Utilizing Automatic Speech Recognition to Improve Deaf Accessibility on the Web

Brent Shiver
DePaul University
bshiver@cs.depaul.edu

Abstract

Internet technologies have expanded rapidly over the past two decades, making information of all sorts more readily available. It has created a shift from paper-based media toward electronic media, e-mail, online news, and online videos. Not only they are more cost-effective than traditional media, these new media have contributed to quality and convenience. However, proliferation of video and audio media on the internet creates an inadvertent disadvantage for deaf internet users. Despite technological and legislative milestones in recent decades in making television and movies more accessible, there has been little progress with online access. A major obstacle to providing captions for internet media is the high cost of captioning and transcribing services. It is virtually impossible to manually caption every single video or audio clip on internet due to the staggering cost. A possible alternative lies in automatic speech recognition (ASR). This paper investigates possible solutions to Web accessibility through utilization of ASR technologies. It surveys previous studies that employ visualization and ASR to determine their effectiveness in the context of deaf accessibility.

1 Introduction

Internet technologies have expanded rapidly over the past two decades, making information of all sorts more readily available. It has created a shift from traditional media such as mail, newspapers, books, television and movies toward e-mail, online news, e-books, and online videos. Not only they are more cost-effective than traditional media, these new media have contributed to quality and convenience. However, proliferation of video and audio media on the internet creates an inadvertent disadvantage for deaf internet users.

Despite technological and legislative milestones in recent decades in making television and movies more accessible, there has been little progress with online access. A major obstacle to providing captions for internet media is the high cost of captioning and transcribing services. A recent development which has the potential for positive change is the passage of the Twenty-first Century Communications and Video Accessibility Act of 2010, which was signed into law by President Obama [1], Part of the law's purpose is to help make online more accessible for the deaf. However, even with help from the legislative front, it is virtually impossible to manually caption every single video or audio clip on internet due to the staggering cost. A possible alternative lies in automatic speech recognition (ASR). This paper investigates possible solutions to Web accessibility through utilization of ASR technologies. It surveys previous studies that employ visualization and automatic speech recognition to determine their effectiveness in the context of deaf accessibility.

2 Challenges of Captioning on the Web

The major challenge of manually captioning is cost. Captioning a video costs approximately $10 - $30 per minute [2]. A cost-effective alternative to manual captioning is ASR technologies. The term describes systems that translate audio content to text material. Applications of ASR include uses in the military and healthcare, as well as automated call centers and for people with mobile challenges. Google recently added the capability for users to upload English transcripts without time codes and its ASR technology would be applied to synchronize the captions with videos [3]. Ken Harrenstien, a deaf engineer at Google involved with the project believes this feature is a major milestone that could open doors to more accessibility on the Internet, but acknowledges the accuracy issues that are inherent with ASR technology that tries to understand speakers from various backgrounds, as will be discussed in the next section.

The first speech recognizer, developed in 1952 by Davis, Biddulph, and Balashek of Bell Laboratories, appeared that identified single spoken digits [4]. At present, there are two scenarios that permit speech recognition technology to maintain a word error rate (WER) low enough that the results are useful. A large vocabulary that covers a breadth of topics necessitates training by an individual speaker. To be speaker independent, an ASR system has to severely limit the size of the vocabulary. Unfortunately, the task of making web-based media more deaf-friendly requires both speaker independence, since it will need to accommodate speakers in all media, and, since the media are not limited to specific topics, it will need to recognize a large vocabulary. An ASR technology that translates speech into text for better deaf accessibility to the Web cannot have restrictions on either speaker or vocabulary, because Web media contain audio information for thousands of speakers on thousands of topics. Unfortunately, accomplishing the goal of a speaker-independent ASR capable of recognizing a large vocabulary has continued to be a herculean task.

Since 1996 National Institute of Standards and Technology (NIST) has been inviting researchers from companies and universities to participate in the Speaker Recognition Evaluation (SRE) every 1-2 years [5]. The goal is to establish benchmarks and measure progress over time on systems that support large vocabulary without speaker training. Over 40 research sites from all over the world, including Massachusetts Institute of Technology, Carnegie Mellon University, and IBM, have participated and contributed to the trials [6]. Despite collective and collaborative efforts, results are still far from accomplishing a consistent 2-4% WER that is considered within range of typical human error in transcription. According to NIST STT Benchmark Test History graph, the best system could only maintain a WER of ten percent while many state-of-art systems, covering other speech tasks, have much higher WERs.

