Automatic Alignment and Error Correction of Human Generated Transcripts for Long Speech Recordings[†]

Timothy J. Hazen

MIT Computer Science and Artificial Intelligence Laboratory
Cambridge, Massachusetts, USA
hazen@csail.mit.edu

Abstract

In this paper we examine the issues of aligning and correcting approximate human generated transcripts for long audio files. Accurate time-aligned transcriptions help provide easier access to audio materials by aiding downstream applications such as the indexing, summarizing and retrieving of audio segments. Accurate time alignments are also necessary when incorporating audio data into the training data for a speech recognizer's acoustic model. We provide some initial analysis of manual transcriptions which show that there can be significant differences between the "approximate" manual transcripts generated by typical commercial transcription services and what was actually spoken in the recording. We then present a new alignment approach for approximate transcriptions of long audio files which is designed to discover and correct errors in the manual transcription during the alignment process.

Index Terms: Speech alignment, speech recognition, error detection and correction.

1. Introduction

In recent years, improvements in data storage and networking technology have made it feasible to provide Internet users with access to large amounts of multimedia content. For example, many universities are now providing free web-based access to audio-visual recordings of academic lectures (e.g., MIT's OpenCourseWare website: http://ocw.mit.edu). Unlike text, however, audio-visual material is not easy to search and browse without time-aligned transcriptions. Because manual transcription can be a costly and time-consuming process, the development of automatic approaches for transcribing lectures is an obvious step towards making multimedia content more accessible. However, in many cases, *approximate* manual transcriptions (i.e., imperfect transcripts that were generated quickly and/or cheaply) may be available. In this case, performing an automatic alignment of these imperfect transcripts with the speech in the audio file is preferable to automatically generating a new transcription.

The manual generation of accurate transcriptions of large amounts speech has also helped advance the state-of-the-art in automatic speech recognition. However, today's recognition systems are typically trained on hundreds (or even thousands) of hours of speech. The continuous increase in the size of recognition training corpora may make careful annotation of training data prohibitively expensive. The use of quickly-generated approximate transcripts is one means of reducing cost, provided a means for accurately aligning this data and detecting errors in the approximate transcriptions can be devised.

In this paper, we examine a new approach for forced alignment of long approximate transcriptions of spoken lectures which is designed to discover and correct errors in the manual transcripts, hopefully improving the fidelity and usefulness of the transcripts in the process.

2. Human Transcription of Speech

In generating transcripts for arbitrary audio data there is an obvious tradeoff between the time required to generate a transcript and the accuracy of the transcript. The skill of the human transcriber plays an important role in this tradeoff. For some applications, such as real-time closed captioning for television or human transcription of court proceedings, a highly-skilled stenographer is typically employed. Today's stenographers are usually aided by a computer application that converts phonetically-based stenography into a written transcript. Stenographers or *stenographic reporters* can achieve the official title of "Certified Real-time Reporter" when they achieve an average transcription accuracy of 95%. Additional post-transcription editing can further improve a stenographer's accuracy if necessary. Unfortunately, skilled stenographers are in short-supply and can be expensive, thereby making their services prohibitive for many tasks [8].

When producing transcripts of pre-recorded materials, the time it takes to produce the transcript is of less importance than the cost. Thus commercial transcription services often employ lower-waged transcribers who need only be efficient typists and competent in the language of the recording. In our experiences with commercial transcription services, approximately correct transcriptions can be produced at speeds varying from 3 to 5 times real-time depending on the quality of the transcription desired and the skill of the transcriber. These transcripts are often enhanced with appropriate punctuation and capitalization. With minimal additional effort they can also include occasional time stamps to aid in alignment with the audio. Highly accurate transcriptions that account for all speech events (including filled pauses, partial words, etc.) as well as other meta-data (speaker identities and changes, non-speech artifacts and noises, etc.) can take up to 50 times real-time depending on the nature of the data and the level of detail of the meta-data [2, 13].

