Improving and aligning speech with presentation slides

Ranjini Swaminathan
University of Arizona
ranjini@cs.arizona.edu

Michael E. Thompson
University of Arizona
thompy@cs.arizona.edu

Sandiway Fong
University of Arizona
sandiway@cs.arizona.edu

Alon Efrat
University of Arizona
alon@cs.arizona.edu

Arnon Amir
IBM Almaden
arnon@almaden.ibm.com

Kobus Barnard
University of Arizona
kobus@cs.arizona.edu

Abstract

We present a novel method to correct automatically generated speech transcripts of talks and lecture videos using text from accompanying presentation slides. The approach finesses the challenges of dealing with technical terms which are often outside the vocabulary of speech recognizers. Further, we align the transcript to the slide word sequence so that we can improve the organization of closed captioning for hearing impaired users, and improve automatic highlighting or magnification for visually impaired users.

For each speech segment associated with a slide, we construct a sequential Hidden Markov Model for the observed phonemes that follows slide word order, interspersed with text not on the slide. Incongruence between slide words and mistaken transcript words is accounted for using phoneme confusion probabilities. Hence, transcript words different from aligned high probability slide words can be corrected.

Experiments on six talks show improvement in transcript accuracy and alignment with slide words.

1. Introduction

Effective browsing and searching of large video collections is an important problem with many remaining challenges. We suggest that presentation slides associated with most lectures and research talks provide promising methods for indexing the fine grained semantic content available in these collections. For example, SLIC (Semantically Linked Instructional Content) [4] uses an automated approach [8, 7] to link images of presentation slides to video frames, thereby segmenting videos into semantic chunks based on slide use, and indexing them using slides. Slide words have also been aligned with the accompanying transcript automatically to find slide boundaries using curve fitting techniques in [6]. Alternatively, slide alignment can also be achieved at capture time using a variety of mechanical methods as is becoming common [1].

We would like to extend the capabilities of systems like the SLIC system by integrating transcripts generated by Automatic Speech Recognition systems (ASR). As has been developed by others, speech phonemes as well as speech transcripts e.g., [10, 5] can provide usable indexing into video for retrieval. While state of the art speech recognition systems exhibit impressive performance, difficulties remain when integrating them into a system like SLIC that hosts a varied selection of technical lectures by multiple speakers. General purpose ASRs depend on the training vocabulary and our unanticipated technical vocabularies lead to difficulties. Hence, we propose an algorithm to overcome some of these drawbacks and more specifically to:

1. Improve the accuracy of the transcripts generated by an ASR using the accompanying slides.
2. Align the transcripts with the corresponding slide words, thereby identifying when the speaker was likely referring to that part of the slide.

Slide words tend to have disproportionately many words from the subject specific vocabulary which coincidentally are the most important words to get right. Hence using the slide words to correct these errors is promising for improving transcripts and for better closed captioning. Further, propagating these corrections to instances where they are used without slides provides better indexing of important terms.

Aligning the transcripts allows index words from the slides to point more accurately to where they are used in the video. Further, being able to break close caption text by slide elements improves readability for hearing impaired users. Finally, alignment enables automatically

1051-4651/10 $26.00 © 2010 IEEE
DOI 10.1109/ICPR.2010.802

high-lighting and/or magnification of slide text for visually impaired users.

2. Aligning slide words and transcripts

We assume that a video with presentation slides has been segmented into chunks corresponding to each slide. As mentioned above, this can be done automatically, and is also supported at capture time by various systems. Further we assume that slide words have been extracted in the order that they would be read, from the accompanying PowerPoint or PDF slides automatically. Hence our input data consists of collections of ordered words corresponding to slides as used in the presentation and corresponding short audio transcripts.

Our transcript correction algorithm exploits the simple observation that speakers who use slides very often use the words in rough sequence before breaking off to elaborate on the slide content. Hence our approach is to build, for each slide, a Hidden Markov Model (HMM)[9] that explains the transcript as a sequence of phonemes being noisily emitted from slide words and interspersed other words. More specifically, the dynamically built model has phoneme emitting states that are derived from the slide word sequence, interspersed with states for additional phonemes not corresponding to slide words (see Figure 1). We fit this model to find the most likely sequence which both aligns the transcript with slide words, and points to correction sites.

