Measuring the Acceptable Word Error Rate of Machine-Generated Webcast Transcripts

Cosmin Munteanu[1], *Gerald Penn*[1,2], *Ron Baecker*[1,2], *Elaine Toms*[3], *David James*[1]
mcosmin@cs.toronto.edu gpenn@cs.toronto.edu rmb@kmdi.toronto.edu elaine.toms@dal.ca james82@gmail.com

[1]) Department of Computer Science
University of Toronto
Toronto, M5S 3G4, Canada

[2]) Knowledge Media Design Institute
University of Toronto
Toronto, M5S 2E4, Canada

[3]) Faculty of Management
Dalhousie University
Halifax, B4H 4H8, 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 of the hurdles users face when browsing and skimming through archives is the lack of text transcripts of the audio channel of the webcast archive. In this paper, we proposed 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 study where human subjects perform question-answering tasks using archives of webcast lectures, and showed that their performance and perception of transcript quality is linearly affected by WER, and that transcripts of WER equal or less than 25% would be acceptable for use in webcast archives.

Index Terms: Speech recognition, Wizard of Oz, Prototyping, User interface, Webcast.

1. Introduction

Recent years have witnessed an increase in the availability and affordability of broadband Internet connections. This has led to an increase in the use of Internet broadcasting [1], such as on-line lectures. Most such webcast media are archived after being delivered live, and can be accessed by users through interactive systems such as ePresence (`http://epresence.tv/`), illustrated in Figure 1, which serves as framework for this study (a review of webcast systems can be found in [2]).

Without transcripts, humans are faced with increased difficulty in performing tasks that are easily achieved with text documents (such as retrieval of audio and video documents from the archives given a text query, or browsing through a large audio or video document of instead of quickly skimming through a text). Various methods propose improved access to speech recordings [3, 4], however, strong research evidence indicates that transcripts are the most suitable tool for performing tasks that require information seeking from webcast archives [5].

Despite efforts to improve the quality of ASR systems, current systems do not perform satisfactorily in domains such as transcribing lectures or conference presentations. Also, it is expected that such systems will not reach perfect or near perfect accuracy in the near future [6]. Currently, due to the adverse acoustic and linguistic characteristics of lecture speech (large vocabulary, speaker independent, continuous speech, imperfect recording conditions), most lecture recognition systems achieve WERs of about 40-45% [7] (some reports suggest a 20-30% WER for lectures given in more artificial and better controlled conditions [8, 9]).

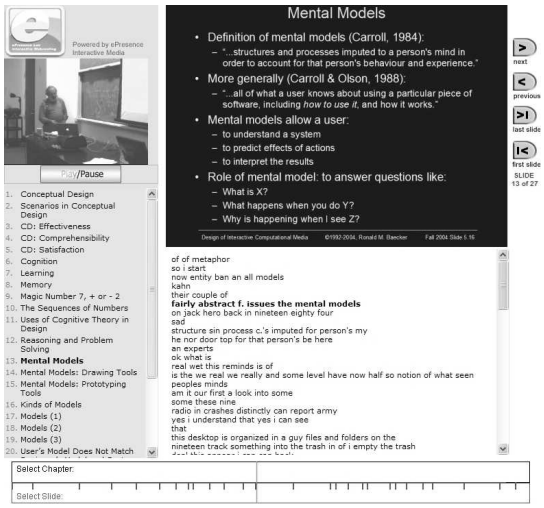

Figure 1: The transcript-enhanced ePresence system.

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 transcripts' quality (and thus, willingness to accept and use transcripts). We also determined what is the minimum level of WER for a transcript to be useful and accepted by users as a feature of webcast systems, and how this compares to the currently or near-future achievable WER for machine-generated lecture transcripts. For this, we designed an ecologically valid experiment, where users performed various tasks using a transcript-enhanced version of the ePresence webcast system. Figure 1 shows the system with transcripts of 45% WER.

ePresence gives users full control of the archive, mainly through the display of the slides used in lectures and the video recording, through interaction with a table of contents (at the left of the screen, which contains "chapter" headings and the title of the slides), and through the timeline (a clickable fine-grained time-progress indicator). For our experiment, transcripts were added to the system. The lines were time-synchronized with the video, by boldfacing the current line of the transcript, thus emulating a closed captioning system, while fully displaying the transcript of the segment of lecture for the current slide. Transcript lines correspond to pauses longer than 200ms. Users can re-synchronize the playback of the video by clicking on a line in the transcript.