For our research, we have employed commercial transcription services to transcribe academic lecture audio data. An examination of the initial transcripts we received yielded several observations. First, they were considerably *cleaner* than the actual speech. Filled pauses, false starts, partial words, grammatical errors and other speech errors were typically removed or

[†]Support for this research was provided in part by the MIT/Microsoft iCampus Alliance for Educational Technology.

Misspelling Examples	Substitution Examples
Furui → Frewey	Fourier → for your
Makhoul → McCool	a priori → odd prairie
Tukey → Tuki	resonant → resident
Eigen → igan	affricates → aggregates
Gaussian → galsian	palatal → powerful
cepstrum → capstrum	Kullback → callback

Table 1: Examples of human errors made in the transcription of several academic lectures on the topic of speech recognition.

	Word Error Rates (%)			
Lecture	Sub.	Del.	Ins.	Total
A	3.4	7.4	1.2	12.0
B	2.0	7.4	0.9	10.2
C	2.0	5.1	0.7	7.7
Average	2.4	6.6	0.9	10.0

Table 2: Word error rates (broken down into substitutions, insertions and deletions) of *approximate* transcriptions provided by a commercial transcription service for three lectures.

corrected in the transcriptions. Unfortunately, we also observed that the transcribers, having little knowledge of the subject-matter contained within the audio data, were often unable to properly transcribe many of the subject-specific terms or proper nouns. In some cases, these errors were easily detected and corrected by locating misspelled words in the transcripts and manually editing them. In other cases though, the transcribers replaced subject specific terms with off-topic words that sounded similar.[1] These substitutions are more problematic because they are harder to find and correct. Table 1 shows a few example misspellings and word substitutions we observed in the transcriptions of three lectures on speech recognition.

To gain better insight into the accuracy of the approximate transcriptions, we manually corrected transcriptions of three lectures from a speech recognition course (roughly 3.75 hours of speech). We paid special care to producing high-fidelity representations of what was actually said, including all filled pauses, partial words and other speech artifacts. We then compared these high-quality transcripts against the approximate transcriptions in order to estimate their error rates. Even after correcting all misspellings and ignoring all filled pauses and partial words, the word error rate of the approximate transcriptions was 10.0%. Similar error rates have been observed in closed captions generated for TV news broadcasts [6].

Table 2 shows a breakdown of the error rates of the approximate transcriptions for the three speech recognition lectures (each from a different lecturer). Deletions accounted for an absolute 6.6% of the 10.0% error rate. The deleted words were often false starts, mistakenly repeated words, or common conversational insertions (such as *okay* or *right*). However, 2.4% of the 10.0% error rate resulted from word substitutions.

3. Semi-Automatic Transcription of Speech

3.1. Overview

The analysis of the error rates of the rough transcriptions leads one to wonder if improved transcriptions can be obtained through a semi-automatic means (without increasing the effort of the human transcribers), or if semi-automatic methods can be used to reduce the human transcription time (without harming transcription accuracy). In the latter case, it has been found that human transcribers can correct the output of an automatic speech recognition system more efficiently than they can transcribe the same speech audio from scratch, provided the initial error rate of the automatically generated transcript is not too severe [1, 4].

In this work, we seek to investigate the former approach. If we have already obtained an human-generated approximate transcription of a lecture we might ask two questions: (1) "Can approximate transcriptions be accurately time-aligned to the speech?", and (2) "Can approximate transcriptions be automatically edited to correct human errors?".

The automatic alignment of found transcriptions to lengthy speech recordings has been studied by others in the past. Caseiro et al. have implemented a single pass approach for text alignment of spoken books [5]. This approach is successful for spoken books because the speech is typically a verbatim reading of the text and there are no difficult audio conditions present that often accompany other types of audio (e.g., background noises and non-speech events, far-field speech for secondary speakers, etc.). A sequential single-pass approach is not feasible for many types of audio because it is too easy for the recognizer to become desynchronized with the approximate transcript during a difficult audio segment and then fail to resynchronize itself.

Moreno et al. developed a recursive method to address the problem [10]. Their approach is based upon the discovery of reliable anchor points found via automatic speech recognition. The use of anchor points helps alleviate problems associated with waveform regions containing poor audio quality that are difficult to align automatically. Lamel et al. also compared automatic speech recognition output against errorful closed captions as a means of aligning the closed captions within their *lightly supervised* training procedure [9].