Sequential model details. Each transcript segment is modeled as containing one or more of the corresponding slide words SW_i in sequence interspersed with some non-slide words NSW_i. We characterize the states in the HMM as slide-word phoneme states and non-slide word phoneme states as shown in Figure1. Transitions from a slide word are possible to any succeeding slide word and the immediate next non-slide word state. Transitions from a non-slide word state can be to any succeeding non-slide word state or to the immediate next slide word state. This set of transitions thus accounts for out of order words as well as consecutive sequences of slide words. A non-slide word state is also allowed to transition to itself, which permits one or more non-slide words of different lengths between consecutive slide word states. For instance, the sequence $\{SW_0, NSW, NSW, SW_4\}$ can be achieved by the following transitions: $SW_0 \rightarrow NSW_1 \rightarrow NSW_3 \rightarrow SW_4$.

We model the transitions between two slide word or non-slide word states as Poisson distributions with parameters $\lambda_{sw,sw}$ and $\lambda_{nsw,nsw}$ respectively which are the Maximum Likelihood Estimators computed from the training data as shown in equation 1.

$$t_{p_{sw_{i,n}},p_{sw_{j,1}}} = \frac{\lambda_{sw_i,sw_i}{}^{(j-i)} e^{-\lambda_{sw,sw}}}{(j-i)!}, j > i \quad (1)$$

$$t_{p_{sw_{i,k}},psw_{i,k+1}} = 1$$

$$t_{p_{nsw_i},p_{nsw_j}} = \frac{\lambda_{nsw_i,nsw_i}{}^{(j-i)} e^{-\lambda_{nsw,nsw}}}{(j-i)!}, j > i$$

where $p_{sw_{i,n}}$ is the last phoneme in the slide word i and $p_{sw_{j,1}}$ is the first phoneme in the slide word j. The self transitioning probability of a non-slide word phoneme is computed as the average length of a phoneme sequence between slide words.

Every state linked to a phoneme has an observation probability associated with every possible observed phoneme. These phoneme confusion probabilities tell us, for example, how likely it is that the ASR recognizes phoneme *ch* as *ch* itself or as *sh*. These are the observation probabilities for our HMM which are computed from the training data by aligning the ASR generated transcript words with the ground truth using dynamic programming. The unmatched words between the matched words in the two sequences are then decomposed into their phonemes and these phoneme sub-sequences are once again aligned. The probability of seeing each unmatched ground truth phoneme is now distributed over all the corresponding unmatched phonemes from a sub-sequence in the automatically generated transcript.

Inference for transcription correction. For every slide under consideration, we use the HMM to predict the Viterbi path [9] which is the most likely sequence of phonemes to have generated the transcript. We then use this path to determine the errors in the transcript and the appropriate slide words to replace them with. An error in the transcript is a word(s) that is recognized inaccurately by the ASR and almost always sounds very similar to the actual word(s) spoken in the talk. For instance, in a talk about crops grown in South America, the word *chia*, a crop grown in that region, is often recognized as *shiya*, a word that has absolutely no relevance to the talk. If we break the words up into their constituent building blocks or phonemes, we see in our example, that *chia:ch-iy-ah* becomes *shiya:sh-iy-ah* in the transcript - a case where one phoneme was replaced by a very similar sounding but different phoneme. By breaking up the words in the transcript into phonemes, we transform the problem of correction into one where a group of phonemes in the audio or speech is regrouped or replaced by similar sounding phonemes in the transcript.

