Usable speech recognition: toward improved access to webcast lectures

Cosmin Munteanu
Department of Computer Science
University of Toronto
Toronto, ON, M5S 3G4, Canada
mcosmin@cs.toronto.edu

Gerald Penn
Department of Computer Science &
Knowledge Media Design Institute
University of Toronto
Toronto, ON, M5S 3G4, Canada
gpenn@cs.toronto.edu

Ron Baecker
Department of Computer Science &
Knowledge Media Design Institute
University of Toronto
Toronto, ON, M5S 2E4, Canada
rmb@kmdi.toronto.edu

CHI 2008, April 5 – April 10, 2008, Florence, Italy
Workshops and Courses: Usable Artificial Intelligence.

Abstract
A growing number of lecture webcasts are archived after being delivered live. In the absence of transcripts, users are faced with increased difficulty in performing tasks easily achieved with text documents (retrieval, browsing, skimming). Unfortunately, speech recognition systems do not perform satisfactorily when transcribing lectures. In this paper, we present an overview of the ePresence lecture transcription project, whose goal is to improve the usefulness and usability of automatically-generated transcripts of webcast lectures. We achieve this by integrating novel speech recognition techniques specifically addressed at increasing the accuracy of webcast transcriptions with the development of an interactive collaborative interface that facilitates users' contribution to the improvement of machine-generated transcripts. We conclude by discussing the challenges (and possible solutions) to successfully integrate transcripts into archives of webcast lectures.

Keywords
Webcasting, automatic speech recognition, text transcripts, wiki, field study

ACM Classification Keywords
I2.7 Natural Language Processing, H5.1 Multimedia Information Systems, H5.2 User interfaces

Introduction

Webcasts are increasingly used to deliver information-rich media over the Internet (such as on-line lectures). Most such media are archived after being delivered live, and can be accessed through various interactive systems. Research evidence indicates that transcripts are one of the most needed additions to webcast systems that would facilitate users' information seeking tasks [2]. Unfortunately, the main challenge in integrating text transcripts into archives of webcast lectures is the poor performance of Automatic Speech Recognition (ASR) systems when transcribing lectures.

ePresence (http://epresence.tv) gives users full control of the archive, mainly through the display of the slides used in lectures and a video recording of the lectures themselves, through interaction with a table of contents (containing "chapter" headings and the title of the slides), and through a timeline (a clickable fine-grained time-progress indicator). The ASR-generated transcripts are time-synchronized with the video (current line in boldface); users can also re-synchronize the playback of the video by clicking on the transcript.

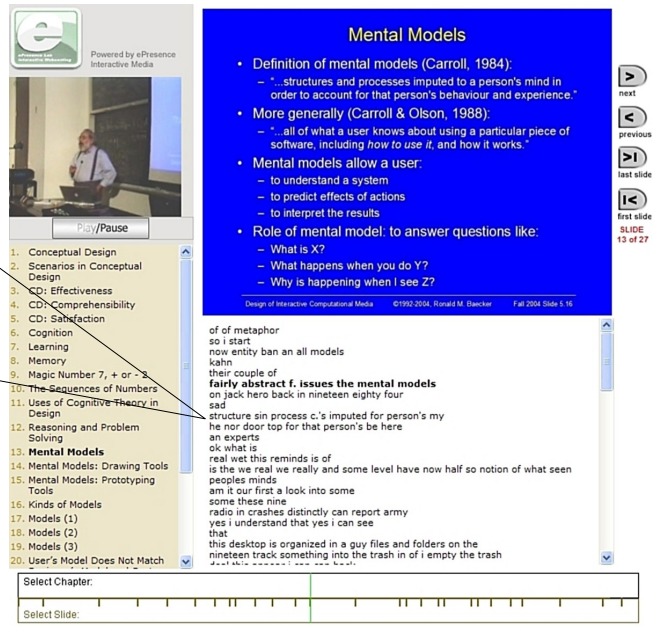

Figure 1: The ePresence webcast system, displaying an archive of a webcast lecture, enhanced with machine-generated transcripts.

Due to adverse acoustic and linguistic characteristics (large vocabulary, speaker independent, continuous speech, imperfect recording conditions), currently-available ASR systems do not perform satisfactorily in domains such as lectures or conference presentations. Most lecture recognition systems achieve Word Error Rates (WER – the edit distance between the automatic and "true" transcripts) of about 40-45% [5] (some reports suggest a 20-30% WER for lectures given in more artificial and better controlled conditions [6]). Moreover, it is expected that such systems will not reach perfect or near-perfect accuracy in the near future [7]. Therefore, in our research we measured the acceptable WER of webcast lecture transcripts, and we developed ASR- and HCI- based solutions to reduce the current WER to desirable values.

Making speech recognition usable – the lecture transcription project

The main focus of the University of Toronto's lecture transcription project is to improve the usefulness and usability of transcript-enhanced webcast archives. We are using as research platform the ePresence system (Figure 1). By exploiting the webcasts' voice channel to improve users' experience in browsing and information searching, we combine novel approaches in Human-Computer Interaction (HCI) and Artificial Intelligence (AI) that would ultimately see transcripts of lectures being integrated into webcast archives.

ASR: how good is good enough?

Through an extensive user study we investigated the user needs for transcription accuracy in webcast archives. For this, we designed a within-subjects study [4], in which 48 participants were exposed to multiple levels of WER in their interaction, in a typical webcast-

use scenario (students writing a quiz based on a lecture). The results showed that users' performance and transcript quality perception is linearly affected by WER, with transcripts of WER equal to or less than 25% being useful and accepted by users of webcast archives. We determined this by assessing users performance in a question-answering task, their perception of transcript quality, as well as users' confidence in their performance and their perceived level of task difficulty. We also found that for most browsing scenarios, users prefer having transcripts even if the quality of those transcripts is less than optimal.

