Online Sequence Alignment for Real-Time Audio Transcription by Non-Experts

Walter S. Lasecki, Christopher D. Miller, Donato Borrello and **Jeffrey P. Bigham**
University of Rochester Department of Computer Science
160 Trustee Rd, Rochester, NY. 14627
{wlasecki, jpbigham}@cs.rochester.edu, {c.miller,donato.borrello}@rochester.edu

Abstract

Real-time transcription provides deaf and hard of hearing people visual access to spoken content, such as classroom instruction, and other live events. Currently, the only reliable source of real-time transcriptions are expensive, highly-trained experts who are able to keep up with speaking rates. Automatic speech recognition is cheaper but produces too many errors in realistic settings. We introduce a new approach in which partial captions from multiple non-experts are combined to produce a high-quality transcription in real-time. We demonstrate the potential of this approach with data collected from 20 non-expert captionists.

Introduction

Real-time speech transcription is necessary to provide access to mainstream classrooms and live events for deaf and hard-of-hearing (DHH) people. While visual access to spoken material can be achieved through sign language interpreters, many DHH people do not know sign language. Captioning can also be more accurate in many domains because it does not involve transliterating to another language, but instead transcribing an aural representation to a written one.

Real-time transcription is currently limited by the cost and availability of professional captionists, and the quality of automatic speech recognition (ASR). Professional captionists are trained to type at natural speaking rates with few errors using stenographer keyboards. These individuals are paid up to $200/hour (Wald 2006), and must be scheduled in advance, for time blocks of 1 hour. ASR generally works quickly and cheaply, but has high error rates. If the speaker has trained the ASR and wears a high-quality, noise-canceling microphone, the accuracy can be above 90%. When recording a speaker using a standard microphone on an untrained ASR, accuracy rates plummet to far below 50%. Additionally, errors made by ASR often alter meaning since they lack human understanding of the context. We explore using groups of non-expert captionists to transcribe speech in real-time. Our hypothesis is that while workers will not be able to type everything they hear, their partial captions can be combined to produce an accurate transcript.

A Crowd of Captionists

We define *the crowd* as a dynamic pool of workers available on-demand. These workers vary in reliability, but we assume that workers do not try to reduce the quality of results. Workers can be recruited from marketplaces such as Mechanical Turk, or other sets of willing participants. Our approach only requires workers to be able to hear and type.

People are able to understand spoken language with relative ease, but most lack the ability to record it at sufficient speed and accuracy. An average person can type roughly 38-40 words per minute(wpm) , but the average English speaker speaks at around 150 wpm (Pashek and Brookshire 1982). Thus, it is unlikely to get a complete transcription using only a single worker. Since partial captions are additive, multiple workers, each captioning different words in the speech, can exceed the quality of any individual worker. We are investigating how workers can be encouraged to transcribe disjoint segments of an audio stream, but in this paper, our tests only rely on naturally occurring differences between workers.

ASR works well under ideal conditions, but degrades quickly in many real-world settings. Current ASR systems are speaker-dependent, have difficulty recognizing domain-specific jargon, and adapt poorly to changes, such as when the speaker has a cold (Elliot et al. 2008). Speakers will make mistakes, speak unclearly or away from the mic, use unfamiliar terms, and otherwise make it difficult to hear certain parts of the speech. Unlike ASR, people have the advantage of understanding the context the word was spoken in. This makes people less likely to mistake a word for another that does not fit the current context. ASR is also unlikely to recognize specialized terms, or those defined during the presentation, whereas people may have prior knowledge of the topic, and can learn new terms on-the-fly.

Our approach uses the crowd, but can also leverage the speed and affordability of ASR systems by using them as additional workers. Thus, as ASR improves, our system can rely more on it and less on humans – further reducing the cost and increasing the availability of transcription.

Background

As workers input partial captions, they will be merged into a single stream of output for DHH users. This is difficult since the ordering of received captions between workers does not accurately reflect the true ordering due to varied typing and connection speeds. Models must also be able to correct for mistakes made by both human and ASR workers.

The problem of regenerating the original content of continuous speech from a set of n workers can be seen as an instance of the general problem of Multiple Sequence Alignment (MSA). While this problem can be solved with a dy-

namic programming algorithm, the time and space complexity is exponential in the number and length of sequences being combined (n workers submitting k words in our case). This complexity means that existing MSA algorithms alone are unlikely to be able to solve this problem in real-time. MSA also cannot align ambiguously ordered words, thus requires a level of coverage that eliminates uncertainty.