In this work we have taken an approach similar to the Moreno approach, but with two primary differences. First, our approach does not use recursion to progressively reduce the alignment process to operate over smaller and smaller segments. Instead, our system uses a fixed number of recognition passes each performing a specific type of refinement to the proposed transcription alignment. Second, our system explicitly searches for discrepancies between the provided manual transcription and the observed acoustics in the audio file and attempts to correct these discrepancies. The different stages of our processing are explained in detail in the sections that follow.

3.2. Stage 1: Automatic Speech Recognition

In the first stage of our approach, we use the SUMMIT automatic speech recognition (ASR) system to automatically generate a transcript for a long audio file [7, 11]. Using the manually generated transcript, we strongly bias the recognizer's language model to the content of the transcript. In our case, we implement a mixture trigram model that combines a trigram model trained on the transcript with a topic-independent trigram trained on a collection of data including the Switchboard corpus and transcribed lectures covering a variety of topics. The mixture trigram was created using the SRI language modeling tool kit [12]. We provided a weight of 0.99 for the transcript trigram and 0.01 for the topic-independent trigram.

The speech recognition result will typically be less accurate than the original manual transcript. However, by string aligning the speech recognition result against the manual transcript and

[1] When properly instructed, transcribers used by speech researchers typically mark words they cannot understand rather than blindly guess.

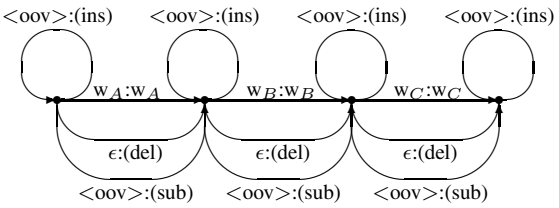

Figure 1: Example FST network used during the pseudo-forced alignment stage for the word sequence (w_A, w_B, w_C) with insertions, deletions, and substitutions allowed.

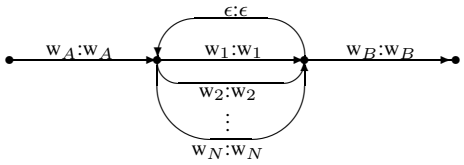

Figure 2: FST network which allows the editing stage to insert any word sequences between words w_A and w_B.

finding matching word sequences between the two transcripts, anchor points for a refined forced-alignment stage can be determined. In our case, we found that matching sequences of only two words provided sufficient initial anchor points for the following stages. When longer sequences of words are matched, we select the actual anchor points at locations where pauses in the speech (if any) are present within the matched sequence. In our experiments anchor points only rarely fell more than 10 seconds apart, and never occured more than 30 seconds apart. Of course, exceptions to these numbers could occur if an audio file contained long non-speech regions (which did not occur in the audio files in our experiments).

3.3. Stage 2: Pseudo-Forced Alignment

After obtaining anchor points from the first stage recognition, the second stage produces a *pseudo-forced alignment* of the manual transcript across the speech segments spanning between the first stage anchor points. We call it a *pseudo-forced alignment* because we do not force the recognizer to align the exact string present in the manual transcription. Instead, we assume that errors in the transcript are possible and we allow insertions of new words and substitutions for existing words through the use of a phonetic-based out-of-vocabulary (OOV) word filler model [3]. We also allow any word in the transcript to be deleted. This process is realized through the composition of a pseudo-forced alignment finite state transducer (FST) with an underlying lexical FST. An example alignment FST allowing insertions, substitutions and deletions is shown in Figure 1. The rates of the insertions, substitutions and deletions can be controlled using penalty weights to insure that correctly transcribed words are rarely replaced or deleted.

