Automatic Speech Recognition for Webcasts: How Good is Good Enough and What to Do When it Isn't

Cosmin Munteanu[1] Gerald Penn[1,2] Ron Baecker[1,2] Yuecheng Zhang[1]
mcosmin@cs.toronto.edu gpenn@cs.toronto.edu rmb@kmdi.toronto.edu jacey.zhang@utoronto.ca
[1]) Department of Computer Science
University of Toronto, Canada
[2]) Knowledge Media Design Institute
University of Toronto, Canada

ABSTRACT

The increased availability of broadband connections has recently led to an increase in the use of Internet broadcasting (webcasting). Most webcasts are archived and accessed numerous times retrospectively. One challenge to skimming and browsing through such archives is the lack of text transcripts of the webcast's audio channel. This paper describes a procedure for prototyping an Automatic Speech Recognition (ASR) system that generates realistic transcripts of any desired Word Error Rate (WER), thus overcoming the drawbacks of both prototype-based and Wizard of Oz simulations. We used such a system in a user study showing that transcripts with WERs less than 25% are acceptable for use in webcast archives. As current ASR systems can only deliver, in realistic conditions, Word Error Rates (WERs) of around 45%, we also describe a solution for reducing the WER of such transcripts by engaging users to collaborate in a "wiki" fashion on editing the imperfect transcripts obtained through ASR.

Categories and Subject Descriptors

I.2.M [**Artificial Intelligence**]: Misc.; H.5.1 [**Multimedia Information Systems**]: Evaluation/methodology; H.5.2 [**User interfaces**]: Natural language,; H.5.3 [**Group and Organization Interfaces**]: Computer-supported Cooperative Work

General Terms

Algorithms, Design, Experimentation, Performance

Keywords

Webcasts, Automatic speech recognition, Collaboration

1. INTRODUCTION

As webcasts become a more common means of broadcasting live events (lectures, presentation, etc.) over the Internet, more of these media are being archived and accessed by users through interactive systems such as ePresence (`http://epresence.tv/`), illustrated in Figure 1, which serves as the framework for our research. Without transcripts, humans are faced with far greater difficulty in performing tasks that are easily achieved with archives of text documents, such as retrieval, browsing, or skimming. Research evidence indicates that transcripts are the most suitable tool for performing tasks that require information-seeking from webcast archives [1].

Currently, due to adverse acoustic and linguistic characteristics (large vocabulary, speaker independent, continuous speech, imperfect recording conditions), ASR systems do not perform satisfactorily in domains such as lectures or conference presentations. Most lecture recognition systems achieve WERs of about 40-45% [2] (some reports suggest a 20-30% WER for lectures given in more artificial and better controlled conditions [3, 4]).

In our research, we have introduced manually and semi-automatically generated transcripts into webcast archives, and investigated how WER influences both users' performance in a question-answering task and their perception of transcript quality (and thus, willingness to accept and use transcripts). We also determined that the minimum level of WER for a transcript to be useful and accepted by users is 25%. For this, we designed an ecologically valid experiment[1], where users performed various tasks using a transcript-enhanced version of ePresence, an extension of the basic functionality of the ePresence system (playback control, slide display, table of content and timeline navigation). While transcripts are fully displayed for a segment of lecture corresponding to one slide, the transcript lines are synchronized with the playback. The lines are also clickable, allowing users to cue the video playback to the corresponding location.

As it is expected that such systems will not reach perfect or near perfect accuracy in the near future [6], we are also proposing alternative tools to reduce current WER levels to the 25% level determined acceptable by our study. For this, we have developed a collaborative tool that extends ePresence functionality by allowing users to edit and correct, in a wiki-like manner, the webcast transcripts. The editing tool is seamlessly integrated into the regular archive viewing mode of ePresence, allowing users to make "on-the-fly" corrections while viewing an archived webcast.

