includes prosodic effects that go beyond the sentence level. In addition there are often more expressive speaking styles in combination with character voices in which a speaker tries to mimic the voice of a certain character occurring in a story.

When a contemporary speech synthesizer, trained on isolated sentences, is used to synthesize a large coherent block of text like a novel it often lacks the expressiveness of a human reading. One of the reasons for this shortcoming is certainly the lack of appropriate training material.

Methods to use large coherent speech corpora such as audiobooks for building synthetic voices have been proposed in e.g. [1], [2] and [3]. Since there are a considerable number of audiobooks available in the public domain (e.g. Librivox.org) this data provides a rich source of coherent speech material with associated text in large quantities. In order to make this huge amount of data accessible for research and synthetic voice building it is necessary to align the transcript to the speech and identify any mismatches between them.

In this paper, a method is introduced that automatically provides high quality transcriptions and associated alignments of large coherent speech corpora to be obtained. The method identifies regions in the large audio files that correspond to sentences in the associated text and defines their location by timestamps. It also provides means to adjust the amount of word accuracy that is used to select sentences from the associated text.

The next section introduces the main issues in automatic alignment of large speech recordings. Then the proposed automatic alignment method using lightly supervised recognition techniques is presented. This is followed by an evaluation of the method using a public domain audiobook. The paper is concluded by a discussion and an outlook into future projects.

2. Automatic alignment of audiobooks

In automatic alignment, orthographic texts are time aligned against a large coherent speech recording without manual interaction. The start and end of units such as phones, words and sentences are marked against the audio recording.

There are a number of practical considerations that need to be taken into account in any method of automatic alignment. The method needs to be able to cope with differences in text and speech. A typical problem is text in the transcript which is not in the speech recording. An example is a table of contents which is not read aloud by the speaker. On the other hand the speaker might have added speech which is not in the transcript. Therefore the automatic alignment approach needs to provide means to detect insertions, deletions and substitutions made by the speaker.

To illustrate the amount of overlap between recording and associated transcript in a large coherent speech corpus the publicly available American audiobook “A Tramp Abroad” by Mark Twain is used. It contains more than 15 hours of speech and the book text includes 7 parts with 50 chapters and 6 appendices (>160,000 words). It is read by an American narrator who uses a wide range of speaking styles. Speech recordings were manually checked and the gold standard transcription generated (GOLD). The book text was automatically normalized (see section 3) and pruned of any meta-data like punctuation.

<table>
<thead>
<tr>
<th></th>
<th>Sub</th>
<th>Del</th>
<th>Ins</th>
<th>Err</th>
</tr>
</thead>
<tbody>
<tr>
<td>Words</td>
<td>0.4%</td>
<td>1.4%</td>
<td>3.6%</td>
<td>5.4%</td>
</tr>
<tr>
<td>Sentences</td>
<td>4.9%</td>
<td>2.5%</td>
<td>6.0%</td>
<td>13.4%</td>
</tr>
</tbody>
</table>

Table 2.1. Word and sentence overlap for the comparison of GOLD vs. book text. There are 157,451 words or 7672 sentences in GOLD.

The word and sentence overlap were both measured by aligning the two word/sentence lists and finding the longest common subsequence. Table 2.1 shows the number of substitutions (Sub), deletions (Del), insertions (Ins) and error rates (Err) for word and sentence overlap.

1 Sources are http://librivox.org/a-tramp-abroad-by-mark-twain/ for the audio (using 128kbps MP3-files) and http://www.gutenberg.org/ebooks/119 for the text.
Less than 2% of the words in GOLD are missing or different from the book text - this shows that there is a strong overlap between the book text and what the speaker actually recorded. However, there is a considerable amount of inserted text: more than 5600 words (3.6%) in 462 sentences (6.0%) originally occurring in the book text, but not read by the narrator. These inserted sentences are mainly a result of each of the seven parts of the book being preceded by a table of contents and a list of illustrations, but none of these are read aloud by the speaker. What is listed as deletions is text spoken by the reader which was not in the book text, i.e. introductory sentences for each chapter ("This is a Librivox recording."). The number of substitutions reveals the amount of differences between the normalized book text and what the speaker actually said, these differences include text normalization errors or simple word-differences like ‘dueling’ instead of ‘due’. When discarding all the inserted and deleted sentences there is a word error rate of 0.61% on the remaining sentences. This figure provides the upper limit of performance for any alignment which could detect all inserted and deleted sentences and which used the book text as alignment text.