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 carefully designed Wizard of Oz experimental framework.

2. Related Work

Studies on the use of archived webcasts [5, 10] indicate that transcripts are a much needed tool to aid navigating through a webcast. Research is thus needed to establish what is a satisfactory quality for archive transcripts, what are the users' expectations from transcripts and how imperfect transcripts should be integrated into a highly-interactive webcast system, but also to develop ASR systems that deliver transcripts with lower WERs.

The task of recognizing speaker-independent, large-vocabulary, continuous, and noisy speech is very challenging. While significant efforts have been spent on improving speech recognition for lectures and presentations [9, 8, 7, 11], the quality of the transcripts (typically WERs of 30-40%, at most 20% in particular conditions) is still below that for other domains, such as broadcast news transcriptions. Unfortunately, the research that investigates how humans deal with such error-ridden transcriptions and which accuracy rates can be deemed acceptable is scarce.

Among the few existing studies, Wizard of Oz experiments showed that humans can only perceive differences in WER greater than 5-10% when directly rating transcripts' quality [12], and that users' expectations of accuracy vary with how critical the domain of the application is [13]. A study that assessed human ability to use transcripts [14] for news recordings retrieval and summarization revealed that users performed better on several measures (time to solution, solution quality, amount of audio played, rate of abandoning the transcripts) when transcripts accuracy was better. A follow-up study in the context of skimming through voicemail messages [15] showed that users performed their tasks faster when simultaneously browsing speech and text, but that performances were lower for keywords not properly transcribed (most critical were phone numbers and names). However, users' performance, when faced with an errorful transcript in a speech browsing interface, can be improved by providing additional information-mining tools [6].

Although these studies provide valuable insights into the users' handling of errorful transcripts, they do not study the relation between performance and WER, nor determine the acceptable WER for a transcript to be included in a browsing interface. Therefore, we have decided to conduct a Wizard-of-Oz-like study to determine these relations, as this simulation method is one of the most appropriate for studying the natural language-based human-computer interaction [16, 17]. Although Wizard of Oz's drawback resides in the need for a skilled human wizard, this method is preferred (instead of prototyping), since the cost of building a full-featured natural language prototype is often prohibitive. However, as it will be shown in Section 4, our proposed simulation method provides the convenience of Wizard of Oz setups while behaving like a true prototype system, with no on-line wizard intervention.

3. Experimental Setup of the User Study

We designed a within-subjects study (a complete description is found in [18]) in which 48 participants were exposed to multiple levels of WER in their interaction, in a typical webcast use scenario – that of the undergraduate student responding to a quiz about the content of a class lecture. We assessed the effect of the WER[1] at four levels: 0% WER (manual transcription), 25% WER (the WER that current ASR systems are able to achieve for broadcast news transcriptions), 45% WER (the WER reported in the literature for the task of transcribing lectures and conference talks, in real-life conditions), and no transcripts (baseline case).

Table 1: The variable used to control the training (overfitting) of the lecture language models.

Variable	Values
Size (in sentences) of lecture corpus	20, 50, 100, 200, all
Modified lecture sentence lengths	1, 5, 7, original
Number of added HUB-4 sentences	0, 650, all
Modified HUB-4 sentence lengths	1, 5, 7, original

Each participant completed a 12-minute long quiz consisting of five factual questions (no lecture comprehension required) for each webcast viewed (one for every level of WER, each on a different 38-minute long 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 the errors in the transcripts. We also collected subjective user data trough post-quiz questionnaires: confidence in their own performance, perception of task difficulty, and impression of transcripts' helpfulness.

4. ASR for a Wizard-of-Oz-like Simulation

As we aimed to evaluate user performance at four pre-determined levels of WER (see Section 3 for the rationale of choosing these levels), we also wanted to maintain a realistic scenario for the Wizard of Oz simulation, as it is recommended for studying natural language-based human-computer interaction [19]. For this, we designed an ASR system that allowed for controlling the level of WER, 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 version 2.0.3 [20]. Transcripts of 25% and 45% WER[2] were obtained by overfitting models to each lecture (in particular, to segments of lectures containing a variable number of sentences). This section describes the design and setup of a WER-controlled ASR system, as well as details about the audio material used in testing.