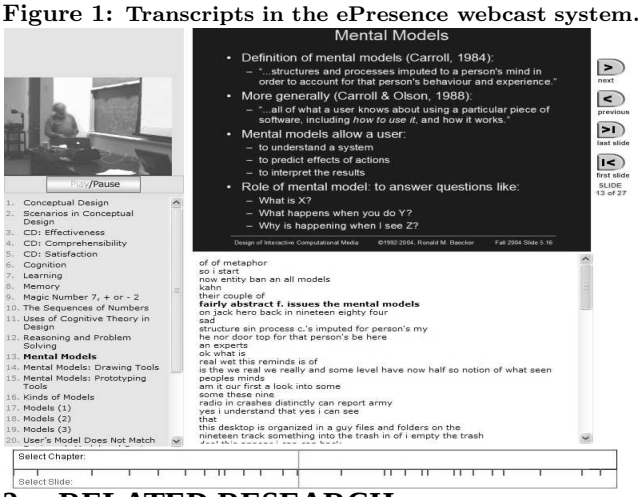
This paper focuses on our method for measuring the acceptable WER of webcast transcripts. We achieve this by combining a procedure for carefully controlling the WER of realistic output within a specially designed Wizard of Oz (WOZ) experimental framework. We then present

[1]A complete description of the study can be found in [5].

ICMI'06 November 2–4, 2006, Banff, Alberta,Canada.

our solution for reducing the currently-achievable WER of lecture webcast transcripts to acceptable levels.

Figure 1: Transcripts in the ePresence webcast system.



2. RELATED RESEARCH

The task of recognizing speaker-independent, large-vocabulary, continuous, and noisy speech is very challenging. While significant effort has been spent on improving speech recognition for lectures and presentations [4, 3, 2, 7], the quality of the transcripts (typically with WERs of around 40%) is still below that of other domains, such as broadcast news transcription. Unfortunately, the research on how humans deal with such error-ridden transcripts and on which accuracy rates can be deemed acceptable is scarce.

Among the few existing studies, one that assessed human ability to use transcripts [8] for news recording retrieval and summarization revealed that users performed better when transcripts accuracy was better. A follow-up study in the context of skimming through voicemail messages [9] showed that users performed their tasks faster when simultaneously browsing speech and text, but performances were lower for improperly transcribed keywords (phone numbers and names). However, users' performance can be improved by providing additional information-mining tools [6].

While these studies provide valuable insights into the users' handling of errorful transcripts, they do not study the relation between performance and WER, nor do they determine the acceptable WER for beneficially including transcripts in browsing interfaces. Therefore, we have conducted a WOZ-like study to determine these relations, as this simulation method is one of the most appropriate for studying natural-language-based human-computer interaction [10]. Although WOZ's drawback resides in the need for a skilled human wizard, this method is preferred (to prototyping), since the cost of building a full-featured natural language prototype is often prohibitive. However, as it will be shown in Section 3.2, our proposed simulation method provides the convenience of WOZ setups while behaving like a true prototype system, with no on-line wizard intervention.

As the currently-achievable WERs for webcast ASR is below the acceptable WER, alternative solutions are needed to reduce this gap. A readily available, though expensive, solution is human intervention. Unfortunately, to our knowledge, no research exists that address the cost of this approach for reducing the WER of transcripts. However, in various other scientific areas, computer-supported collaboration has emerged as an alternative. For example, it was shown in two separate studies ([11], [12]) that the task of indexing and labelling a large collection of images for query-based retrieval can be carried out using web-based collaboration. Collaboration has also been successfully applied to various other tasks, from controlling a mechanical robot over the Internet [13] to open source software development [14]. In Section 4, we will describe our development of a collaborative interface that facilitates the reduction of WER for transcripts of webcast lectures.

3. MEASURING THE ACCEPTABLE WER

3.1 A User Study

We designed a within-subjects study (a complete description is found in [5]) in which 48 participants were exposed to multiple levels of WER in their interaction, in a typical webcast-use scenario (undergraduate students responding to a quiz on a lecture). We assessed the effect of WER[2] on four levels: 0% (manual transcription), 25% (the WER that current ASR systems are able to achieve for broadcast news transcription), 45% (the WER reported in the literature for transcribing lectures and conference talks, in real-life conditions), and no transcripts (the baseline case).