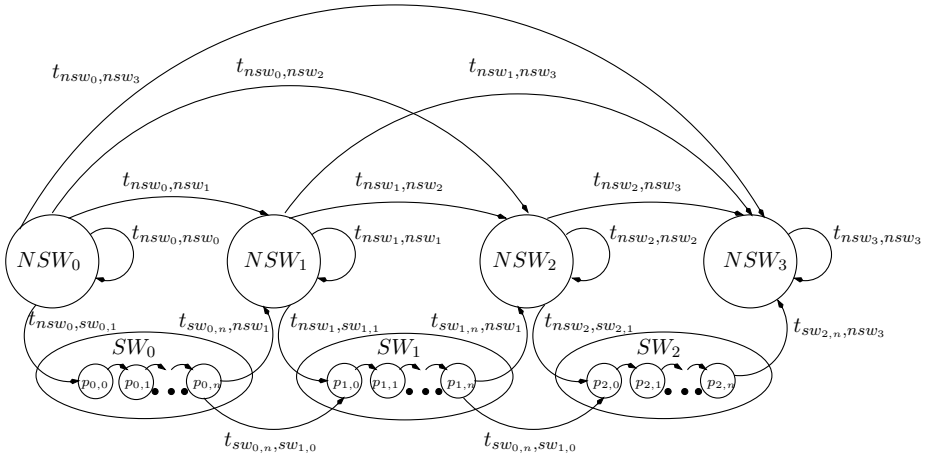

Figure 1. Hidden Markov Model for transcript correction. Sequential transitions are allowed from a slide word state SW_i to subsequent slide word states SW_j or the immediately following non-slide word state NSW_{i+1}. Similarly, transitions from a non-slide word state NSW_i are permitted to the immediately following slide word state SW_i, subsequent non-slide word states NSW_j, or to itself. Slide word to slide word and non-slide word to non-slide word transitions are modeled as Poisson distributions.

3. Evaluation and Results

Our motivation for using a sequential model for transcript correction is both to correct the errors in the transcript as well as to index the transcript with the slide words. We therefore evaluate the performance of our algorithm using two measures: **1) Accuracy score** or the number of words that are correct in the transcript. We align the transcript with the ground truth data and count the number of words that match. We consider different forms of the same word {plurals, -ing, -ed} to be strictly different. Ideally we would like to restrict our evaluation to a subset of the slide words which likely make a significant difference to understanding the talk and make good indexes. Here, we obtained this set of words by removing all stop words or short common words such as prepositions and articles from the list of slide words. **2) Alignment score** is the number of words on the slide that can be aligned with the transcript.

We experimented with six talks from the SLIC database. All the talks are about 75 minutes in length and have about 45 slides. We discarded slides with only images, tables, or figures. We used the state of the art IBM Hosted Transcription Service [3] to generate the speech transcripts for all the talks. The *n* slides from a talk were partitioned into *n* subsets with one slide as test data and the rest as training data in each subset. After computing the transition and observation probabilities using the training data, we used the publicly available Matlab HMM toolbox [2] to compute the Viterbi path of phonemes that is most likely to have generated the test phoneme sequence. Using this phoneme path sequence we determined the positions of words in the transcript that closely resemble the words from the slide and the corresponding slide words to replace them with.

Table 1 shows the accuracy counts computed for the six talks. The alignment algorithm predicts the position of a slide word in the ASR generated transcript, which limits the number of possible corrections of multiple instances of the same error to the number of times the word occurs on the slide. This is addressed with correction propagation below.

We see that accuracy improved significantly in talks T1 and T3 where there were a large number of words outside the vocabulary of the ASR, somewhat in talk T4, and not at all for the others which had many commonly used words. The presence of many such common words resulted in less than optimal alignments for the correction.

Table 1. Word accuracy counts for different talks, for the ASR generated transcript (first row); the corrected transcript (CT, second row); and the transcript with corrections propagated (CP, third row). The numbers in parentheses are the subset of words in the evaluation set

Talk	T1 (210)	T2 (140)	T3 (207)	T4 (126)	T5 (118)	T6 (290)
ASR	141	116	148	61	81	160
CT	155	116	159	64	79	157
CP	163	116	162	60	80	162

Table 2. Word alignment scores for different talks, with the number of slide words aligned with the ASR generated transcript (first row) and the corrected transcript (CT, second row). The numbers in parentheses are the total number of slide words in the talks.