3.4. Stage 3: Alignment Editing

After the second stage is complete, the transcript is fully aligned against the speech, and regions containing potential substitutions, deletions and insertions are marked. From this transcript, we can re-run our speech recognizer over local segments containing these marked regions. Thus in the third stage, we allow the recognizer to *edit* the manual transcript by hypothesizing any word(s) in the recognizer's full vocabulary as a replacement

Alignment Stage	Alignment WER (%)
(1) ASR	24.3
(2) Forced Alignment	10.3
(3) Editing - Iter 1	9.3
(3) Editing - Iter 2	8.8
Oracle w/o Editing	10.0

Table 3: Word error rates for the different stages of our automatic alignment and error correction procedure.

for the marked substitutions and insertions. This is achieved by creating a FST network which allows insertions and substitutions discovered in the pseudo-forced alignment to be replaced by any word sequence allowed by the recognizer. An example FST is shown in Figure 2 in which the pseudo-forced alignment of a word sequence (w_A, w_B) hypothesized an insertion between the words w_A and w_B. The resulting FST network allows any sequence of the recognizers N vocabulary items to be inserted between w_A and w_B. This FST network is further composed with a general topic-independent trigram model FST to provide language model constraint. We also allow the recognizer to reconsider any deletions proposed in the previous stage. This process allows the system to correct some of the errors of the human transcriber.

One potential problem that can be encountered during the editing stage, is that the recognizer can enter into a recognition word loop (such as the one in Figure 2) and fail to leave this word loop before encountering the end of the utterance. This is possible under the condition that recognizer's pruning algorithm prunes all paths leaving the word loop to reenter the desired word sequence (e.g., word w_B in Figure 2). This is generally not a problem for utterances that are short in duration and/or contain very clean speech, but it becomes more likely when the audio becomes noisy, distorted or reverberant, or if the lecturer's speech changes dramatically from standard speech (e.g., the lecturer is shouting or laughing while lecturing). We have found that these types of errors can be mitigated by running two passes of editing. Because the process of correcting errors in the transcript improves the fidelity of alignment, we allow the system to reset the anchor points after the first editing stage. We also loosen the recognizer's pruning threshold in a second editing stage for regions that failed to find a valid full path through the editing FST during the first editing stage.

3.5. Results

Table 3 summarizes our experiments on the three lectures discussed in Section 3.3. The ASR first stage achieved a word error rate (WER) of 24.3% across the three lectures. This result is achieved with a basic speaker independent recognizer tuned to run in real time on a single processor. Better results can certainly be achieved with more sophisticated multi-pass processing and more careful control of the language model smoothing parameters, but this ASR result is adequate enough to provide a set of suitable anchor points for the subsequent stages.

The second stage forced alignment result is reported without consideration for the substitutions, insertions, and deletions proposed by the recognizer. Insertions proposed by the OOV word filler model are ignored in the WER calculation, and transcribed words that are deleted or substituted are retained at the time locations where the substitutions or deletions occured. The word error rate is reported using string alignment within manually determined pause separated speech utterances. Thus, minor

Utterance 1	Ground truth:	they give you predictable there are lots of things are predictable about
	Human Transcript:	there are lots of things are predictable about
	ASR:	they give you predictable their velocities are predictable about
	Pseudo-Forced Alignment:	(insertion) there are lots of things are predictable about
	Edited Forced Alignment:	they give you predictable there are lots of things are predictable about
Utterance 2	Ground truth:	either you try to adapt not adapt but change the input
	Human Transcript:	either you try to change the input
	ASR:	either you try to adapts not adapt but change the input
	Pseudo-Forced Alignment:	either you try to (uh) (insertion) (uh) (insertion) change the input
	Edited Forced Alignment:	either you try to adapts not adapted change the input

Table 4: Examples of two utterances processed by the system. For each utterance, the table shows the *ground truth* transcription, the approximate transcript provided by a commercial transcription services, and the output of the three stages of our alignment process.

discrepancies in word alignment are not penalized, but incorrect placement of words across pause boundaries is penalized. The forced alignment WER is directly comparable to the oracle result showing the minimum error rate achievable by an optimal string match alignment of the approximate transcript against the true transcript. The automatic alignment procedure shows only a minor 0.3% degradation in performance (from 10.0% to 10.3%) compared with the optimal alignment. This indicates that our system is producing accurate alignments.

