SCAN: Designing and evaluating user interfaces to support retrieval from speech archives

Steve Whittaker, Julia Hirschberg, John Choi, Don Hindle, Fernando Pereira, Amit Singhal
ATT Labs-Research, Shannon Labs, 180 Park Avenue,
Florham Park, NJ, 07932, USA.
+1 (973) 360 8339

stevew/julia/choi/hindle/pereira/singhal@research.att.com

ABSTRACT

Previous examinations of search in *textual* archives have assumed that users first retrieve a ranked set of documents relevant to their query, and then visually scan through these documents, to identify the information they seek. While document scanning is possible in text, it is much more laborious in *speech* archives, due to the inherently serial nature of speech. Yet, in developing tools for speech access, little attention has so far been paid to users' problems in scanning and extracting information from within "speech documents".

We demonstrate the extent of these problems in two user studies. We show that users experience severe problems with *local navigation* in extracting relevant information from within "speech documents". Based on these results, we propose a new user interface (UI) design paradigm: *What You See Is (Almost) What You Hear*, (WYSIAWYH) - a multimodal method for accessing speech archives. This paradigm presents a *visual analogue* to the underlying speech, enabling visual scanning for effective local navigation. We empirically evaluate a UI based on this paradigm. We compare our WYSIAWYH UI with a visual "tape recorder", in relevance ranking, fact-finding, and summarization tasks involving broadcast news data. Our findings indicate that an interface supporting local navigation multimodally helps relevance ranking and fact-finding, but not summarization. We analyze the reasons for system success and identify outstanding research issues in UI design for speech archives.

Keywords

Speech indexing and retrieval, field/empirical studies of the information seeking process, comparing interfaces for information access, user studies.

1. INTRODUCTION

Recently, there have been major increases in the amounts of data stored in digital speech archives. Broadcasting companies have made radio programs available, public records such as the US Congressional Debates are being archived, and large private archives of audio conferences and voicemail can be cheaply stored for subsequent reference. Such archives are potentially highly valuable, as speech has been shown to be both ubiquitous and critical for the execution of many workplace tasks [3,21]. However, these archives are currently under-utilized, in large part due to the absence of effective user-centered techniques for archival access. Although a number of speech retrieval systems have been built for TREC [20], these systems have generally paid little attention to user requirements, or to the development of UIs. We consequently lack systematic information about: the processes by which people currently access information from speech archives, general principles for designing UIs to speech archives, and methods for evaluating such interfaces. This study addresses those issues.

A natural starting point for identifying how people might access information from *speech* archives is the large body of research on *text* retrieval. Yet with few exceptions, such as Hearst [8], and the interactive track of TREC [20], text retrieval research has focused on document *search*, where the retrieval engine's goal is simply to identify a *ranked set of documents* relevant to the user's query. Subsequent scanning within these documents to actually locate information, e.g. extracting specific facts, or identifying relevant paragraphs, are behaviors generally not addressed. It is usually assumed that, for more detailed information seeking, users can easily scan and browse the retrieved texts (although Hearst's [8] *Tilebars* is an important exception).

In the context of a *speech* corpus, however, it is apparent that UIs supporting only document search are insufficient, because of the problems for users of scanning and browsing speech data. A story in the NIST Broadcast News corpus, for example, can be 25 minutes long. Given the sequential nature of speech, it is extremely laborious to scan through multiple speech stories to obtain an overview of their contents [1], or to identify specific information of direct relevance within speech [5,6]. Interfaces for accessing speech archives therefore need to support *local navigation* within "speech documents", as well as relevance based search.

The structure of the paper is as follows: we present two user studies of voicemail that: (a) examine user problems of local navigation in accessing speech; and (b) identify the strategies users employ to overcome these problems. From these studies we derive a new paradigm for the design of speech access systems: *What You See Is (Almost) What You Hear (WYSIAWYH)*. This paradigm presents a *visual analogue* to the underlying speech, enabling visual scanning for effective local navigation. We describe a new UI designed according to that paradigm. The interface is to SCAN, a system that accesses a broadcast news archive. We empirically evaluate a UI based on this paradigm. We compare the SCAN UI with a visual "tape recorder", in relevance ranking, fact-finding, and summarization tasks involving broadcast news data. Our findings indicate that an interface

supporting local navigation multimodally helps relevance ranking and fact-finding, but not summarization. We analyze the reasons for system success and identify outstanding research issues in UI design for speech archives.

