ASR Satisficing: The effects of ASR accuracy on speech retrieval

Litza Stark[1]*, Steve Whittaker*[2]*, and Julia Hirschberg*[2]

[1]University of Delaware,
Newark, DE, 19716
[2]AT&T Labs–Research
Florham Park, NJ, 07932, USA
litza@udel.edu, {stevew, julia}@research.att.com

ABSTRACT

We examine how differences in the accuracy of Automatic Speech Recognition transcripts affect users' ability to use these in tasks requiring the retrieval of speech "documents". We compare performance measures, processing strategies, and preference data for subjects using transcripts and speech data to perform a series of relevance judgment and summary tasks on transcripts with different levels of accuracy. Results show effects for transcript quality on solution accuracy, time to solution, amount of speech played for the task, likelihood of subjects abandoning use of a transcript, and subject perceptions of task difficulty, transcript utility, readability, and comprehensibility.

1. INTRODUCTION

Research on Automatic Speech Recognition (ASR) generally assumes perfect transcription accuracy to be its holy grail. The more accurately a system can transcribe an utterance, the better that system performs (modulo time factors) Recently this assumption is being challenged, however. Different metrics are being discussed in the context of new applications for ASR technologies, such as spoken dialogue systems and speech retrieval systems. Should metrics such as concept accuracy (for dialogue systems) or system performance on certain classes of words (for speech retrieval systems) be preferred in training a recognizer for these tasks? So far, however, these questions have primarily been asked when the transcript is being used **system internally**, for example by the dialogue manager in an interactive system or the information retrieval component of a speech browser. Little research has been done on the effects of transcription accuracy on humans using the transcripts to carry out real world tasks. This paper addresses this issue in the context of a speech retrieval system, SCAN, Spoken Content-Based Audio Navigator.

2. THE SCAN SYSTEM AND ITS USER INTERFACE

SCAN operates on the NIST TREC SDR corpus, a subset of the DARPA HUB-4 Broadcast News corpus, which includes news broadcasts from the major networks and CNN, hand-segmented by NIST labelers into news stories. SCAN uses ASR techniques to produce an errorful transcript of each story, operating on speech segmented into audio paragraphs, or, PARATONES. Stories relevant to a text query are retrieved by an information retrieval (IR) system — in SCAN, a modified version of the SMART system [2, 1]. Results of the recognition and retrieval stages are then passed to SCAN's graphical user interface (GUI).

The SCAN GUI was designed to support local navigation within speech documents, as well as document retrieval. Based upon empirical studies of search in the voicemail domain [4], we proposed a new paradigm for speech retrieval interfaces: "what you see is (almost) what you hear" (WYSIAWYH) [3], a multimodal approach exemplified in Figure 1. WYSIAWYH is based on the hypothesis that humans will find a *visual analogue* to the speech recording useful in search and browsing. To this end, we employ

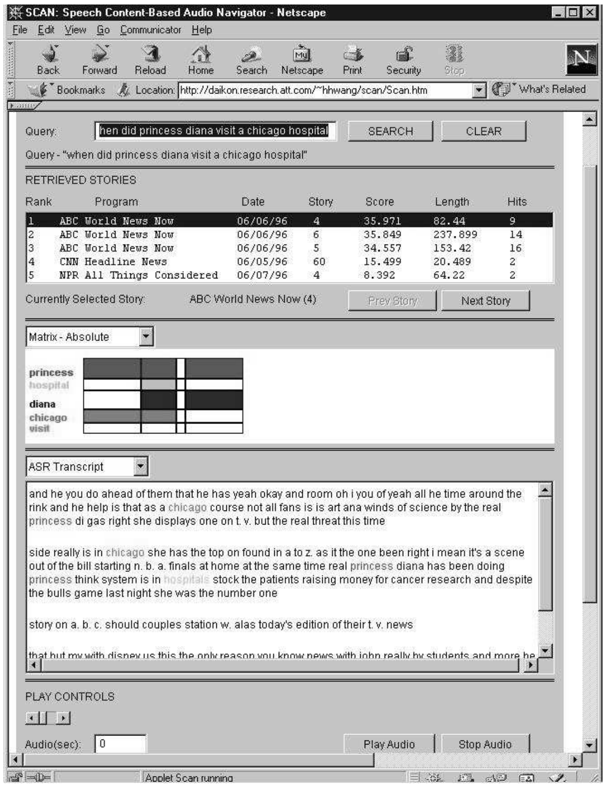

Figure 1: WYSIAWYH Browser

well-understood text formatting conventions (such as headers and paragraphs) to provide useful analogues for speech browsing.