4.1. Acoustic Models

The acoustic model (AM) that is part of the SONIC toolkit was used in our experiment. The decision tree state-clustered HMMs model is built on 30 hours of data from 283 speakers from the WSJ0 and WSJ1 subsets of the 1992 development set of the Wall Street Journal (WSJ) Dictation Corpus [21]. The WSJ Dictation Corpus is a collection of microphone recordings (1 channel, 16-bit, 16KHz sampling rate) of WSJ news texts read by journalists (not necessarily with experience in dictation). Both for the AMs and for the recognition process the acoustic vectors were represented using SONIC's default[3] Perceptual Minimum Variance Distortionless Response (PMDVR) cepstral coefficients, with a

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

[2] WERs that are usually reported in the literature for current broadcast news (25%) and lecture speech (45%) ASR systems. Future work will take in consideration finer-grained levels of WER.

[3] Overall, we were pleased with most of the out-of-the-box features.

Table 2: The training (overfitting) variables' values for the target WERs of 25% and 45%.

	Lecture 1		Lecture 2		Lecture 3		Lecture 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

39-dimensional feature vector (12 PMVDR parameters) computed over 10ms audio frames and 20ms Hamming windows.

4.2. Language Models

In order to have a greater control of the overfitting process, the training sentences were mixed with the transcripts of the 1997 LDC Broadcast News (HUB-4) Corpus [22] 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 in new sentences of equal length). A summary of the variables used to control the training/overfitting process is presented in Table 1. The tri-gram LMs were built in ARPA format using the CMU-Cambridge Statistical Language Modeling toolkit [23] and converted to the SONIC binary format.

4.3. Lexicon

The pronounciation dictionary was build to cover all words found in the manual transcription, thus, there were no out-of-vocabulary items. Individual lexicons were built for each segment of the lecture corpus on which LMs were trained. The CMU Pronouncing Dictionary v.0.6 [24] was used to extract the pronounciations for the lecture words. For technical words not in the dictionary, the SONIC's `sspell` lexicon access tool was used to generate pronounciations using letter-to-sound predictions from a decision tree which we trained on the entire CMU Pronouncing Dictionary.

4.4. Recordings

The recordings used for our study were collected in a large, amphitheatre-style, lecture hall (200 seating capacity), using the AKG C420 head-mounted directional microphone. The lecturer is male, early 60s, and a native speaker of English. The recordings were not intrusive, and no alterations to the lecture environment or proceeding were made. The 1-channel recordings were digitized using the TASCAM US-122 audio interface as uncompressed audio files with 16KHz sampling rate and 16-bit samples.

4.5. Recognition

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 presented in Table 1. The variables' values for the target WERs of 25% and 45% are outlined in Table 2.

The SONIC decoder performs recognition in two passes. The first pass decoding uses the specified AMs and LMs. After the first pass is complete, an unsupervised Maximum Likelihood Linear Regression (MLLR) of the AM is performed using the output of the first pass (ASR hypotheses are labeled with confidence scores). The second pass uses the MLLR-adapted AM. Since the recognition is performed in two passes, each of them producing its own hypotheses, we also considered using as one of the WER-controlling variables the pass from which we selected the ASR output. However, as mentioned in [20], SONIC's MLLR adaptation in the second pass usually produces an output of a slightly lower WER. Thus, we found that the output of the first pass was always 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 an existing webcasting system that additionally provided the following components: video of the presentation, slides, table of contents, and timeline. This setup allowed us to design an ecologically valid experiment as in a Wizard of Oz simulation, without the need for the on-line intervention of a human wizard.

5. Results of the User Study

While a complete analysis of the data collected through the user study is presented in[18], we will summarize here the key findings. With respect to quiz scores, our study revealed that transcripts of WER of 25% were marginally better than having not transcripts in the webcast system, while 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%.

Through a post-session questionnaire, users indicated that they rather have transcripts with errors than no transcripts and would use such a system for most academic tasks. Navigational features such as the 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.

6. Conclusions and Future Work

One of the major drawbacks for the users of audio/video archives (such as those of webcast lectures and presentations) is the difficulty in performing operations typically associated with archived text. While manual transcription is not time and cost-effective, for lecture and presentation speech, the poor accuracy of ASR-generated transcripts makes their use questionable.

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 common simulation techniques (prototyping and Wizard of Oz) 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 where subjects used a fully-featured webcast browsing tool, while answering quizzes based on archives of webcast lectures. The study revealed that WER linearly influenced users' performance, and that for transcripts with a WER equal to or less than 25%, users' task performance was better than that of using no transcripts. WER also influenced (linearly) the users' perception of transcript quality and task difficulty, and transcripts of WER of 25% were better in this respect that using no transcripts.