2. LOCAL NAVIGATION: SCANNING AND INFORMATION EXTRACTION STRATEGIES

To identify how users currently browse and search speech corpora, we conducted two studies of voicemail access. Voicemail represents a real-world domain with experienced users who have evolved strategies for dealing with important speech data. It is therefore a good starting place for studying local navigation in speech.

We examined local navigation strategies under two controlled laboratory conditions. We gave users two types of graphical interfaces to a voicemail archive of 8 messages whose average length was about 30s [23]. Users were given two types of access tasks, derived from interviews and surveys conducted with voicemail users [22]. The tasks were to *summarize a message* or to *extract specific information* (e.g. a name) from a message. Subjects experienced serious problems with local navigation, even for a small archive of short messages. They learned the global structure of the archive but were unable to remember specific message contents. Information extraction tasks were extremely hard, particularly when multiple facts had to be retrieved: users repeatedly replayed material they had just heard, suggesting problems with remembering local message structure. In a post-hoc memory task, users also showed poor recall for message contents.

A second study [22] used a combination of interview and survey methods to investigate voicemail retrieval strategies. 148 high volume users (recipients of more than 10 messages per day) experienced two main problems in accessing voicemail: (a) *scanning* - navigating to the correct message or relevant part of the message; (b) *information extraction* - accessing specific facts from within the message. Note-taking was a key processing strategy, with 72% of users reporting that they 'almost always' took notes. Users described two different note-taking strategies: (a) *full-transcription*, attempting to produce a verbatim transcript of messages to avoid later replayings; (b) *message indexing*, abstracting only key points (such as caller name, caller number, reason for calling, important dates/times and action items). Typically, users kept originals as a backup, in case their notes were insufficient. Many users kept sequential notes, so they could use this temporal index to locate a message in their archive. Finally we identified cues used in processing voicemail, such as the importance of *intonation* to indicate or clarify speaker intentions.

Both studies illustrate the problems of local navigation with users finding it hard to scan messages and extract information from within messages. The different note-taking strategies indicate the methods that users currently employ for local navigation: *Indexing* provides an abstract overview of each message, presenting its key points and serves as a guide for archive scanning, i.e. locating one message in relation to others. *Full-transcription* provides a (labor-intensive) textual rendering of speech to facilitate subsequent information extraction. User comments indicate, however, that they prefer *not* to rely wholly on a text transcription, being concerned about losing the extra intonational information provided by the original speech.

These data together suggest general principles for improved UIs to speech archives, and indicate a potential solution to the problem of local navigation. By taking notes, users construct *visual analogues* of voicemail messages (in the form of transcripts and indices) allowing them to visually scan and index into the corresponding speech. We therefore need interfaces that: (a) address the problem of local navigation; (b) provide visual analogues to underlying speech. Furthermore, (c) these interfaces must be multimodal: people want access to the original speech, as well as this visual information.

3. THE SCAN USER INTERFACE

While the IR literature focuses on global *search*, our initial experiments show an additional critical role for local scanning and information extraction. Both tasks are particularly difficult for speech data. The SCAN (Spoken Content-based Audio Navigation) UI for accessing broadcast news data attempts to support both local and global navigation (see Figure 1). The underlying system provides access to a corpus of 47 hours of broadcast news from the NIST/DARPA test set [4]. This is a set of recorded radio and TV news. It is made up of programs such as current affairs discussions, breaking news and headlines. The stations and programs include: *NPR: All Things Considered*, *ABC: World News Tonight*, *CNN: Early Primetime News*, *NPR Market Place* (Figure 1 provides more instances of programs).

The user interface consists of three elements depicted in Figure 1, namely Search, Overview and Transcript. Search is intended to support global access to "speech documents", while Overview and Transcript elements address local navigation. We describe each UI element in turn.