The SCAN GUI includes a SEARCH COMPONENT, which allows users to input their queries, at the top of the browser. Search results are depicted in the RESULTS panel immediately below, which presents a relevance-ranked list of 10 audio documents, with information on program name and story number, date, relevance score, length (in seconds), and total hits (number of instances of query words in document). The VISUAL OVERVIEW component provides high-level information about the current document. Each query term is color coded, and each paratone represented by a column in a matrix, where column width shows the relative length of that paratone in the story. If a query term occurs in a paratone then the matrix cell corresponding to term and paratone is depicted in the relevant color (in Figure 1 the query term "princess" is depicted in cells representing the first, second and fourth paratones). The ASR TRANSCRIPT provides detailed (if sometimes inaccurate) information about the contents of a story. Query terms in the transcript are highlighted and color-coded, using the same coding scheme used in the Visual Overview. Users can play a given paratone by clicking on the corresponding column in the Overview or paragraph in the Transcript. Finally, a simple PLAY BAR represents a single story, which users can access randomly within the bar, plus start and stop buttons to control play for this component and others.

2.1. Evaluating the GUI

In evaluating the SCAN GUI we compared it with a simpler version which contained only the Results panel and the Play controls — i.e. with neither overview nor ASR transcript. We found that users generally performed better with the WYSIAWYH browser in terms of time to solution, solution quality, perceived task difficulty, and users' rating of browser usefulness, for document relevance judgments (subjects were asked to determine which of a set of five retrieved documents is most relevant to a given query), and fact-finding (subjects were asked to find the (factual) answer to a specific question). However, when the task was to summarize a given document, there was no difference between the WYSIAWYH browser and the simple browser.

A central concern in our experiments was the role of the ASR transcript in supporting browsing. As noted above, the transcript (the same used for IR) is intended to support information extraction, providing detailed information about the contents of a story. The transcript panel displays a transcript of the selected story. Since the overall ASR word accuracy on the corpus in this example was 69% on this corpus, transcripts usually contain errors (in paragraph 2 of the transcript in Figure 1, "in a tizzy" is transcribed as "in a to z"). So, transcripts may be more or less helpful. In regions where it is mostly accurate, it is intended that users find relevant information simply by reading – without the time burden of listening to the audio. Where it is less accurate, it should at least enable rapidly visual scanning to find relevant regions to play in the audio. The transcript also should provide local contextual information: users can decide whether to play a particular paratone by reading surrounding paragraphs to determine its likely relevance. Finally, overall transcript quality should help users assess the likely accuracy of all the transcript, search and overview information. For example, bizarre phrases like "buster and those ties and assess state...." (Figure 1, paragraph 2) indicate the transcript is inaccurate. They may thus also suggest that query terms in the overview may have been misrecognized, and, thus, that the overall retrieval of relevant documents may be flawed.

Subject's comments and our observations indicated 3 main uses of the transcript: (a) DIRECT READING without accessing the underlying audio; (b) VISUAL SCANNING to relevant regions and then playing these for accurate information; (c) CONCURRENT PLAYING AND LISTENING, where subjects listened to a given paragraph while scanning nearby paragraphs for contextual information relating to what they were playing. In these experiments, the mean word accuracy over all audio "documents" was 67%, ranging from a maximum of 88% to a minimum of 35%. According to our subjects, the choice and efficacy of transcript-based strategies was highly dependent on the accuracy of the transcript. Direct reading was only possible with accurate transcripts, while visual scanning and concurrent playback tended to be used for transcripts judged to be comprehensible but errorful, where the audio could be used to check the transcript accuracy.