Despite advancements in ASR technologies, they are geared toward customers who have the benefit of being able to hear. While the software produces the transcriptions, the hearing users are able to catch errors while listening to own self or the recording, and then are able to make corrections as needed. Unfortunately, deaf users do not have this benefit and will not know whether the transcribed text is accurate or incorrect, thus making it difficult or nearly impossible to trust the veracity of the transcription. However, there is additional information that could be a valuable tool. All ASR software use probabilities or confidence levels to determine translations, but they are discarded and users do not have access to them. This has the potential to benefit deaf users because it would include information on which passages are likely to be accurate and which other portions are likely to be incorrectly transcribed. This data would help add context to the translated text and empower the users to make decisions.

3 Speech Recognition Visualization

At times, when processing sound, speech recognition software may not be able to identify the words. This may be due to a speaker talking too rapidly or using an atypical or unexpected word. In these cases, there may be multiple interpretations of the word being spoken. Deaf users do not have the option of reviewing the recorded speech and checking it against the recognized text for ambiguities or errors. Typical commercial speech recognition software does not indicate that an ambiguity exists through visual means such as annotated text or listing possible translation alternatives. Despite enormous potential benefits to deaf population, the available literature that focuses on ASR and deafness is scarce. Virtually every study that evaluates visualization strategies of text created via ASR involves hearing users.

Vertanen & Kristensson (2008) investigated possible benefits for hearing users of employing an underlying visualization to emphasize low-confidence. This approach was created in the hope of lowering the cost of creating transcripts for spoken speech. An initial transcript was captured through ASR, and a human editor then read and corrected errors in the transcript while listening to the recorded speech. The goal of the approach was to help the editor to catch more errors in a time-effective manner. Figure 1 shows how visualization techniques were employed to indicate the speech recognition engine's confidence in the produced text . Red underlines indicated words with low confidence, and the darkness of the underlines was proportional to the lowness of the confidence. This visualization helped users identify potential errors in only a limited number of cases. The users would catch errors more often only if low-confidence text were correctly flagged. On the other hand, if text was incorrectly identified and not underlined as low-confidence, chances were greater that the users would miss the problem. The authors concluded that it was possible that the users placed too much faith in recognizer's ability to present annotations accurately.

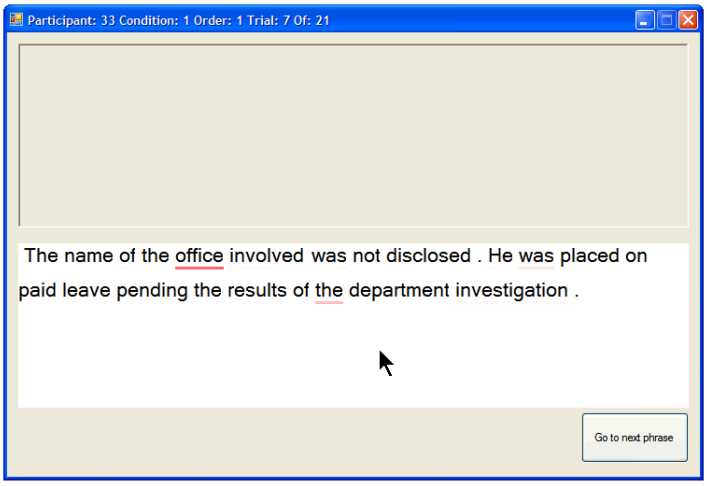

Figure 1. Shades of red underline is applied to words with low-confidence. The word "office" has a lower confidence than the word "was" that appears in the second sentence.

A case study conducted by Collins, Penn, & Carpendale (2007) focused on uncertainty visualization through utilization of lattices to support decision-making which involved a multilingual chat application that used an automatic translation engine. The goal was to provide possible choices through visualization and empower users to choose a translation that makes most sense or discard it altogether. Lattices were generated as representation of possible

translations and included confidence levels through fill color and border thickness. Although the study involved spoken language translation, it has a close resemblance to automatic speech recognition. Although user testing was not performed, they collected informal user feedback. The participants expressed general interest in the visualization of uncertainty so they can make appropriate decisions.

4 Utilizing Speech Recognition to Aid Comprehension

Wald (2006) explored the possibility of utilizing Automatic Speech Recognition (ASR) to aid classroom learning specifically for students with disabilities including deaf, hard-of-hearing, blind, and dyslexic. This is the only extant study that explores any aspect of ASR visualization for increased deaf accessibility. It also investigated the benefits of using ASR to enhance quality of learning and teaching for students without disabilities. It found that one of the problems with real-time speech-to-text synthesis was a lack of punctuation. Without punctuation, the ASR-created transcripts were difficult to read and understand. A workaround was to add single and double spaces to the transcripts as visual cues of brief pauses and long pauses respectively.