3.1 Search

The SCAN interface's Search component provides access to sets of relevant "speech documents" in response to user queries. We identify these sets by applying information retrieval methods to errorful textual transcriptions of each "document", that have been generated by automatic speech recognition (ASR). To generate the ASR transcriptions, we first segment the speech into *paratones* ("audio paragraphs"), using acoustic information [9], classify the recording conditions for every paratone (narrowband or other) and apply ASR to each. We combine results for each paratone so that for every "speech document" we have a corresponding (errorful) ASR transcript. Terms in each transcript are then indexed for retrieval by the SMART IR engine [2, 17]. When the user types a query ("What is the status of the trade deficit with Japan?") into the text box at the top of the browser (labeled Query), the system searches the errorful transcripts for relevant documents. The search results are depicted in the Results panel immediately below, as a relevance-ranked list of 10 "speech documents", corresponding to the 10 most relevant news stories. We also present additional information about each news story, including program name, date, story number (to distinguish the multiple stories occurring within a program), relevance score, length (in seconds), and total *hits* (number of instances of query words). The user selects a story by clicking on it.

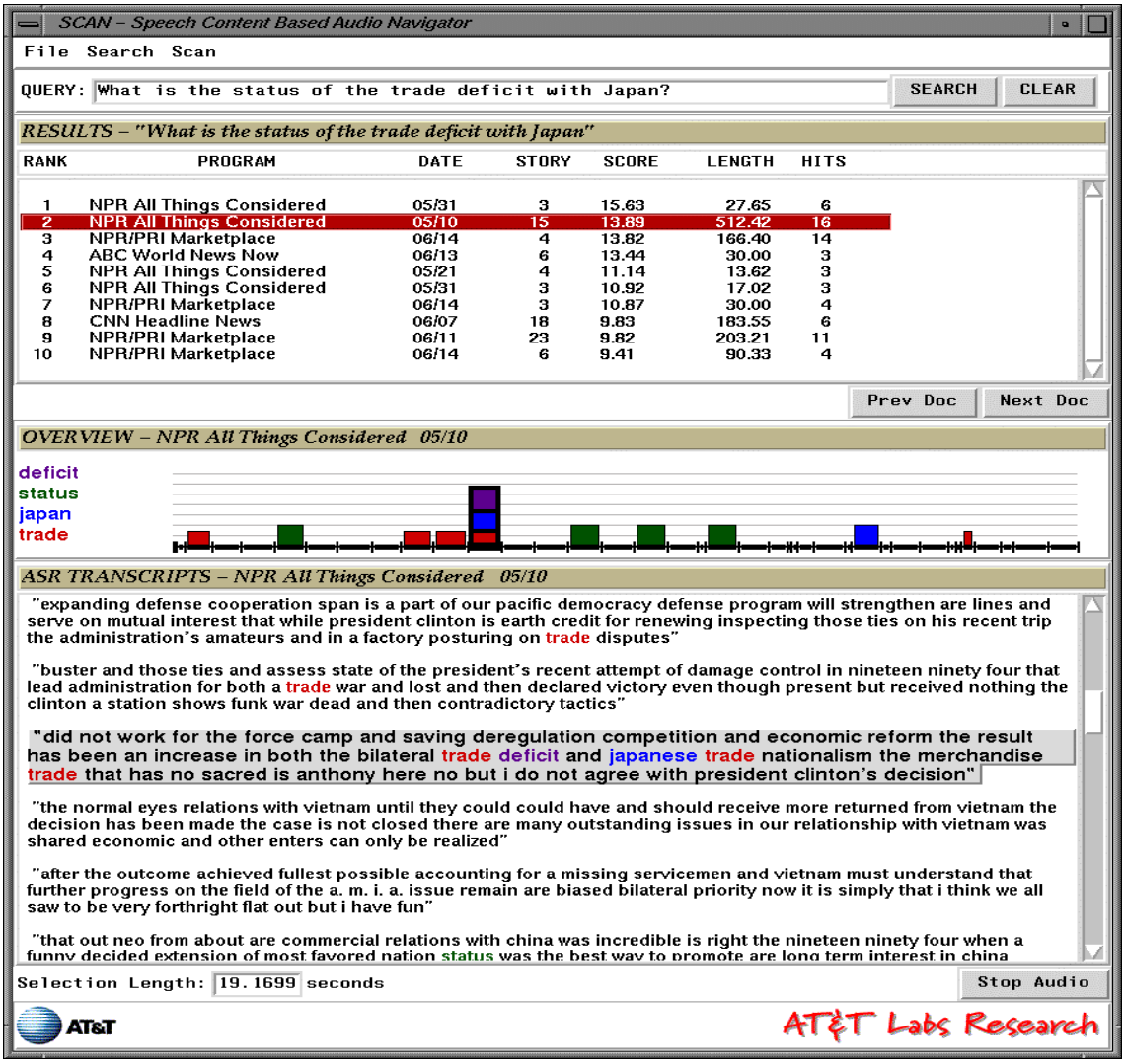

Figure 1: The SCAN user interface

3.2 Overview

The Overview component provides high level visual information about individual "speech documents". Users can rapidly *scan* this to locate potentially relevant audio regions. It displays which query terms appear in each paratone of the story. Each query word is color coded, and each paratone is represented by a vertical column in a histogram. Thus the word "trade" occurs in the second paratone and hence in the second histogram column. The width of the histogram bar represents the relative length of that paratone. The height of each bar in the histogram represents the overall query word *weights* (the term weighted indices of the query terms for the corpus normalized for the paratone length). Different query terms are combined within the same histogram column, so that column 11 in the Overview in Figure 1 contains instances of each of the words, "trade", "japan", and "deficit". The co-occurrence of these terms suggests a potentially highly relevant region within the "document". Users can also locate specific query terms by examining color distributions across paratones. A similar technique is used for textual documents in [8]. Users can directly access the speech for any paratone by clicking on the corresponding column of the histogram. Selecting a column initiates play from the start of the corresponding paratone. This component also supports global comparison between "speech documents". Comparing Overviews for multiple documents can reveal which "documents" have a greater density of query terms and hence contain potentially more relevant regions.

3.3 Transcript

