Two Experiments Comparing Reading with Listening for Human Processing of Conversational Telephone Speech•

Douglas Jones[1], Wade Shen[1], Elizabeth Shriberg[2], Andreas Stolcke[2], Teresa Kamm[3], Douglas Reynolds[1]

[1]MIT Lincoln Laboratory
Lexington, MA USA
{daj,swade,dar}@ll.mit.edu

[2]SRI International
Menlo Park, CA USA
{ees,stolcke}@speech.sri.com

[3]Department of Defense
Fort Meade, MD USA
tkamm@acm.org

Abstract

We report on results of two experiments designed to compare subjects' ability to extract information from audio recordings of conversational telephone speech (CTS) with their ability to extract information from text transcripts of these conversations, with and without the ability to hear the audio recordings. Although progress in machine processing of CTS speech is well documented, human processing of these materials has not been as well studied. These experiments compare subject's processing time and comprehension of widely-available CTS data in audio and written formats – one experiment involves careful reading and one involves visual scanning for information. We observed a very modest improvement using transcripts compared with the audio-only condition for the careful reading task (speed-up by a factor of 1.2) and a much more dramatic improvement using transcripts in the visual scanning task (speed-up by a factor of 2.9). The implications of the experiments are twofold: (1) we expect to see similar gains in human productivity for comparable applications outside the laboratory environment and (2) the gains can vary widely, depending on the specific tasks involved.

1. Introduction

Over the last decade, automatic speech recognition technology has made consistent progress. We observe a drop in word error rate from 90% in 1993, for original Switchboard data, to 15% in 2004 for Rich Transcription data based on the Switchboard and Fisher corpora available at the Linguistic Data Consortium (LDC), as shown in Figure 1 [6]. Although there has been significant research into the benefits of automatic speech recognition for human users of applications such as voice mail and audio search [4] and for human readers of various types conversational telephone speech transcripts [5], we know of no experiment that directly compares Switchboard- and Fisher-type CTS transcripts with audio in terms of human processing.

In the following sections we describe two experiments that directly compare CTS transcripts with audio. We hypothesize that if subjects are faster at processing telephone conversations by reading than by listening, we can observe differences in terms of task time completion and accuracy in a controlled experimental setting. More specifically, if subjects are faster at reading than listening without degrading comprehension or task performance, then it follows that consumers of audio data can realize efficiency benefits from transcription. Two experiments were run that test this hypothesis: one that involves careful reading and one that involves visual scanning.

Written experiment materials were available from an earlier reading experiment that tested comprehension of CTS items from the Rich Transcription 2003 (RT-03) data sets [1] in various text conditions [5], but those experiments did not include audio stimuli. Both of the experiments reported in this paper use Fisher audio and text files: around 750 words for around 4½ minutes of audio. We selected 18 of the 36 Fisher transcripts from the RT-03 evaluation set for Experiment I and the disjoint set of 18 files for Experiment II.

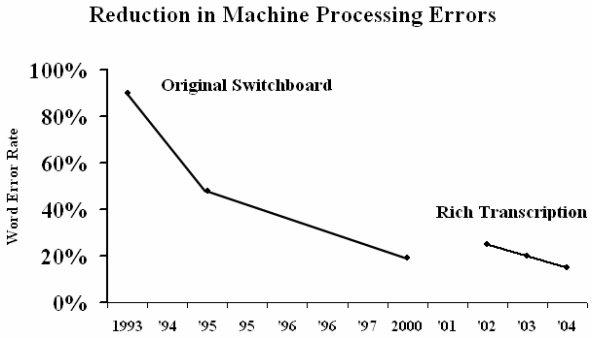

Figure 1: Improvements in Machine Processing for Conversational Telephone Speech

2. Experiment I – Careful Reading

In Experiment I, subjects were asked to read or listen to conversations and answer multiple choice questions about them. Subjects were asked to maximize their accuracy. The test items they were given consisted of a conversation (in text, audio or text+audio conditions) and three multiple-choice questions designed to elicit careful reading/listening by subjects. Test items were presented in a latin square design so that each item was seen by every subject and each item only appeared in one condition for any given subject.

2.1. Materials and Processing

We selected 18 Fisher audio files and transcripts with associated multiple choice questions for each subject to

• This work is sponsored by the Department of Defense under Air Force Contract F19628-00-C-0002. Opinions, interpretations, conclusions and recommendations are those of the author and are not necessarily endorsed by the United States Government.

process (approx. 6 per condition). Figure 2 shows a transcript fragment and one sample question.