Each participant completed a 12-minute long quiz consisting of five factual questions (no overall lecture comprehension required) for each webcast viewed (one for every level of WER, each on a different 38-minute lecture). Users had full control of the lecture during the quiz. At least two of the five quiz questions did not have the answers on slides and were obscured by transcription errors. We also collected subjective user data trough post-quiz questionnaires: confidence in their own performance, perception of task difficulty, and transcripts' helpfulness.

3.2 ASR System Setup

As we aimed to evaluate user performance at four pre-determined levels of WER, we also wanted to maintain a realistic scenario for the WOZ simulation, as it is recommended for studying natural language-based human-computer interaction [15]. For this, we designed an ASR system that allowed us to control the WER level, by developing language models (LMs) and vocabularies that were over-fit to each lecture. Transcripts of 0% WER were obtained through manual transcription.

To achieve the desired levels of less-than-perfect WERs, the ASR system was built using the SONIC toolkit [16]. Transcripts of 25% and 45% WER were obtained by overfitting models to each lecture (in particular, to segments of lectures containing a variable number of sentences). SONIC's accompanying acoustic model (AM) was used in our experiment. This model is built on 30 hours of data from the Wall Street Journal Dictation Corpus [17], a collection of microphone recordings of news texts read by journalists.

In order to control the overfitting process, the training sentences were mixed with the transcripts of the 1997

[2]The WER of a transcript was computed as the average WERs of the utterances (transcript lines) of length at least 3 words. Most 1 and 2-word lines were just breathing noises or repetitions.

Table 1: The training (overfitting) variables' values for the target WERs of 25% and 45%.

Lecture	1		2		3		4	
Number of sentences in lecture	1280		928		811		972	
Variables / values for WER=	25%	45%	25%	45%	25%	45%	25%	45%
Size (in sentences) of lecture corpus	100	20	200	20	100	20	50	20
Modified lecture sentence lengths	original	5	original	5	original	5	original	5
Number of added HUB-4 sentences	0	650	0	650	0	650	0	650
Modified HUB-4 sentence lengths	-	1	-	1	-	1	-	1

LDC Broadcast News (HUB-4) Corpus [18] Evaluation Set. Although tri-gram LMs were built on the training corpora, further variability was introduced in the training process, by altering the length of the training sentences (this was achieved by concatenating all sentences in the corpus and then splitting them into new sentences of equal length). The pronounciation dictionary (built separately for each lecture corpus) was built to cover all words found in the manual transcription (no out-of-vocabulary items).

The recordings used for our study were collected in a large, amphitheatre-style lecture hall using a head-mounted directional microphone. The recordings were not intrusive, and no alterations to the lecture environment or proceedings were made. The recognition was performed on each set of sentences using the language model that was trained on data consisting of or containing the same set. For an individual lecture, a set of models that produced the desired average WER was chosen, such that all models in that particular set were trained using the same values for the training variables outlined in Table 1. The SONIC decoder performs recognition in two passes, each producing its own hypotheses; since the second pass usually produces an output of a slightly lower WER, we found that the output of the first pass was a better choice for our purpose.

Besides allowing for a greater control of the WER variable, the method we used to generate lecture transcripts ensured that users were exposed to transcripts generated by a real ASR system. Transcripts with these levels of WER as well as no transcript were integrated into the ePresence webcasting system. This setup allowed us to design an ecologically valid experiment as in a WOZ simulation, without the need for a human wizard's on-line intervention.

3.3 User Study: Key Findings

While a complete analysis of the data collected through the user study is presented in [5], we will summarize here the key findings. With respect to quiz scores, our study revealed that transcripts with 25% WER were marginally better than not having transcripts in the webcast system, that WERs of 45% lead to lower quiz scores than no transcripts, and that the overall relation between performance (quiz scores) and WER is linear. We also found that users' confidence in their performance, as well as their perceived level of quiz difficulty, were in the same linear relation with WER as the quiz scores. However, users perceived transcripts as being very helpful roughly the same for manually-generated transcripts as for transcripts with WER of 25%. Participants indicated (through a post-session questionnaire) that they would rather have transcripts with errors than no transcripts and would use such a system for most academic tasks. Navigational features such as a "table of contents" and the ability to playback selected transcript lines were favoured by participants as the most helpful tools to compensate for transcription errors.