The SCAN ASR Transcript supports *information extraction*, providing detailed, if sometimes inaccurate, information about the contents of a story. These are the same ASR transcripts that were used to support search. The transcript panel displays a transcription of the selected story. The transcript in Figure 1 has been scrolled so that the first visible paragraph does not correspond to the start of the "speech document". Because the transcript has been generated automatically, it usually contains errors (in paragraph 4 of the transcript in Figure 1, "to normalize" is transcribed as "the normal eyes"). When the speech recognizer makes errors, they are deletions, insertions and substitutions of

the recognizer's vocabulary, rather than the types of non-word errors that are generated by OCR. If the target speech contains large numbers of words that are not in the recognizer's vocabulary (the Out-of-Vocabulary Problem), this leads to multiple word substitution errors. In addition, recognition errors often cascade: the underlying language model explicitly models inter-word relationships, so that one misrecognition may lead to others. Finally function words tend to be misrecognized more than content words.

Query terms in the transcript are highlighted and color-coded, using the same coding scheme used in the Overview panel (e.g. the word "trade" is highlighted in paragraph 1). Users can play a given paratone by clicking on the corresponding paragraph in the transcript.

The transcript has several potential functions. First, in regions where it is mostly accurate, users can find relevant information simply by reading -- without listening to the audio. Like the overview, it supports rapidly visual scanning to find relevant regions in the audio. The transcript also provides *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 can help users assess the likely accuracy of transcript, search and overview information. For example, bizarre phrases like "buster and those ties and assess state…." (beginning of paragraph 2) indicate the transcript is inaccurate. They also suggest that query terms in the overview may have been misrecognized. If errors are prevalent, then users may rely more on the speech than transcripts.

3.4 Player

The current SCAN interface also provides random access to "speech documents" using a simple play bar representing a single story. The UI is analogous to a tape-recorder. Users can insert the cursor at any point in the bar to indicate where to begin playing. Start and stop audio buttons are available to control play and may also be used for the overview and transcription panels. The player is not visible in Figure 1, but users can scroll down to it below the Transcript.

3.5 "What You See Is (Almost) What You Hear" Principles For Speech Retrieval UIs

Together, the elements of the UI support a new paradigm for speech retrieval interfaces: *"What you see is (almost) what you hear"* (WYSIAWYH). A key principle of this design paradigm is to provide a *visual analogue* to the underlying speech, using text formatting (such as headers and paragraphs) to exploit well understood text conventions in order to present useful local context for speech browsing. By depicting the abstract structure of "audio documents" in the Overview, and by providing a formatted Transcript, we hope to make visual scanning and information extraction fast and effective, addressing the problems of local navigation identified in our user studies.