What to do when ASR is not good enough – the AI approach

Spoken Language Processing has long focused on methods and technologies that would ultimately mimic human performance when transcribing speech to text. In this vein, we have worked on improving the accuracy of ASR systems for lectures by building statistical predictive models of language specific to the topic and genre presented. Such models are needed to narrow down the possible choices of words when transcribing speech to text.

One of the common challenges when transcribing lectures is the mismatch between the language used in a lecture and the predictive language models employed by the ASR system. We propose a solution that addresses this issue through a novel information retrieval technique that exploits lecture slides by automatically mining the World Wide Web for documents related to the presented topic, and using these to build a better-matching language model. Such

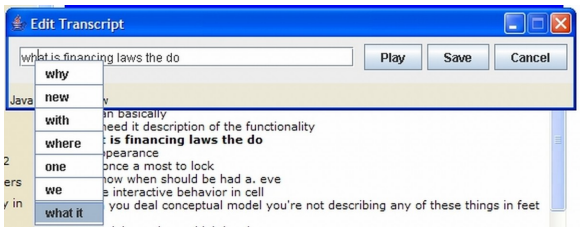

Figure 2: Wiki-like editing of imperfect transcripts.

approach can overcome the need to build two models (a large, topic-independent, one and a smaller, topic-specific, model). This solution achieves a relative WER reduction of 11% [5].

What to do when ASR is not good enough – the HCI approach

Despite improvements to the ASR performance for lecture transcriptions, there still exists a significant gap between the desirable and actual WER. To reduce (and possibly eliminate) this gap, we have turned our attention to HCI – in particular, to Computer-Supported Collaborative Work, a solution often used to overcome limitations of AI systems, such as in [1].

For this, we have developed and evaluated an iterative design of a collaborative tool that extends ePresence's functionality by allowing users to edit and correct the webcast transcripts in a wiki-like manner. The editing tool (Figure 2) is integrated into the archive viewing mode of the webcast system, allowing users to make corrections "on-the-fly" while viewing an archived webcast. The pop-up containing the editing tool is triggered by right-clicking on individual transcript lines. Users can also switch to an extended editing mode that allows for the correction of multiple lines. All edits are instantly reflected in the webcast transcripts.

The editing tool was evaluated iteratively by integrating it with other educational resources available to the students in two Computer Science courses. Our field studies showed that this is a feasible solution for alleviating the errors of ASR webcast lecture transcripts. We have not only analyzed the improvements in transcript quality brought by the editing tool, but also found that wiki-editing is well

received by webcast users. Moreover, our studies revealed that students see the ability to correct transcripts as an enhancement of the classroom experience and are willing to contribute to the improvement of lecture transcripts.

What to do when ASR is not good enough – putting it all together

One of the research questions that remains open is determining the appropriate motivational method for eliciting the necessary user participation that leads to satisfactory correction of webcast transcripts in more general scenarios (particularly outside the classroom environment). However, research evidence suggests that further WER reductions are possible in subsequent lectures when manual transcripts of earlier lectures in a series are used to re-train an ASR system [3]. For this, the corrected transcripts can be employed to address one of the main problems that ASR systems face in large-vocabulary and unconstrained domains (such as lectures): the lack of previously-collected data on the same topic and from the same speaker. These allow for more accurate tuning of all the system's parameters. Therefore, we are currently working on developing self-adaptive ASR training methods to exploit users' partial corrections of early lectures in a course. By pursuing this approach, the user will no longer be a simple recipient of an AI system's output, but rather an active contributor to the system's internal operation.

Conclusions

Improving access to archives of webcast lectures is a task that, by its nature, requires research efforts common to both HCI and AI. By combining research in these areas through integrating novel speech recognition techniques with the development of an interactive collaborative interface, we have shown that the usefulness and usability of automatically-generated transcripts of webcast lectures and presentations can be improved. More work is still needed to determine how to enhance current interfaces with other, more compact, textual projections of webcast lectures, such as automatically-generated summaries or tables of contents.

Acknowledgements

This research is funded by the NSERC Canada Network for Effective Collaboration Technologies through Advanced Research (NECTAR).

References

[1] L. von Ahn and L. Dabbish. Labeling Images With a Computer Game. in *Proc. ACM CHI 2004*, 319–326

[2] C. Dufour, E. Toms, J. Lewis, and R. Baecker. User Strategies for Handling Information Tasks in Webcasts. In *Proc. ACM CHI 2005*, 1343–1346

[3] J. Glass *et al.* Recent Progress in the MIT Spoken Lecture Processing Project. In *Proc. Interspeech - Eurospeech 2007*, 2553–2556

[4] C. Munteanu, R. Baecker, G. Penn, E. Toms, and D. James, The effect of speech recognition accuracy rates on the usefulness and usability of webcast archives, in *Proc. of CHI 2006*, 493-502

[5] C. Munteanu, G. Penn, and R. Baecker. Web-Based Language Modelling for Automatic Lecture Transcription. In *Proc. Interspeech – Eurospeech 2007*, 2353–2356

[6] I. Rogina and T. Schaaf. Lecture and Presentation Tracking in an Intelligent Meeting Room. In *Proc. ACM (IEEE) ICMI 2002*, 47–52

[7] S. Whittaker and J. Hirschberg. Look or Listen: Discovering Effective Techniques for Accessing Speech Data. In *Proc. British HCI 2003*, 253–269