When we examined the effect of differences in transcript recognition accuracy explicitly, we found some effect on quality of solution and users' perception of task difficulty, but, surprisingly, no effect on amount of speech played or on time to completion of task. There were also task-specific effects. For fact-finding, better quality ASR led to higher quality solutions ($r(22) = 0.42$, $p < 0.05$), and there was a trend towards lower perceived task difficulty ($r(22) = 0.35$, $p < 0.07$). User comments also suggested that, with higher quality transcripts, users were able to extract more information from the transcript alone, reducing the amount of speech they needed to play, and allowing them to be more precise about what they played. Where transcript quality was poor, they were forced to do more listening: "I wanted to scan the transcript but I found a massive number of errors in the speech recognition, so I decided to listen." However, we could find no objective evidence for reduction in playing time with more accurate ASR ($r(22) = 0.25$, $p > 0.05$), nor were users faster to solve the task ($r(22) = 0.05$, $p > 0.05$). There were also no effects of ASR quality on any measure for the summary task. This is consistent with our earlier finding — that in general the SCAN UI did not help with the summary task.

Why did transcript quality fail to affect outcome and process measures more directly, and why did we find task specific effects for some of our measures? One possibility is that transcript quality was confounded with story, since we had only one transcript per task. To control for this and to better understand the effects of transcript quality, we conducted a second experiment.

3. EFFECTS OF TRANSCRIPT QUALITY

3.1. Experimental Design

In the previous study, transcript quality was confounded with story. For our new experiment, we controlled for this by choosing transcripts from the output of eight different ASR systems with different degrees of accuracy. The overall word accuracy levels of these systems ranged from 10% correct to 99.2% correct. So, for each story used in our experiment, we obtained four different levels of word accuracy — 100% correct (human transcript), 84%, 69%, 50%. We could therefore compare users performing the same tasks but with different levels of word accuracy in the materials they used. Eight users performed four instances of both the summarization and relevance judgment tasks described for the previous study. Tasks and transcript quality were randomized within subject. In this experiment, subjects saw only the search, results, transcript, and play components of the GUI, not the overview.

For each question we collective objective data, including time to solution and quality of solution (as assessed by two independent judges), as well as number, type, and duration of browsing and play operations. We also observed subjects to identify different strategies they were using to carry out the task (e.g. reading only, playing only, scanning and playing, etc.). We collected subjective data via having subjects "think aloud" during the experiment and recording their statements, and via brief post-task questionnaires (task difficulty, usefulness of the player, utility of the transcript, overall transcript readability and comprehensibility and criteria people used for judging transcript quality).

3.2. Hypotheses

The experiment was designed to test three aspects of the possible effect of transcript quality (measured in word accuracy) on user behavior and perception: Whether subject **performance** would differ (H1 and H2); whether the way subjects **process** documents would differ (H3a, H3b, and H4); and whether subjects would **subjectively perceive** their tasks differently (H5 and H6). These were the hypotheses tested:

1. High quality transcripts, i.e. those with higher word accuracy, afford ease of accessing information by scanning and reading without the need for playing the story. Subjects should be faster to generate solutions when transcripts are more accurate.
2. Subjects should therefore produce higher quality solutions with high accuracy transcripts for similar reasons.
3. (a) More accurate transcripts allow straightforward access by scanning and reading. So, subjects should play less of the story when transcript quality is high.

   (b) More accurate transcripts allow straightforward access to information by scanning and reading, while information extraction is harder with low quality transcripts. Subjects should therefore spend less time reading stories when transcript quality is high.
4. Given the difficulty of extracting information from a low quality transcript, subjects should be more likely to abandon use of the transcript when transcript quality is poor, relying exclusively on playback.
5. Subjects should perceive the task to be easier when transcripts have higher word accuracy.
6. Subjects should perceive high quality transcripts to be more useful for carrying out the tasks than low quality transcripts.