While we believe this visual information will be helpful in local navigation, however, we do not think it will eliminate the need to access the original speech. There are two reasons why visual information alone is insufficient. First, ASR errors mean that the Transcript frequently diverges from the underlying speech. There were about 30% word errors for the SCAN corpus, and it seems unlikely that error-free ASR will be available within the foreseeable future. This is especially true for domains like news that have spontaneous speech, unforeseen recording conditions, and large numbers of out-of-vocabulary items (because of constantly changing content). A second reason for needing the original speech is the importance of intonation. Our voicemail users stressed the importance of preserving original speech messages, so that they could fully interpret their handwritten notes. Voice quality and intonational characteristics are lost in transcription, and intonational variation has been widely shown to change even the semantics of simple sentences [12]. So, transcription alone is unlikely to be an effective substitute for multimodal access.

4. EVALUATION STUDY

4.1 Method

To test our hypotheses about the usefulness of our WYSIAWYH paradigm in supporting local browsing, we compared the SCAN browser, with a control interface that supported only *search.* This control gave users only the search panel and the player ("tape recorder") component described above. Users used the search panel to find stories, as with the SCAN browser, but had only the random access player ("tape-recorder") for browsing within "documents".

From our previous user interviews and experiments [22,23], we developed a task taxonomy for retrieval. We wanted to compare different retrieval situations along several important task dimensions, including: making global judgments about sets of "speech documents", locating specific information from within a "document", and extracting the overall gist of a "document". We therefore collected data experimentally to compare the two interfaces on the following 3 tasks:

- *relevance judgments* - compare five news stories to determine which was most relevant to a given topic (e.g. *"how good was Valujet's safety record prior to the Florida accident?"*);
- *fact-finding* - extract factual information from a story to answer a specific question (e.g. "*who starred in the Broadway musical 'Maggie Flynn'?"*);
- *summarization* - produce a 4-6 sentence summary of a given story (e.g. for a story on a bombing in Manchester).

Because our focus was on browsing behavior rather than search, we wanted all users to access the same set of stories. So, rather than spontaneously generating their own queries, users were given the queries to type in to the search panel for each task. In the relevance task, users were asked to consider five stories, but for the fact-finding and summary tasks, they only accessed one story. We attempted as far as possible to normalize story length across the 3 tasks.

The experimental design was randomized within subjects. Twelve subjects were given a total of 12 questions each (4 of each of the 3 task types). For half the questions they used the SCAN browser, and the control browser for the other half. For each question we measured *outcome* information: *time to solution* and *quality of solution* (as assessed by two independent judges). We also collected information about the *processes* by which people answered each question: number, type, and duration of browsing and play operations. We also collected subjective data. After each question we had subjects judge task difficulty. Because we were interested in browsing strategies and processes, we encouraged

subjects to "think aloud" as they carried out the tasks, and we recorded their statements. We also administered a post-test survey probing relative task difficulty, how well the SCAN UI supported each task, overall browser utility, how the browser might be improved, quality of the transcript, and what factors led users to evaluate the transcript positively or negatively.

	Variable	SCAN (mean)	Control (mean)	Prediction confirmed?
Outcome	Time to solution (secs.)	414.7	500.7	Yes
	Solution quality (maximum score = 100)	78.3	66.7	Yes
Subjective ratings	Perceived task difficulty (scale of 1-5, 1 is "hard")	3.51	2.77	Yes
	Perceived browser utility (scale of 1-5, 1 is "very useful")	1.67	4.08	Yes
Process measures	Number of operations	10.2	6.0	No
	Amount of audio played (secs.)	108.2	275.3	Yes

Table 1: Effects of browser type on local navigation

4.2 Hypotheses

- *Supporting local navigation*: We expected the SCAN browser to support local navigation better than the control for the two *outcome* measures (time to solution and solution quality). Users should evaluate tasks as easier using the SCAN browser, and rate the SCAN browser as better overall. We expected our *process measures* to show the SCAN browser supported more efficient retrieval: users should require fewer operations to complete tasks, and play less audio with the SCAN browser;
- *Task differences*: In terms of solution time, solution quality, and perceived task difficulty, we expected the fact-finding task to be easier than the summary task, which in turn would be easier than the relevance task, based on the amount of information users had to access to perform the task. Fact-finding requires access to part of a single document, whereas summaries require access to an entire document, and relevance judgements require access to multiple documents;
- *ASR Transcript quality*: We predicted that ASR transcript quality (as assessed by word error rates) would influence performance. High quality ASR should improve solution quality, reduce solution time, reduce perceived task difficulty and reduce the amount of speech played.