> **... B:** I think in a way it's good. I was divorced. I got married too early. And then I was divorced. And now I'm married again. And it's good. It's much better than if I had stayed with that first person. I think the only thing is it makes some people maybe not take marriage as seriously as they would have otherwise **...**
>
> **Questions:**
> **1. What is the current marital status of one speaker?**
> **(A) married** (B) divorced (C) engaged (D) widowed **...**

Figure 2: Sample RT-03 EARS Readability Test Item

In preparation for Experiment I, both the text transcripts and the audio files were enhanced relative to previous RT-03 test materials. The reason for enhancing the transcripts is obvious – we did not want accidental errors in the reference transcripts or irrelevant artifacts of the text-rendering process to interfere with human readability.

The reason for enhancing the audio files is less obvious but important nonetheless for establishing a realistic baseline: 1) we would not be providing subjects with a sophisticated audio browsing environment 2) our subjects were not trained in audio browsing. The type of audio processing that we performed could be expected, given basic off-the-shelf tools and skilled audio users. Thus, we believe it to be a reasonable approximation of basic audio performance. The most straightforward improvement was simply to remove all silences above a certain length so that the conversation can be heard in a shorter amount of time. This has been done for the materials in Experiment I.

2.1.1. Audio Preparation

Audio files of each conversation were preprocessed for listening so as to eliminate extended non-speech regions. Although this editing was performed by making use of the reference transcripts, it was felt that it was reasonably close to automatic processing capabilities, and would avoid an unfair disadvantage for the listening condition. Even in the absence of automatic preprocessing, but assuming state-of-the-art playback equipment, listeners could skip over such regions quickly. Non-speech regions were inferred from forced alignments of the NIST RT-03 reference transcripts to the waveforms, using SRI's CTS recognizer. The time marks of all speech tokens were extracted, and regions not containing speech on either channel for more than 0.3s were marked for deletion. On average, waveforms were shortened by 21% as a result. The edited signals where then added into a single channel for playback. Spot-checking of the output confirmed that this procedure yielded waveforms that did not eliminate any speech but was devoid of typical conversational pauses, without sounding unnatural. The beginning and end of each conversation was also truncated so as to correspond exactly to the transcribed data used as stimuli in the experiments.

2.1.2. Transcript Selection and Cleanup

We informed our selection of the 18 conversations from the RT-03 files available using the following approach. The conversations were sorted by overall speaking rate, after removing long pauses. Because future analyses may want to examine effects of speaking rate on modality differences, we selected the 6 slowest, 6 fastest, and 6 most medial-speed conversations based on overall speaking rate, (collapsing over both talkers).

Original transcripts from these conversations were annotated (punctuation was added), and cleaned up (disfluencies and backchannels were removed) using both the official Linguistic Data Consortium's reference annotations [7] and a second human pass. We used LDC's reference transcriptions as well as LDC's Meta Data Extraction (MDE) annotations, based on reference RTTM files [1]. A description of the metadata markup can be found on LDC's web site [8].

Based on this input, we: extract each Sentential Unit; remove all filler words (filled pauses and discourse markers); remove all words in the edit regions of other disfluencies; remove incomplete Sentential Units; capitalize the first character of each Sentential Unit; add punctuation (period, question mark); remove some backchannels. The mean number of words removed was 17% (standard deviation 5%, minimum 9%, maximum 27%). In a second, automated pass, we converted the revised annotated transcripts to enhance readability even further by removing all backchannels and conflating adjacent turns by the same speaker.

The result was a transcript composed of more conventional-appearing paragraphs preceded by speaker labels. In a third, hand-editing pass, which did not significantly modify the overall word statistics above, we removed a small number of remaining backchannels (not marked in the LDC RTTM file) and cleaned up a variety of small problems due to isolated misannotations in the original input files.

2.2. Test Platform and Human Subjects

We delivered the test to subjects using a simple web-based test-taking environment. The subjects were able to scroll through the text transcripts and to use simple audio navigation buttons (play, stop, fast forward, rewind, and audio slide bar). Passages containing both text and audio did not provide any linkage between the two modalities. Instead, subjects were allowed to choose any combination of audio or text for question-answering.

Thirty college-level adults from the Boston area participated in this experiment. All subjects participated in both Experiment I and Experiment II.

2.3. Results

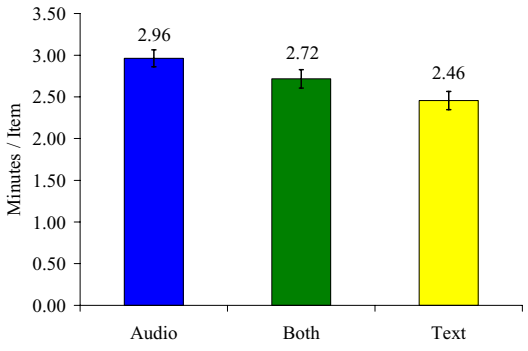

Figure 3: Speed for Experiment I (Careful Reading)

We observe in Figure 3 that subjects were slightly faster when reading texts (2.46 minutes per item) than when listening to audio (2.96 minutes per item), by a factor of 1.2. The combination of text and audio (2.72 minutes per item) is faster by a factor of 1.1. Although the differences across conditions are modest, they are statistically significant, both and by subject (F (1, 29): 410.848; P: 0.000) and by item (F(1,17): 147.878, P: 0.000).