Existing research on ASR for lectures and presentations shows promising results that can lead to a further reduction of error rates for these domains: while current lecture-dedicated systems can achieve WERs of 40-45%, emerging ASR systems can, in certain conditions, reduce the WER up to 20-30%. We are currently focused on developing better ASR systems that will be able to deliver WERs of 25% for real-life lecture conditions. Also, since current measures of speech recognition accuracy (mainly WER) might not fully reflect user needs for transcript quality, we are working on developing other more appropriate measures of quality.

7. Acknowledgements

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

8. References

[1] P. Ritter, "The business case for on-demand rich media," Wainhouse Research Whitepapers, 2004.

[2] R. M. Baecker, "A principled design for scalable internet visual communications with rich media, interactivity, and structured archives," in *Proc. of CASCON*, 2003, pp. 83–96.

[3] B. Arons, "SpeechSkimmer: A system for interactively skimming recorded speech," *ACM Transactions on Computer-Human Interaction*, vol. 4, no. 1, pp. 3–38, 1997.

[4] N. Sawhney and C. Schmandt, "Nomadic radio: Speech & audio interaction for contextual messaging in nomadic environments," *ACM Transactions on Computer-Human Interaction*, vol. 7, no. 3, pp. 353–383, 2000.

[5] C. Dufour, E. G. Toms, J. Lewis, and R. M. Baecker, "User strategies for handling information tasks in webcasts," in *Proc. of CHI*, 2005, pp. 1343–1346.

[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, pp. 253–269, Springer-Verlag.

[7] E. Leeuwis, M. Federico, and M. Cettolo, "Language modeling and transcription of the TED Corpus lectures," in *Proc. of the IEEE Conference on Acoustics, Speech, and Signal Processing*, 2003, pp. 232–235.

[8] I. Rogina and T. Schaaf, "Lecture and presentation tracking in an intelligent meeting room," in *Proc. of the International Conference on Multimodal Interfaces*, 2000.

[9] K. Kato, H. Nanjo, and T. Kawahara, "Automatic transcription of lecture speech using topic-independent language modeling," in *Proc. of the International Conference on Spoken Language Processing*, 2000, pp. 162–165.

[10] E. G. Toms, C. Dufour, J. Lewis, and R. M. Baecker, "Assessing tools for use with webcasts," in *Proc. of the Joint Conference on Digital Libraries*, 2005, pp. 79–88.

[11] A. Park, T.J. Hazen, and J.R. Glass, "Automatic processing of audio lectures for information retrieval," in *Proc. of the IEEE Conference on Acoustics, Speech, and Signal Processing*, 2005, pp. 497–500.

[12] R. Van Buskirk and M. J. LaLomia, "The just noticeable difference of speech recognition accuracy," in *CHI Mosaic of Creativity: The Conference Companion on Human Factors in Computing Systems*, 1995, p. 95.

[13] M. J. LaLomia, "User acceptance of handwritten recognition accuracy," in *The Conference Companion on Human Factors in Computing Systems*, 1997, p. 107.

[14] L. Stark, S. Whittaker, and J. Hirschberg, "ASR satisficing: The effects of ASR accuracy on speech retrieval," in *Proc. of the International Conference on Spoken Language Processing*, 2000, pp. 1069–1072.

[15] 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 the SIGCHI Conference on Human Factors in Computing Systems*, 2002, pp. 275–282.

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

[17] A. Life, I. Salter, J.N. Temem, F. Bernard, S. Rosset, S. Bennacef, and L. Lamel, "Data collection for the MASK kiosk: WOz vs prototype system," in *Proceedings of the International Conference on Speech and Language Processing*, Philadelphia, Pennsylvania, USA, 1996, pp. 1672–1675.

[18] 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 *Proceedings of the ACM Conference on Human Factors in Computing Systems – CHI*, 2006.

[19] N. Dahlbäck, A. Jönsson, and L. Ahrenberg, "Wizard of Oz studies – why and how," in *Proceedings of the International Workshop on Intelligent User Interfaces*, Orlando, Florida, USA, 1993, pp. 193–200.

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

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

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

[23] P.R. Clarkson and Rosenfeld R., "Statistical language modeling using the CMU-Cambridge Toolkit," in *Proceedings of ESCA Eurospeech*, 1997, vol. 1, pp. 2707–2710.

[24] "The CMU Pronouncing Dictionary v. 0.6," http://www.speech.cs.cmu.edu/, 1998.