4.2 Results

4.2.1 Supporting local navigation

Users performed better with the SCAN browser than the control. We conducted multiple independent ANOVAs with users, task type, and browser type as the independent variables. The dependent variables in each ANOVA were: time to solution; solution quality; perceived task difficulty; users' rating of browser usefulness. Results are summarized in Table 1, and the data for solution quality, solution time and perceived task difficulty depicted in Figures 2, 3 and 4. Our predictions were confirmed for outcome measures: solution time ($F_{(1,72)} = 7.05$, $p < 0.01$), solution quality ($F_{(1,72)} = 8.40$, $p < 0.005$), and also for subjective ratings: perceived task difficulty ($F_{(1,72)} = 19.50$, $p < 0.0001$), perceived browser utility ($F_{(1,72)} = 35.04$, $p < 0.00001$).

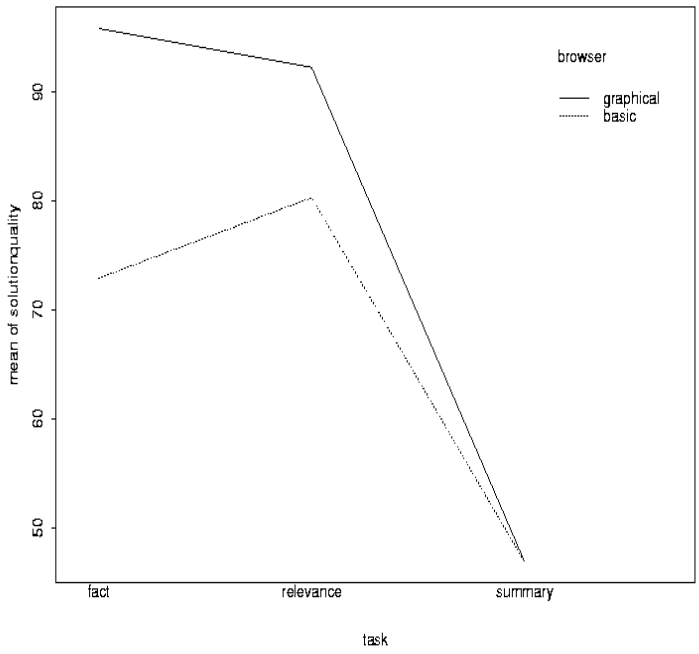

Figure 2: Effects of browser and task on solution quality

Qualitative data: How did the SCAN UI provide support for local navigation? With only the "tape-recorder" browser, users reported several problems. Although listening to an entire story was highly tedious, users were forced to do so because they lacked clues to information structure: *"It's so painful to try and find specific information that I'm going to surrender and listen to the whole thing"*. Lack of structural information meant they could not skip over parts of the story, or they ran the risk of missing significant information. *"It doesn't help to skip forward because I don't know where this section ends, so it means that I have to listen to the whole thing"*. Users also complained that with the "tape-recorder" that on occasion audio came "*too fast*", so they had to listen to passages multiple times. *"I actually played the answer but I didn't hear it. I realized later that I'd heard it and then had to go back"*. Going back to relevant regions was also a major problem: *"I missed an explanation and I knew that it was slightly before halfway so I moved the cursor one third of the way back and listened to all that again"*.

The SCAN UI addressed these problems. There were 3 ways that it reduced the amount of speech subjects played. The UI enabled global relevance judgments based on the overview or the transcript alone: *"Listening is clearly too slow – I don't want to listen to every story, so I'm just looking here (in the overview) for stories that talk about the topic in a broad sense"*. If the transcript was high quality, users could avoid listening entirely: "*I'm just reading the transcripts to determine relevance – even without listening to it, I'm sure that story 4 is the best match"*. Even errorful transcripts, however, like the overview, gave users greater precision in judging which parts of the "speech

document" they needed to play. "*For scanning, the transcript is really useful. Also I can just buzz around the overview to find when I'm in the right area … I'm going to play the places where the transcript is awful"*.

The SCAN interface was often used multimodally, with simultaneous playing and reading. People scrolled forward and backward around the paratone they were playing, reading the surrounding transcript paragraphs to obtain context for what they were hearing. On other occasions they would listen to an important part of the "speech document" (e.g. the beginning) to set some context, while scrolling the transcript to visually scan the remainder of the story. *"I was trying to get the story opening through audio and then look ahead for the rest of it in the transcript"*.