4. MANAGING TRANSCRIPTS WITH LESS-THAN-ACCEPTABLE WER

Current ASR systems deliver transcripts of webcast lectures and presentations of 40-45% WER, while the acceptable WER threshold is 25%. To reduce this gap, we have developed a collaborative editing tool that allows users to correct and edit the transcripts. It extends the basic functionality of the system without burdening the user at the same time.

4.1 Wiki-like Editing of Transcripts

During regular playback of a webcast archive, users can right-click on any transcript line (not necessarily the one currently being played back), and an edit box (Figure 2) is displayed, allowing users to make corrections to the selected line. This line becomes highlighted in red, which potentially differentiates it from the current line, which is bold-faced. Besides colour-highlighting, the edit box is popped up on the screen about two transcript lines above the selected line, to maintain a visual connection with the transcript context.

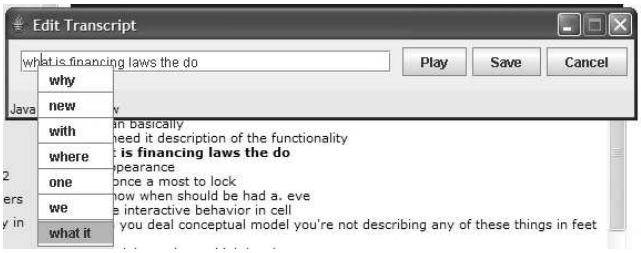


Figure 2: Wiki-like editing of imperfect transcripts

To avoid editing conflicts, a server-side locking mechanism prevents users from simultaneously editing the same line. When trying to edit a locked line, users are informed that the line is being edited by a different user, and that a browser refresh might be needed to update the transcript (webcasts need accurate time synchronization between all components, so regularly checking for transcript updates is not possible).

This on-the-fly editing mode has the advantage of being light-weight on the users – the tool is invisible unless invoked – while at the same time allowing users to carry out corrections to the transcripts without explicitly loading a different interface (the webcast playback is resumed automatically after the edit pop-up is closed).

4.2 The Transcript Editing Tool

The editing tool has several features that facilitate correcting errors in transcripts.

Edit area: allows users to freely make corrections to the displayed transcript line.

Suggestion drop-down: when right-clicking on words in the edit box, a list of possible replacement words is displayed. These are choices under consideration by the ASR system during the recognition process, and extracted

from the word lattices produced by the ASR system – only words that overlap by more than 70% in time alignment with the original word in the lattice are considered as alternatives.

Play button: plays the audio recording corresponding to the selected transcript line, extracted off-line from the original recording (before processing and compression of the streaming video) to ensure optimum quality.

Save: both the transcripts in the webcast window and the originals stored on the webcast server are instantly updated.

Other collaborative features: users can verify the amount of editing work they carried out, quantified as the number of word-level edit actions (deletions, insertions, and substitutions). Also, editing access can be restricted to certain users up to the level of transcripts corresponding to certain slides, which is useful for defining a collaboration model of student lecture transcript editing.

5. CONCLUSIONS, DISCUSSION, AND FUTURE WORK

In this paper, we proposed a procedure for prototyping an ASR system that generates realistic transcripts of any desired WER. Our procedure addresses the drawbacks of the two most common simulation techniques (prototyping and WOZ) used in natural-language-based human-computer interaction studies: it eliminated the need for a skilled human wizard that intervenes in the simulation, while avoiding the costly (sometimes even technologically impossible) solution of prototyping a fully-functional natural language system. Using our WER-controlled ASR system, we conducted a user study which revealed that WER linearly influences both users' task performance and users' perceptions of transcript quality and task difficulty, and that transcripts with a WER equal to or less than 25% were better in all respects than using no transcripts.