Talk	T1 (386)	T2 (225)	T3 (323)	T4 (156)	T5 (191)	T6 (424)
ASR	180	145	135	39	72	103
CT	216	150	167	119	76	134

In the case of talk T6, we see that the ASR does less well in absolute accuracy compared to the other talks, and that the correction algorithm does not improve matters. This talk uses many abbreviated forms of organization names such as USGS and NSF which get mistaken for phonemes coming from stop words.

To further improve transcript accuracy using what we learned from the HMM output, we propagated the replacements suggested by the correction algorithm to other parts of the talk. This made it possible to correct multiple instances of the same error within a transcript segment as well as in other transcript segments. Our error propagation scheme finds the error patterns in the transcript and replaces them with the corresponding slide words. We remove stop words from our list of slide words since some of the smaller stop word phonemes are very frequently found in the transcript as a part of larger words. Propagating corrections improves the accuracy over the corrected transcripts in three talks - T1, T3 and T5. In the case of talk T2, there were no other instances of the error patterns that could be replaced and propagation did not change the results. Talk T4 had one slide where the propagated changes introduced a good many errors which brought down the net accuracy. A manual inspection of the corrections revealed that including some acceptable alternative forms of the words being evaluated, will further increase the accuracy counts.

Table 2 shows the alignment scores, or the number of slide words aligned correctly with words in the ASR generated transcript and the corrected transcript. Propagated changes are not relevant here. We see that the corrected transcript show significant improvement in alignment for all the talks in our data set.

One reason why alignment can work well even if we are not able to correct many words is as follows. We have noticed that the sequential model sometimes replaces a correct word in the transcript with a different form of the same word from the slide. For example, *foods* in the transcript, might get replaced by *food* from the slide since they have similar phoneme patterns. If the ground truth has the word *foods* in the corresponding position, then this actually introduces an error in the transcript but benefits the alignment. This suggests that sensibly accounting for different forms of a word may be very helpful. However, this is a challenge in itself and we do not address that issue in this paper. However, such (mis)replacements do not take away much from the readability of the transcript and at the same time improve the alignment with the slide words.

4. Conclusions and Future Work

We show that it is feasible to correct speech transcripts using slide words, despite the limited overlap of words between the two sources. We achieve this using an alignment approach that can absorb the many spurious words. The method is effective because many of the correctable words are both important and difficult for speech recognizers. Further, the alignment of these two disparate data streams is beneficial for improving access to educational video, especially for users with disabilities. Future work will include testing different transition models depending on whether one is within a slide fragment (e.g., bullet point), integrating simple language models to support choosing between alternative forms of multiple proposed correction words, and using laser pointer detection as a further cue as to which part of the slide the speaker is focused on.

References

[1] Camtasia Studio. `http://www.techsmith.com/camtasia.asp`.

[2] Hidden Markov Model toolbox for matlab. `people.cs.ubc.ca/~murphyk/Software/HMM/hmm.html`.

[3] IBM Hosted Transcription Service (build 08/05/2009). `antemural.watson.ibm.com/SLWeb/`.

[4] The SLIC browsing system. `slic.arizona.edu`.

[5] YouTube. `www.youtube.com`.

[6] Y. Chen and W. J. Heng. Automatic synchronization of speech transcript and slides in presentation. In *Proceedings of ISCAS Vol. II, pp. 568-571*, 2003.

[7] Q. Fan, A. Amir, K. Barnard, R. Swaminathan, and A. Efrat. Temporal modeling of slide change in presentation videos. In *ICASSP*, 2007.

[8] Q. Fan, K. Barnard, A. Amir, A. Efrat, and M. Lin. Matching slides to presentation videos using sift and scene background matching. In *Proceedings of 8th ACM SIGMM International Workshop on MIR*, 2006.

[9] L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. In *Proceedings of the IEEE, vol77, no2,*, pages 257–287, 1989.

[10] S. Srinivasan and D. Petkovic. Phonetic confusion matrix based spoken document retrieval. In *Proceedings of ACM SIGIR.*, pages 81–87. ACM, 2000.