In addition, we investigated which aspects of the transcript led subjects to question its accuracy, by gathering subjective ratings of perceived comprehensibility and readability of the transcript. We also examined how perceptions of various lexical, syntactic and semantic characteristics of the transcripts were related to these. For example we asked users whether they felt that "bizarre syntax", "odd terms", "incomprehensible words" or "strange proper names" had contributed to their perceived comprehensibility or readability of the transcripts. These questions and descriptors were based on an analysis of subjects' characterizations of transcript errors in our previous studies.

3.3. Analysis and Results

We conducted multiple separate ANOVAs with the following independent variables: subjects, transcript quality (with two levels: "high" for word error rates above 84%, and "low" for error rates below 70%), task type (relevance judgments versus summarization), and order (whether this was the first or second time that a subject had encountered a transcript of this quality). The dependent measures were: solution quality (as evaluated by two independent judges), time to solution, total amount of time spent playing speech for the task, total amount of time spent reading, perceived difficulty, and the likelihood of a subject abandoning use of the transcript, to rely solely on playing. We report interactions and post hoc tests where these are relevant to the hypotheses.

Table 1: Effects of ASR quality on processing

Measure	High ASR Quality	Low ASR Quality	Hyp
Time to Solution	335s	390s	Conf
Norm'd Solution Quality	0.22	0.19	Conf
Am't Played	37.9s	52.4s	Conf
Reading Time	297s	337s	Conf
Prob. Abandoning Transcript	0	0.16	Conf
Perc'd Task Diff	2.32	3.12	Conf
Perc'd Transcript Utility	2.78	4.34	Conf

Table 1 shows the results. Analysis of the effect of word accuracy differences on performance shows that higher quality transcripts only affect time to solution, not solution quality. Hypothesis H1, that higher word accuracy should lead to faster solutions was confirmed. People answered questions more quickly with high quality transcripts ($F(1,16) = 7.34$, $p < 0.02$). Post-hoc tests showed that this advantage for transcript quality only occurred with the relevance task however. Hypothesis H2, that higher quality transcripts should produce produce higher quality solutions by providing easier access to information was not confirmed ($F(1,16) = 0.60$, ns). One possible reason for the failure to find effects is that there were individual differences in strategy, with more diligent subjects spending more time trying to optimize solutions. We therefore normalized solution quality scores by dividing them by the time to complete the task. The resulting analysis showed that high quality transcripts produced better solutions faster, but that this effect was limited to the first instance of where a subject was presented with that quality of transcript ($F(1,16) = 26.98$, $p < 0.0001$). One interpretation is that subjects evolved strategies for dealing with low quality transcripts, when they received them later in the experiment.

Analysis of how transcript differences affected subjects' processing strategies revealed that indeed subjects play less recorded speech when the quality of the transcript is high (Hypothesis H3a); $F(1,16) = 11.67$, $p < 0.002$. In addition, people spent less time reading high quality transcripts (Hypothesis 3b); $F(1,16) = 3.98$, $p < 0.05$). However, post-hoc tests again showed that transcript quality reduced reading time only for relevance judgments but not summaries. Finally, we found that subjects indeed did abandon transcripts more quickly when transcript quality was poor (Hypothesis 4), relying simply on play commands. Our first measure was based on the likelihood that people would stop using the transcript in the first paragraph. As predicted, people were much less likely to stop using high quality transcripts in the first paragraph. ($F(1,16) = 8.33$, $p = 0.001$). We also looked at the likelihood that users would abandon using the transcript at any point (not just in the first paragraph), and these results are shown below. Transcript

Table 2: Abandonment of Transcript by Quality of Transcript

Transcript Quality	No	Yes
perfect	14	2
good	13	3
moderate	3	13
low	6	10

accuracy was a strong predictor of whether or not people would persist with use ($\chi^2 = 21.84$, 2df, $p < 0.0001$).

Analysis of the effect of word accuracy differences on subjects' perception of the task and of the utility of the transcript also confirmed our hypotheses H5 and H6. Subjects did perceive tasks to be easier with high quality transcripts (Hypothesis H5); $F(1,16) = 11.66$, $p < 0.002$). And subjects did perceive that high quality transcripts were more useful for carrying out the task (Hypothesis 6); $F(1,16) = 65.70$, $p < 0.00001$).