Process data: The process data (number, type and duration of play operations) support these qualitative observations. Our prediction that the SCAN UI would prove more efficient was confirmed. With the SCAN browser, people played much less speech ($F_{(1,72)}$=106.07, $p < 0.000001$). However, contrary to our expectations, we found that subjects used more operations with the SCAN browser than with the control ($F_{(1,72)}$ =15.69, $p < 0.0001$). User behavior suggested the reason: in the control condition, with no effective means of scanning, users often played a "speech document" from beginning to end. In contrast, SCAN users might play brief parts of several regions within a "document" to quickly identify relevant portions, and then sometimes listen to these multiple times.

4.2.2 Task differences

The results did not support our predictions. There were main effects for task for two of our dependent variables: solution time ($F_{(2,72)}$ = 19.0, $p < 0.00001$), solution quality ($F_{(2,72)}$ = 40.9, $p < 0.000001$), but not for perceived task difficulty ($F_{(2,72)}$ = 2.94, $p > 0.05$). The effects of task and the impact of the browser, for each of these variables are shown in Figures 2, 3 and 4. Planned comparisons showed that solution time was lower for fact-finding than relevance and both were faster than summaries. For solution quality, fact-finding was equivalent to relevance judgments, and both were better than summaries. Furthermore, there were interactions between task and browser: for solution time ($F_{(2,72)}$ = 2.92, $p < 0.05$), and solution quality ($F_{(2,72)}$ = 3.62, $p < 0.05$), with the SCAN browser producing higher quality, quicker solutions for fact-finding and relevance tasks but not for summaries. For perceived task difficulty, there was a significant interaction ($F_{(2,72)}$ = 4.47, $p < 0.02$), with the browser only affecting the relevance tasks.

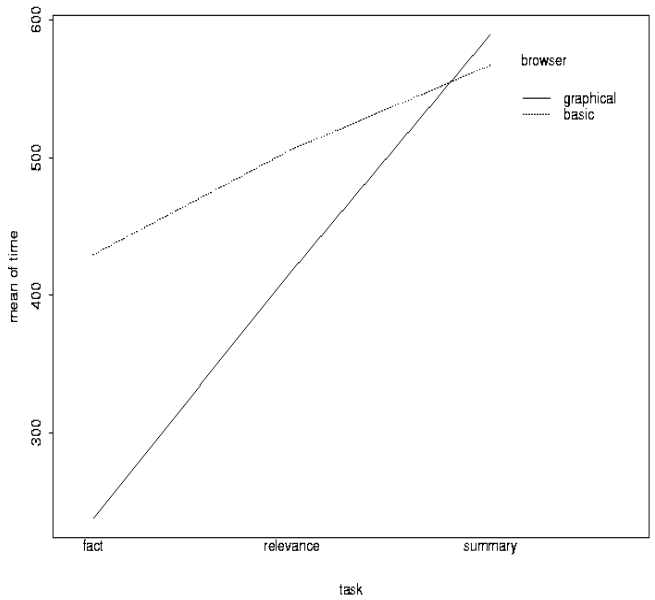

Figure 3: Effects of browser and task on solution time

Why was the summary task so hard, and why did SCAN local browsing capabilities fail to improve performance? From user behavior and comments, we conclude that the SCAN overview failed to improve the summarization task for two reasons. First, the even distribution of query terms that often occurred throughout a story provided no clue to which particular term-highlighted region provided the best summary information. In consequence, users were unable to focus their activities on specific parts of the story. Second, highly relevant regions for summarizing were sometimes not highlighted at all because synonyms were used instead of the actual query terms: *"I'm not going to zoom in on various paragraphs (from the overview) because the whole story is about the topic"*. As for the transcripts, most users felt that they did not offer accurate enough information for summarizing. *"First I thought the transcript would help, but it didn't get the slant of the story. The transcript only helps with information extraction. … To get the whole slant I need to listen to it all"*. The transcription errors also disrupted a smooth reading of the story. *"I just couldn't parse it, everything was so disjointed I couldn't make sense of it"*.

4.2.3 Effects of ASR quality

ASR quality varied significantly among the "documents" retrieved, from a maximum of 88% words correctly recognized to a minimum of 35%, with a mean of 67%. We correlated ASR transcript quality (percentage of words correctly recognized) with process and outcome measures for summary and fact-finding tasks.. It is unclear how to allocate a single ASR quality score to multiple documents, so the relevance task was not included in this analysis. We also restricted the analysis to the SCAN condition – the only one in which the transcript was available.