One method to align text and speech is to identify the stretch of text which best corresponds to a large audio file. In [1] a specially adapted forced alignment approach is presented to align large multi-paragraph speech recordings. Manual interaction was required for these alignments as text was added to match the beginnings and ends inserted by the speakers. This kind of procedure is typically needed in alignment approaches to handle differences between transcription and speech. Other methods proposed in the literature use weighted finite state transducers (WFST, e.g. [3], [4]) or factor automata [5]. The forced alignment approach will simply find the best match in terms of state sequence between text and a block of audio and not necessarily the correct sequence. It relies on simple beam pruning to identify erroneous matches. Using forced alignment schemes with no manual interaction often fail or align deleted text. This was found to be the case for the data in Table 2.1 in a pilot study.

In this paper a method is presented that overcomes the limitations of forced alignment approaches. The method is applied to a publicly available audiobook and evaluated. The next section will describe this method in more detail.

3. Lightly supervised alignment

The proposed approach can be described as lightly supervised recognition [6] plus sentence selection (henceforth called LightlySupervised). The basic idea of this approach is to run a recognizer over the whole audio data and to use the recognized word sequences and their locations in time for automatic alignment. In order to support the recognizer in this task a language model (LM) is created based on the text of the book.

The LightlySupervised approach runs fully automatically and therefore differs significantly from approaches which need to adapt the text to the audio in order to work reliably. The flowchart for the LightlySupervised approach is illustrated in Figure 1.

In TextNormalization first any automatically identifiable headers and footers are stripped from the text. Then the text is pruned of some special characters/formatting styles which can cause text normalization problems (e.g. sequences of asterisks/dashes). The text is then automatically split into sentences and normalized using the Toshiba speech synthesis engine ToSpeak [7]. ToSpeak is also used to generate the lexicon used in the recognizer. Any phonetic transcriptions for words not in the internal ToSpeak lexicon are automatically generated.

![Flowchart of automatic alignment using lightly supervised recognition and sentence selection.](image)

Figure 1: Flowchart of automatic alignment using lightly supervised recognition and sentence selection.

After these text preparations the language model is created. The LM is used to bias the recognizer towards the book text utilizing the assumed overlap between book text and spoken text. A background LM is trained on the English Gigaword corpus [8]. An in-domain LM is trained on the normalized sentences. These two LMs are then combined by linear interpolation using the CMU SLM toolkit [9]. The LM vocabulary contains the unique words from the book text (e.g. roughly 13,200 words in “A Tramp Abroad”). The LM interpolation weight for the in-domain model was set to 0.9, i.e. a strong bias towards the book text.

The audio conversion step includes any audio file format conversion, down-sampling and acoustic feature file generation. Librivox audio files are available in compressed audio formats (MP3, OGG VORBIS) therefore they are first converted into WAV format. For each large audio file the corresponding MFCC file is created for the recognition stage.

For efficiency reasons, the large audio files (1 min - >1h) were split into small audio chunks (0.5 sec – 30 sec). An audio-only version of the automatic prosody labeling tool, Prosodizer [10], was used for this purpose. It detects prosodic events such as pitch accents and boundary tones. Only major intonational phrase boundary tones are used for automatic chunking. Any other pause detection method could be used for this purpose, such as running a phone loop to identify regions of silence.

The recognition part uses a set of existing acoustic models. These speaker independent models were trained on >150h of read speech data (text type was newspaper, i.e. WSJ0 + WSJ1). These models used a basic setup, i.e. more complex models could improve recognition accuracy. The state-clustered triphone models were maximum likelihood trained and include 650 states and 12 components (mixtures) per state. The acoustic feature vector includes 39 components comprised of 13 cepstrum coefficients including the 0th order coefficient, with the first and second order derivatives. A large vocabulary speech recognition tool (HDecode tool from HTKv3.4.1 [10][11]) is used for recognition. Running the recognizer is done in two steps: first the recognizer using speaker independent models is applied to a single chapter split into small audio chunks. The recognized output is used to adapt the speaker independent models to the audiobook speaker using
Maximum Likelihood Linear Regression (MLLR) adaptation [12] to improve the recognition accuracy. The adaptation uses full MLLR transformations with 32 classes. Second, the recognition is rerun for each chapter using speaker adapted models.

After the recognition stage the sentence selection stage is performed. The recognition output is a long sequence of words without punctuation or other meta-data. To complete the alignment, the recognition sequence must be split into sentences corresponding to the book text, and invalid or questionable text-speech alignments removed, and the recognized text replaced with book sentences in order to retain any punctuation or meta-data in the original text. These steps are now described in more detail below.

The normalized book text is aligned with the word sequence created from the recognition output by finding the longest common sub-sequence. The alignment is then used to split the recognized word sequences into sentences according to the book text.