We also looked at subjects' perceptions of transcript accuracy. One very striking observation is that subjects' ratings of perfect transcripts are low. Only half the subjects thought that perfect transcripts were "very readable", and 57% thought they were "very comprehensible". When asked about this, our subjects noted the conversational nature of the transcribed speech, which included hesitations, incomplete sentences, and grammatical errors not normally found in written text. So, it could be that subjects were holding a spoken genre to a written standard in their evaluations. Despite this lack of absolute accuracy, however, subjects' relative judgments of perceived readability and comprehensibility of transcripts were highly correlated with objective quality. That is transcripts with higher word accuracy were rated as higher quality than those with lower accuracy (for comprehensibility $r(62) = 0.75$, $p < 0.0001$, and for readability $r(62) = 0.73$, $p < 0.\ 0001$).

Our exploratory investigation of which aspects of the transcript led subjects to question its accuracy also show some interesting findings. Subjects' perceptions of the lexical, syntactic and semantic characteristics of transcripts related to their judgments of transcript readability and comprehensibility. For non-perfect transcripts we conducted two analysis in which we regressed perceptions of word comprehensibility, appropriateness of syntax, appropriateness of terms, and appropriateness of proper names against perceived readability and comprehensibility respectively. Both readability and comprehensibility models were significant ($F(4,43) = 12.32$, $p < 0.000001$, and $F(4,43) = 10.08$, $p < 0.000001$). In both analyses, word comprehensibility and appropriateness of syntax were significant predictors of readability and comprehensibility. Finally, we looked at whether perceived readability and comprehensibility affected the likelihood that users would abandon the transcript. For both measures, there is a strong relationship between perceived quality and persistence of use (for readability $\chi^2 = 20.93$, 4df, $p < 0.0001$, for comprehensibility, $\chi^2 = 20.06$, $df = 4$, $p < 0.001$).

4. DISCUSSION

Our study of how ASR transcript quality affects people's ability to use these transcripts in relevance ranking and summarization tasks shows the following major findings: Overall, people completed tasks more rapidly with high quality transcripts. ASR quality was an important determinant of solution quality — but only the first time people encountered a transcript at a given quality level (controlling for time to solution). People played less speech with high quality transcripts. They were more likely to stop using low quality transcripts. People found tasks easier and transcripts more useful with high quality transcripts. High quality transcripts were also seen to be more readable and more comprehensible. While most of the effects we found were restricted to the relevance judgment task, this may be because people tended to rely upon playing rather than reading the transcript for the summary task.

Our findings must of course be interpreted relative to several possible limitations of our approach. The transcripts used in the study were generated by different recognizers, so their error patterns may have been different — both in distribution and in type of errors. The overall word accuracy of entire documents may be too crude a way to define ASR quality for a task. If different portions of a document have different error rates, this may have different effects on user performance; for example, if the transcript at the beginning of a document is accurate, the user may gain useful information about the document as a whole, even if the overall transcript quality is low. Early transcript accuracy might also motivate users to continue using the transcript, as opposed to switching to listening to the story directly. These possibilities will be considered in subsequent experimentation.

We must also consider the question of whether there is a threshold at which improvement in transcript accuracy ceases to facilitate performance; our finding that subjects are rather poor at distinguishing between accurate and somewhat inaccurate transcripts suggests this may be true.

5. REFERENCES

1. C. Buckley. Implementation of the SMART information retrieval system. Technical Report TR85-686, Department of Computer Science, Cornell University, Ithaca, NY 14853, May 1985.

2. G. Salton, editor. *The SMART Retrieval System—Experiments in Automatic Document Retrieval*. Prentice Hall Inc., Englewood Cliffs, NJ, 1971.

3. S. Whittaker, J. Choi, J. Hirschberg, and C. Nakatani. "what you see is almost what you hear: Design principles for accessing speech archives. In *Proceedings of the Fifth International Conference on Spoken Language Processing*, Sydney, 1998. ICSLP98.

4. S. Whittaker, J. Hirschberg, and C. Nakatani. All talk and all action: strategies for managing voicemail messages. In *Proceedings of CHI '98*, Los Angeles, 1998.