An audio browsing study was conducted by Vemuri, et al (2004). It applied time-compression techniques to audio files as constrasted with (Munteau, et al, 2006) who imposed a time limit to complete a quiz. It explored the benefits of visualizing ASR transcripts with varying WERs. The generated transcripts had a WER of 16% to 67% (mean=42%, sd=15%), which was comparable to other commerical ASRs. During the experiment, five different conditions were followed: C1: Perfect transcript; C2: Transcript generated by ASR (using word brightness); C3: Transcript generated by ASR; C4: Completely incorrect transcript; C5: No transcript, audio only. Figure 2 shows the results of the five treatments. The study identified C1 as being the best but costly option and is time-consuming and requires manual intervention. More cost-effective options C2 and C3, which were generated by ASR, were found to do nearly as well as C1. Transcript generated for C2 utilizing word brightness did not show any significant improvement over C3. Finally, as expected, performance under conditions C4 and C5 were the poorest, but interestingly, there was no statistical difference between them. A possible explanation may be that C4's transcript was so bad that the participants ended up ignoring it altogether. The researchers concluded that ASR transcripts improve comprehension when listening to time-compressed speech.

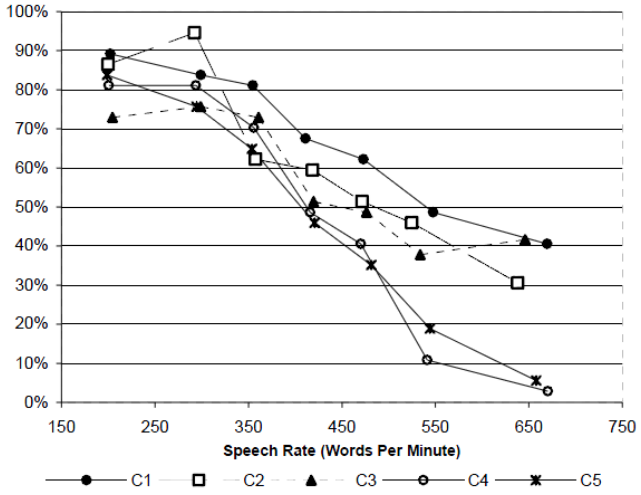

Figure 2. C2 and C3 were shown to do nearly as well as a perfect transcript C1.

5 Discussion

With exception of Wald's efforts, all of the studies placed emphasis on use of ASR transcripts to benefit hearing users. Wald wanted to improve classroom learning for students with disabilities including those with hearing loss. In those studies using ASR for hearing populations, utilizing automated transcripts to retrieve time-compressed audio content and skimming through webcast archives facilitated greater comprehension. The visualization tool offered by Vertanen and Kristensson utilized shaded, red underlines to emphasize words having low-confidence. However, it is reliant on ASR being accurate and when ASR incorrectly recognizes words as having high confidence its effectiveness deteriorates. Although the application created by Collins, Carpendale, and Penn did not involve using ASR, it visualized probabilities and made them available as lattice paths to offer decision-making to users. This approach could be useful for an ASR tool that has the statistics but needs an effective way to output the results. An automatic speech-to-text tool designed specifically for deaf users has the potential to provide a better bridge to audio/video media. Although ASR has accuracy issues, several studies have shown that well-chosen visualizations have the potential to help users to glean additional information from error-laden texts.

6 References

[1] U.S. Congress, S. 3304: Twenty-First Century Communications and Video Accessibility Act of 2010, http://www.govtrack.us/congress/bill.xpd?bill=s111-3304.

[2] Custom Captions, accessed October 20, 2010, http://www.customcaptions.com/#PriceComparisonChart.

[3] J. Sutter. "An engineer's quest to caption the Web," *CNN.com, a*ccessed February 21, 2011, http://articles.cnn.com/2010-02-09/tech/deaf.internet.captions_1_captioning-deaf-people-sign-language?_s=PM:TECH.

[4] B. Juang and L.Rabiner, "Automatic Speech Recognition - A Brief History of the Technology Development," *Elsevier Encyclopedia of Language and Linguistics*, 2nd Edition, 2005.

[5] National Institute of Standards and Technology. Multimodal Information Group - Speaker Recognition. National Institute of Standards and Technology. [Online] 2011. [Cited: March 2, 2011.] http://nist.gov/itl/iad/mig/sre.cfm.

[6] A. Martin and C. Greenberg. "2008 NIST Speaker Recognition Evaluation Slides from Workshop Presentation," June 17-18, 2008.

[7] M. Wald. "An exploration of the potential of Automatic Speech Recognition to assist and enable receptive communication in higher education," *Research in Learning Technology*, 2006, pp. 9-20.

[8] C. Munteanu, R. Baecker, G. Penn, E. Toms, D. James. "The Effect of Speech Recognition Accuracy Rates on the Usefulness and Usability of Webcast Archives," *CHI Proceedings - Visualization and Search*, 2006, pp. 493 - 502.

[9] S. Vemuri, P. DeCamp, W. Bender, C. Schmandt. "Improving Speech Playback Using Time-Compression and Speech Recognition," *CHI Proceedings*, 2004, pp. 295 - 302.