The two iterations of the editing stage reduce the WER of the transcription from 10.3% to 8.8%. This is a 14% relative reduction in word error rate from the automatic forced alignment, and a 12% relative reduction in error rate over the optimally aligned human transcription. An examination of the results shows that almost all of the improvement is gained from the reinsertion of words omitted by the human transcriber. While the fidelity of the corrected transcriptions was improved, a preliminary examination reveals that most of the automatic corrections are not important for human comprehension. The system showed little ability to correct human mistranscriptions of important content words. This is not entirely unexpected as most of the human errors involved substitutions which are phonetically close to the actual words spoken, and hence similarly difficult for the automatic system to discriminate.

Figure 4 shows example outputs of our system for two utterances. In both utterances, the human transcriber deleted speech errors produced by the speaker. In both cases the pseudo-forced alignment stage detected the untranscribed words and the editor hypothesized reasonable word sequences for these regions. One might notice that the ASR and edited forced alignment results differ in some cases. This is the result the recognizer using different starting and ending anchor points in different stages, resulting in different language model contexts at the start and end of each utterance chunk passed to the recognizer.

4. Conclusion

In this paper we have examined the issues of aligning and correcting approximate human generated transcripts for long audio files. We have presented a new alignment approach for approximate transcriptions of long audio files which is designed to discover and correct errors in the manual transcription during the alignment process. Preliminary experiments have shown that the system can reduce the error rates of approximate human-generated transcripts by 12%.

5. References

[1] K. Bain, S. Basson, A. Faisman and D. Kanevsky, "Accessibility, transcription and access everywhere." *IBM Systems Journal*, vol. 44, no. 3, pp. 589–603, 2005.

[2] C. Barras, E. Geoffrois, Z. Wu and M. Liberman, "Transcriber: Development and use of a tool for assisting speech corpora production." *Speech Communication*, vol. 33, no. 1-2, pp. 5-22, January 2001.

[3] I. Bazzi and J. Glass, "Modeling out-of-vocabulary words for robust speech recognition." In *Proc. of ICSLP*, Beijing, October, 2000.

[4] D. Binnenpoorte and C. Cucchiarini, "Phonetic transcription of large speech corpora: How to boost efficiency without affecting quality." In *Proc. of the International Congress of Phonetic Sciences*, Barcelona, August 2003.

[5] D. Caseiro, H. Meinedo, A. Serralheiso, I. Trancoso, and J. Neto, "Spoken book alignment using WFSTs." In *Proceedings of the Human Language Technology Conference*, San Diego, March 2002.

[6] J. Garofolo, C. Auzanne, E. Voorhees, and W. Fisher, "1999 TREC-8 spoken document retrieval track overview and results." In *Proceedings of the 8th Text Retrieval Conference (TREC-8)*, Gaithersburg, MD, November 1999.

[7] J. Glass, "A probabilistic framework for segment-based speech recognition." *Computer, Speech, and Language*, vol. 17, no. 2-3, pp. 137–152, April-July 2003.

[8] Information Technology Support Center, "Appeals transcript - speech recognition feasibility study." Technical report sponsored by the US Department of Labor, December 2004.

[9] L. Lamel, J. Gauvain, and G. Adda, "Lightly supervised and unsupervised acoustic model training." *Computer Speech and Language*, vol. 16, no. 1, pp. 115-129, January 2002.

[10] P. Moreno, C. Joerg, J.-M. Van Thong, and O. Glickman, "A recursive algorithm for the forced alignment of very long audio segments." In *Proc. of ICSLP*, Sydney, Australia, December 1998.

[11] A. Park, T. J. Hazen, and J. Glass, "Automatic processing of audio lectures for information retrieval: Vocabulary selection and language modeling." In *Proc. of ICASSP*, Philadelphia, March 2005.

[12] A. Stolcke, "SRILM - An extensible language modeling toolkit." In *Proc. of ICSLP*, Denver, CO, September 2002.

[13] S. Strassel and M. Glennky, "Shared linguistic resources for human language technology in the meeting domain." In *Proc. of the ICASSP 2004 Meeting Recognition Workshop*, Montreal, May 2004.