Real-time Input Combiner

We present a dependency-graph model to track word ordering and remove ambiguity. The graph is composed of elements that each store a set of equivalent words. These elements track the lowest timestamp of a word and the most common spelling. We assume that each worker will provide captions in the correct order. When a input is received, a lookup table is used to find the best fitting element (based on time stamp) that occurs as a descendant of the previous word input by the same worker. If no matching element is found, a new element is created, the word is added, and a link between the new element and the element containing the last word entered by the user is added. Finally, the graph is updated to ensure only the longest path between each pair of elements is maintained. The graph can then use statistical data to merge the branches in the graph back together to form the caption. To prevent unbounded growth, we prune elements with timestamps older than 15 seconds from the actively updating graph and write them to a permanent transcript. This graph thus allows new input to be added incrementally, with updates taking less than 2ms on average.

We analyzed worker input streams and found many submit semantically equivalent captions that inadvertently differ from other workers. Our data showed that differences were often the result of writing style, use of contractions , or simple misspellings. To account for this, we use a set of rules aimed at homogenizing the input without altering meaning. We use aspell (aspell.net) to correct misspellings, and use a simple filter for common abbreviations and contractions.

Experiments

We recruited 20 undergraduate students to act as non-expert captionists. These students had no special training, or previous formal experience transcribing audio. We asked participants to transcribe four three-minute audio clips from MIT OpenCourseware lectures (ocw.mit.edu). These inputs were aligned offline with an expert-transcribed baseline using the Needleman-Wunsch dynamic sequence alignment algorithm. We compared workers with Nuance's Dragon Dictate 2 ASR on three measures: (i) *coverage*, the number of words spoken that were transcribed by some worker, (ii) *accuracy*, the number of words entered by workers that corresponded to a spoken word, and (iii) *latency*, the average time taken for some worker to input a correct caption.

Our results show that workers can outperform ASR, and that more workers lead to better results. The average latency across these tasks was 3.87 seconds which we expect to be sufficient for real-time captioning. Tests with our combiner show that both the coverage and accuracy can be improved over that of a single worker. Since our approach can filter

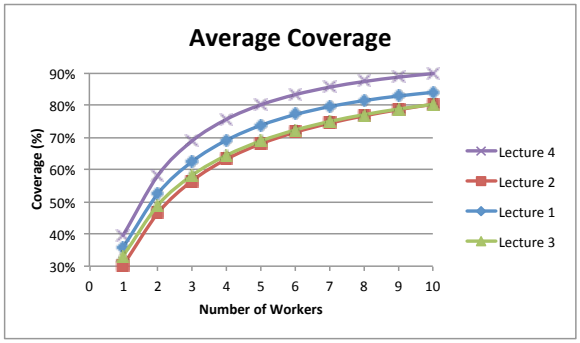

Figure 1: Average coverage for groups of 1-10 workers on four different classroom audio clips. ASR averaged 17.99%

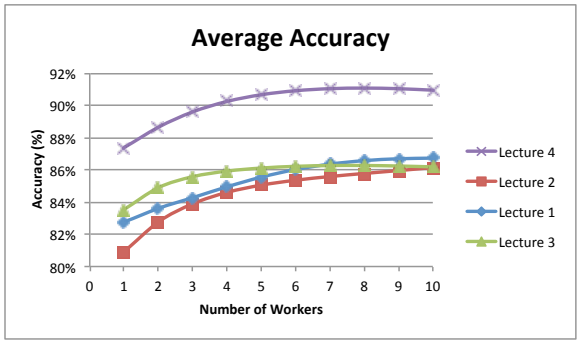

Figure 2: Average accuracy for groups of 1-10 workers on four different classroom audio clips. ASR averaged 43.94%

worker input, we can achieve higher accuracy than groups of workers using alignment. In one example with 5 workers, we averaged 94% accuracy compares to 86% using alignment.

Conclusion and Future Work

We demonstrated the potential for groups to outperform both individual workers and ASR in real-time transcription in terms of coverage and accuracy and introduced a new approach for aligning partial captions to create high-quality transcription on-the-fly. Since coverage is bounded by the union of the the workers' input, we are investigating interfaces that encourage workers to each caption different parts of the audio – improving the chances that someone will type each word. It also may be possible to use separate crowds to correct captions in real-time to further improve quality.

References

Elliot, L. B.; Stinson, M. S.; Easton, D.; and Bourgeois, J. 2008. College Students Learning With C-Print's Education Software and Automatic Speech Recognition. In *American Educational Research Association Annual Meeting*.

Pashek, G. V., and Brookshire, R. H. 1982. Effects of rate of speech and linguistic stress on auditory paragraph comprehension of aphasic individuals. *Journal of Speech and Hearing Research* 25:377–383.

Wald, M. 2006. Creating accessible educational multimedia through editing automatic speech recognition captioning in real time. *Interactive Technology and Smart Education* 3(2):131–141.