Question-answering accuracy was 90% for audio, 88% for text+audio, and 92% for text-only but the differences were not significant across conditions.

We conducted an exit interview of the subjects and asked them their preference for different media types during testing. The results are shown in Figure 4. A little over half of the subjects preferred text-only.

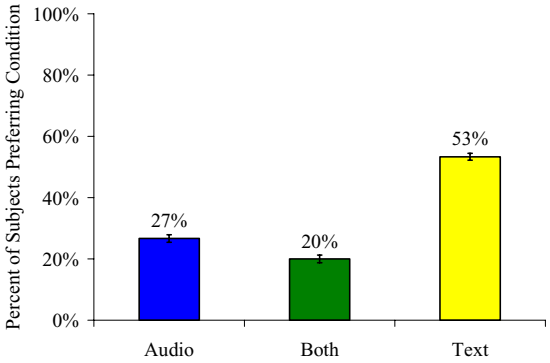

Figure 4: Preferences for Experiment I (Careful Reading)

One thing that subjects mentioned was that it was very boring to read these telephone conversations and that listening forced them to think about the questions while listening for answers. We expect that more information-rich materials might have results that show more preference for the text condition.

3. Experiment II – Visual Scanning

In order to test the relative performance of transcripts more thoroughly, we conducted a second experiment which involved a simple, uniform test question for all of the transcripts. The second experiment was designed to place more emphasis on text skimming than on careful reading. We used the remaining 18 Fisher transcripts from the RT-03 evaluation set, but rather than using multiple choice questions like those used in Experiment I, we asked a uniform question of each conversation: *Does either speaker mention where they are calling from or where they live or have lived in the past?* Six of the 18 transcripts contained reference to a speaker's origin or location.

3.1. Design, Test Platform and Subjects

The overall design and test platform for Experiment II is identical to Experiment I – 18 conversations in three conditions: audio-only, text-only, and text combined with audio in a latin square design. All of the Experiment II subjects participated in Experiment I.

3.2. Materials

Neither text nor audio were enhanced for Experiment II due to scheduling constraints. The conditions for conversational materials for Experiment II are as follows: Transcript Only, (not enhanced); Audio Only (not enhanced); Both Reference Transcript and Un-enhanced Audio. We feel that the differences in the quality of materials in Experiment I and Experiment II did not significantly impact the overall results.

3.3. Text Transcripts

We used the transcripts as annotated by the LDC for the RT-03 evaluation data. These transcripts omit the introductory remarks that the callers make, which typically involve an exchange of greetings, names and locations. Although many speakers do not provide their location, either at the beginning of the call or later on, the ones who do typically mention it in their greeting. Excluding the greeting provided better test data since the mention of origin or location would not be predictable, i.e., appearing at the very beginning of the conversation.

3.4. Audio Files

Since we used the official transcript regions provided for the RT-03 evaluation, we trimmed the audio files to match these regions exactly.

3.5. Results

There was a substantial speed-up for the two text conditions over the audio-only condition for this experiment, as shown in Figure 5.

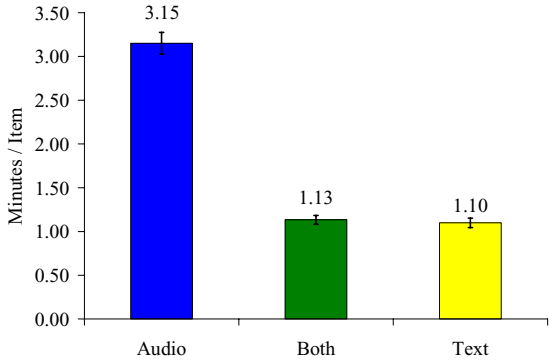

Figure 5: Speed for Experiment II (Visual Scanning)

For the audio-only condition, subjects spent 3.15 minutes per item (comparable to the 2.96 minutes per item in Experiment I). But we see that skimming for speakers' origins and locations at 1.10 minutes per item is faster by a factor of 2.9. The combination of text and audio (1.13 minutes per item) is faster than audio alone by a factor of 2.79. The differences across conditions is significant, both by subject (F(1,29): 424.377; P: 0.000) and by item (F(1,17): 150.346; P: 0.000). There was no significant difference in accuracy for the three conditions, as in Experiment I, although overall the accuracy was slightly lower: 83% for audio, 87% for audio+text, and 85% for text-only.