Unfortunately, current ASR systems yield error rates of 40-45%, below the determined threshold of usability and usefulness. As a solution to bridging the WER gap, we have developed a collaborative tool that extends the basic functionality of a transcript-enhanced webcast system by engaging users to collaborate in transcript editing and correction for webcast lectures and presentations. This tool seamlessly integrates with the webcast interface and allows for on-the-fly corrections during normal viewing of the archived webcast.

We are currently conducting a user study[3] aimed at quantifying the WER reductions brought directly by the user-performed transcript corrections. We are also using these corrections as a source of ASR re-training and fine-tuning that will further improve transcript quality under the same acoustic conditions. The next stage of the study will address research questions related to wiki-like collaboration, such as how to better motivate students to correct transcripts and how to provide intuitive feedback on other users' edits.

6. ACKNOWLEDGEMENTS

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

[3]In progress by the publishing deadline; we will report on the results during the Conference sessions and on-line at `http://www.cs.toronto.edu/~mcosmin`.

7. REFERENCES

[1] C. Dufour, E. G. Toms, J. Lewis, and R. M. Baecker, "User strategies for handling information tasks in webcasts," in *Proc. of CHI*, 2005.

[2] E. Leeuwis, M. Federico, and M. Cettolo, "Language modeling and transcription of the TED Corpus lectures," in *Proc. of the IEEE ICASSP*, 2003.

[3] I. Rogina and T. Schaaf, "Lecture and presentation tracking in an intelligent meeting room," in *Proc. of IEEE ICMI*, 2000.

[4] K. Kato, H. Nanjo, and T. Kawahara, "Automatic transcription of lecture speech using topic-independent language modeling," in *Proc. of ICSLP*, 2000.

[5] 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.

[6] S. Whittaker and J. Hirschberg, "Look or listen: Discovering effective techniques for accessing speech data," in *Proc. of the Human-Computer Interaction Conference*. 2003, Springer-Verlag.

[7] A. Park, T.J. Hazen, and J.R. Glass, "Automatic processing of audio lectures for information retrieval," in *Proc. of IEEE ICASSP*, 2005.

[8] L. Stark, S. Whittaker, and J. Hirschberg, "ASR satisficing: The effects of ASR accuracy on speech retrieval," in *Proc. of ICSLP*, 2000.

[9] S. Whittaker, J. Hirschberg, B. Amento, L. Stark, M. Bacchiani, P. Isenhour, L. Stead, G. Zamchick, and A. Rosenberg, "SCANMail: a voicemail interface that makes speech browsable, readable and searchable," in *Proc. of CHI*, 2002.

[10] N.O. Bernsen, H. Dybkjær, and L. Dybkjær, *Designing Interactive Speech Systems: From First Ideas to User Testing*, Springer-Verlag, 1998.

[11] L. von Ahn and L. Dabbish, "Labeling images with a computer game," in *Proc. of CHI*, 2004.

[12] T. Volkmer, J. R. Smith, and A. Natsev, "A web-based system for collaborative annotation of large image and video collections," in *Proc. of ACM MM*, 2005.

[13] K. Goldberg, B. Chen, Solomon R., and S. Bui, "Collaborative teleoperation via the internet," in *Proc. of the IEEE ICRA*, 2000.

[14] K. Crowston, H. Annabi, J. Howson, and C. Masango, "Effective work practices for software engineering: Free/libre open source software development," in *Proc. of WISER*, 2006.

[15] N. Dahlbäck, A. Jönsson, and L. Ahrenberg, "Wizard of Oz studies – why and how," in *Proc. of the International Workshop on Intelligent User Interfaces*, Orlando, Florida, USA, 1993.

[16] B. L. Pellom, "SONIC: The University of Colorado continuous speech recognizer," Tech. Rep. TR-CSLR-2001-01, University of Colorado, 2001.

[17] "The Wall Street Journal Dictation Corpus (DARPA-CSR)," The Linguistic Data Consortium, LDC94S13, 1992.

[18] R. Stern, "Specifications of the 1996 Hub-4 broadcast news evaluation.," in *Proc. of the DARPA Speech Recognition Workshop*, 1997.