Word error rates for each sentence alignment are calculated to select or discard sentences using an adaptable threshold value. When a sentence is selected the recognized word sequence is replaced with the words from the book text.

The text matching step is motivated by the assumption that most of the book text is actually read by the speaker and therefore the book text is basically ‘overlaid’ over the recognized word sequences, thereby correcting any (small) misrecognitions. However, the approach does not handle sentences inserted by the speaker, these will be discarded. Other mechanisms could be applied to capture sentences inserted by the speaker, by e.g. evaluating confusion networks for specific sentences and using the results as additional means to select or deselect sentences.

In order to get the timestamps of sentence ends a forced alignment across a whole large audio file is performed. The new word sequence consisting of a mixture of recognized words and words taken from the book text will be aligned to get timestamps of sentence end words.

4. Evaluation

As basis for the evaluation the manually checked American audiobook “A Tramp Abroad” (see section 2 and footnote 1) was used and represents the gold standard (GOLD) in the following evaluation results.

The LightlySupervised approach relies on achieving sufficiently good accuracy at the recognition stage. Table 4.1 shows the word and sentence error rates after the recognition steps have been performed (including speaker adaption of HMM models which yield a 9% absolute improvement in word accuracy). As can be seen reasonably low word error rates are achieved in recognition. It also becomes clear that the approach significantly reduces the number of insertions: word insertions are down to 1% from 3.6% and sentence insertions are down to 0.9% from 6% (cf. Table 2.1 the comparison of GOLD vs. BOOK). However, the number of substitutions went up from 0.4% to 1.8% on the word level and from 4.9% to 11.5% on the sentence level. The overall WER of 3.3% is higher than the 0.6% upper limit WER when using the book text and detecting all inserted/deleted sentences, but better than the WER of the complete book text (5.4%). By combining the two sources of information, i.e. (1) book text and (2) lightly supervised recognition output better WERs can be achieved as shown next.

<table>
<thead>
<tr>
<th>Threshold</th>
<th>Sents Extracted</th>
<th>Extracted WER BOOK</th>
<th>LS_REC</th>
</tr>
</thead>
<tbody>
<tr>
<td>-</td>
<td>97.9%</td>
<td>1.6%</td>
<td>3.3%</td>
</tr>
<tr>
<td>&gt;0%</td>
<td>94.4%</td>
<td>1.1%</td>
<td>2.9%</td>
</tr>
<tr>
<td>50%</td>
<td>92.1%</td>
<td>0.5%</td>
<td>2.6%</td>
</tr>
<tr>
<td>75%</td>
<td>88.9%</td>
<td>0.4%</td>
<td>2.2%</td>
</tr>
<tr>
<td>100%</td>
<td>65.4%</td>
<td>0.2%</td>
<td>0.2%</td>
</tr>
</tbody>
</table>

Table 4.2. Effect of sentence selection word accuracy threshold on % of sentences extracted and LightlySupervised output text accuracy.

When looking at the last line of Table 4.2 it can be seen that the output WER does not reach 0% even at the 100% threshold. There are 3 potential causes for this: (a) errors in the manual transcription, (b) spelling mistake corrections, (c) incorrect recognition heavily biased to book text.

One interesting question related to the selection process is: How different are the selected sentences which are different from GOLD, in terms of their word accuracy? To check, the 75% threshold was chosen. At that threshold 3.4% of the selected sentences differed from GOLD. Their average word accuracy is 91% and their word correct rate is 96%, indicating that these sentences are actually not too far off. This shows that the LightlySupervised approach is able to extract the majority (e.g. 88.9% at a 75% threshold) of sentences with high word accuracy without any manual intervention.

Following the recognition stage the sentences to be kept will be selected and mapped to the book text. For each sentence a threshold based on the word accuracy of the recognized text to the book text was chosen to make this selection. To show the effect of different word accuracy thresholds Table 4.2 presents the variation in % sentences extracted depending on this threshold. Also presented is the WER of the sentences after mapping to the book text (BOOK). For information the WER of the recognized sentences (i.e. text prior to conversion; LS_REC) is also shown. It can be seen that the number of false accepts stays moderately low even at the lowest possible threshold where all sentences which could be aligned to a book sentence are selected. The WERs at no threshold and >0% threshold are higher than the 0.6% upper limit WER because of alignment issues like some book sentences aligned to single recognized words. However, at the 50% threshold the WER of the 92.1% selected sentences drops to 0.5% even lower than the upper limit of 0.6% of the 97.5% sentences (when discarding any deleted/inserted book text sentences). The WER could even be reduced down to 0.2% at the 100% threshold of course at the expense of pruning about 45% of sentences. This shows the effectiveness of the sentence selection stage as compared to the output of the lightly supervised recognition.