As in Experiment I, we conducted an exit interview of the subjects and asked them their preference for the materials. The results are shown in Figure 6. Subjects overwhelmingly preferred text-only, at 87%. Only 3% of the subjects preferred audio only, and 10% preferred to have both audio and text.

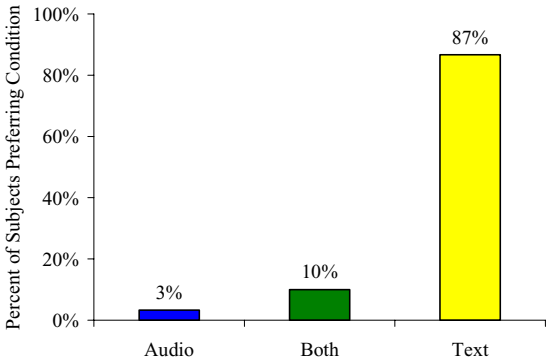

Figure 6: Preferences for Experiment II (Visual Scanning)

A reason that was frequently given for their preference was that they were afraid they might miss a mention of an origin or location in the audio, and it was much faster to skim for this information than to listen.

4. Conclusions

The first experiment employed test items with three multiple-choice questions for testing careful reading or listening of telephone conversations. The second used a uniform question designed to elicit information scanning: we asked whether either speaker in the conversation revealed his or her location or origin. We observe a very modest impact for transcripts over audio for the careful reading task (speed-up by a factor of 1.2) and a much more dramatic impact for transcripts over audio in the visual scanning task (speed-up by a factor of 2.9).

The implications of the experiments are twofold: (1) we expect to see comparable gains in human productivity in comparable applications outside the laboratory setting and (2) the gains can vary widely, depending on the specific tasks involved.

5. Future Work

The experiments described in this report contrast human reading and human listening modalities in processing conversational telephone speech. Although they serve as a baseline for comparison, we believe that more striking contrasts between audio and text transcripts could be shown in conditions that involve massive volumes of data that are simply beyond the ability of individual humans to process in a limited time.

For specific tasks that require human processing of individual telephone conversations, we expect significant enhancements in productivity. The degree of enhancement depends heavily upon the processing task, as we have demonstrated in these two experiments. Overall, timing was not significantly different for text-only compared with the combination of text and audio. Accuracy did not differ significantly, regardless of whether subjects read or listened to the materials.

In future experiments we would like to contrast text with audio in more massive volumes of text. For human-readable quantities, we would like to develop methods for characterizing the task more precisely in terms of reading, listening and reasoning in order to predict which tasks can most benefit from transcripts. We would also like to conduct experiments in which errorful transcripts are used, in order to estimate utility thresholds for fully automatic speech-to-text systems for particular tasks.

6. Acknowledgements

We wish to acknowledge Paul Gatewood and John Tardelli at ARCON Corporation for recruiting subjects, preparing materials and administering the experiment under significant time pressure. Dimitra Vergyri of SRI performed the audio file editing, and Yang Liu of ICSI processed the reference transcripts based on metadata annotations.

7. References

[1] Doddington, G. 2003. RTTM Format Specification. http://www.nist.gov/speech/tests/rt/rt2003/fall/

[2] Garofolo, J. (2003). Rich Transcription Spring 2003 benchmark tests.

[3] Gibson, Edward, et al. 2004. Two New Experimental Protocols for Measuring STT Readability. Report for DARPA/EARS/Rich Transcription 2004 Workshop.

[4] Hirschberg, J., et al. SCANMail: Browsing and Searching Speech Data by Content. Proc. of EuroSpeech, 2001.

[5] Jones, Douglas, Florian Wolf, Edward Gibson, Elliott Williams, Evelina Fedorenko, Douglas Reynolds, Marc Zissman. 2003. Measuring the Readability of Automatic Speech-to-Text Transcripts. Proceedings of EuroSpeech Conference, Geneva.

[6] Pallett, David S. 2003. A Look at NIST's Benchmark ASR Tests: Past, Present and Future. http://www.nist.gov/speech/history/pdf/NIST_benchmark_ASRtests_2003.pdf

[7] Strassel, S. et al. Fisher English Training Transcript Data, http://www.ldc.upenn.edu/Catalog/docs/LDC2004T19.

[8] Strassel, S. (2003). Simple metadata annotation specification Version 5.0. Linguistic Data Consortium, Philadelphia, PA. http://www.ldc.upenn.edu/Projects/MDE

[9] Whittaker, S, J. Hirschberg, J. Choi, D. Hindle, F. Pereira, and A. Singhal, "SCAN: Designing and evaluating user interfaces to support retrieval from speech archives," In Proc. ACM SIGIR '99, 1999.