<table>
<thead>
<tr>
<th>Threshold</th>
<th>Words</th>
<th>Del</th>
<th>Ins</th>
<th>Err</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sentences</td>
<td>11.5%</td>
<td>3.0%</td>
<td>0.9%</td>
<td>15.4%</td>
</tr>
</tbody>
</table>

Table 4.1. Word and sentence error rates for the recognition output before sentence selection. There are 157,451 words or 7672 sentences in GOLD.
85% at the 75% threshold and even at the 100% threshold usually more than 50% of sentences are extracted.

<table>
<thead>
<tr>
<th>Book title</th>
<th># Book sentences</th>
<th>Word accuracy threshold</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>75%</td>
</tr>
<tr>
<td>A Tramp Abroad</td>
<td>7941</td>
<td>88.9%</td>
</tr>
<tr>
<td>Gulliver of Mars</td>
<td>2662</td>
<td>95.6%</td>
</tr>
<tr>
<td>Emma</td>
<td>8195</td>
<td>87.0%</td>
</tr>
<tr>
<td>Sons and Lovers</td>
<td>16500</td>
<td>89.0%</td>
</tr>
<tr>
<td>Walden</td>
<td>4305</td>
<td>86.5%</td>
</tr>
</tbody>
</table>

Table 4.3. Examples for the % of sentences extracted relative to the sentences in the book for two word accuracy thresholds (75% and 100%).

Since one of the goals of this research is to enable the study of paragraph related speech effects it is important to know the number of complete paragraphs aligned, i.e. paragraphs where none of the included sentences have been discarded. Therefore the first five chapters of “A Tramp Abroad” were analyzed and the number of complete paragraphs counted (at the 75% threshold). Across these first five chapters 88% of all paragraphs are complete, indicating that there will be sufficient training material to study paragraph related effects.

To assess the suitability of the extracted sentences for approaches that require the location of sentences in the audio, the accuracy of the sentence boundary timestamps is investigated. This was evaluated by manually labelling a test-set consisting of about 1.5% of all sentences from “A Tramp Abroad”. The test-set sentences were randomly chosen. More than 95% of the boundaries were found to be OK, 1.5% had a bad onset, 3.1% had a bad offset. None had both boundaries off. This shows that the sentence boundaries produced by the LightlySupervised approach are mostly accurate and the number of bad boundaries is small.

5. Discussion

The LightlySupervised approach runs fully automatically, with no need for manual checking of the text-audio correspondence prior to runtime or manual intervention during running. It is independent of the duration of the speech recordings. Since it uses a recognition approach it does need a set of speaker independent HMM models for the recognition stage. The strength of the approach lies in the way it automatically handles differences in text and speech while maintaining high output accuracy. Contrary to many forced alignment approaches manual adaptations of text to fit the speech recordings are not required.

It also has a choice of operating points by adapting parameters like the word accuracy threshold which is used to select or discard recognized sentences. Depending on the threshold set, a higher word/sentence accuracy can be achieved when this is the main goal. A lot of data can be retrieved even at the 100% threshold – over 9 hours for “A Tramp Abroad”. If a certain amount of text–speech differences can be accepted the threshold can be lowered and more sentences are selected. This is particularly interesting for building corpora for synthetic voices or corpora for analysing speech effects unique to large coherent speech recordings.

6. Conclusions

An approach to perform automatic alignment of large coherent speech corpora with associated transcripts has been presented and evaluated. The LightlySupervised approach developed fully automatically extracted large amounts of data suitable for research and synthetic voice building. It uses two sources of information: (1) original book text and (2) lightly supervised recognition output. By combining the two it is possible to obtain high accuracy transcriptions and good sentence boundary accuracy while reducing the number of deletions, insertions and substitutions. The percentage of the extracted sentences and the accuracy of transcriptions can also be controlled.

The approach has been successfully applied to more than 20 American English audiobooks including differences in recording conditions, speakers, speaking styles and amounts of audio data. Future work includes the refinement of acoustic models, the improvement of the language model, and handling of inserted speech. Another major issue concerns the detection of speaking styles, i.e., how to separate sentences in neutral speaking style from other sentences, including e.g. emotionally coloured speaking styles or character voices. This could provide valuable training material for a variety of natural language processing domains, including text normalization, speaking styles, emotion detection, paragraph based speech effects, and synthetic voice building to name just a few.

7. References