As we predicted, 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 transcriptions, they 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 transcription 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 reduced playing with 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 the SCAN UI did not help with the summary task.

Why did transcript quality not affect outcome and process measures more directly? Consistent with our earlier results we found some *task-specific effects*, with no influence of transcript quality on summaries. It may also be that our ASR accuracy measure was too crude to affect user behavior. Our measure was for *overall* ASR quality for entire "documents". It may be, however, that we need to measure ASR in *specific regions* of the "document". For example, if the ASR at the beginning of the "document" is accurate, this not only gives the user useful contextual information for understanding the remainder of the "document", but also motivates them to continue using the transcript, as opposed to switching to listening to the story directly. Future work needs to devise more local measures of ASR quality to examine such effects.

5. CONCLUSIONS

This research has identified a new problem in UIs for speech retrieval, i.e. support for local navigation. We have outlined a new paradigm for interfaces to address this (WYSIAWYH: *"What you see is (almost) what you hear"*), where a multimodal interface provides a visual analogue and straightforward indexing into the underlying speech. Our user evaluation showed that we made have considerable progress in addressing the problem of local navigation. Comparing the WYSIAWYH-based interface with a simple visual tape-recorder interface showed superiority for WYSIAWYH for fact-finding and relevance judgment tasks. The overview and transcript elements of the SCAN UI offer multiple methods for users to reduce the problems of time-consuming serial access to speech. The interface allows users to: use overview and transcript information to avoid playing entire "speech documents" they judge to be irrelevant; extract information from the transcript without playing anything at all; and, finally, if playing is necessary, focus on paratones they judge to be most relevant to their task. Users can also access information multimodally by listening to relevant paratones and reading the relevant transcript simultaneously.

How does WYSIAWYH relate to other interface work on speech access? Similar techniques, using visual handwritten notes to index into recorded speech, have been successful for accessing personal speech data [15,19,24]. Several video retrieval systems have presented key video frames to provide visual overviews to video programs [7,18]. Other broadcast news and meeting recording systems present high level topic or speaker switching information [10,11,16]. However, with the exception of [11] these latter UIs have not been evaluated on access tasks with real users.

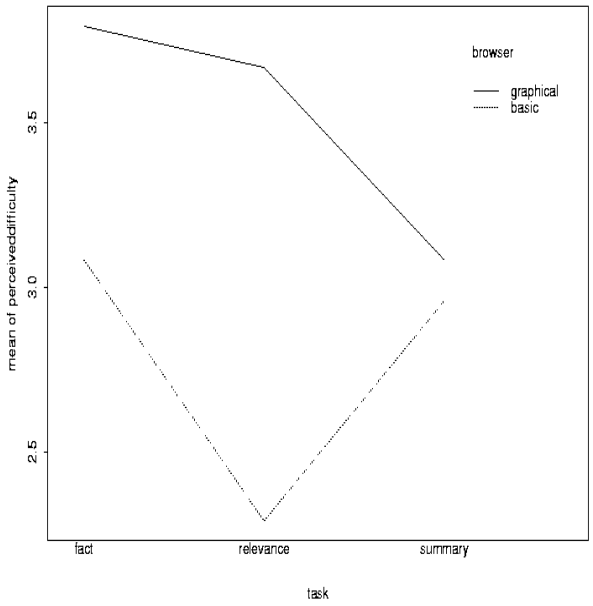

Figure 4: Effects of browser and task difficulty on perceived task difficulty

We find it significant in our studies that the multimodal SCAN interface is beneficial only for certain tasks, such as fact-finding and relevance ranking. For these tasks, users were able to exploit the overview and transcript to extract local facts or to make global judgments. However the summary task required access to the specific content of an entire document. All sections of the document were potentially relevant to the summary. It was therefore difficult to judge what was important information without a good transcription of the document, or actually listening to what was said. In general, the transcript was too inaccurate to allow users to identify important summary information, and they were forced to play entire documents. Even when the word accuracy rate is as high as 88%, this problem persists. How might we then improve summarization? First, of course, even higher word accuracy might help, but it is unclear from our data how close to perfect a transcript must be for subjects to trust it fully for summarization. Second, automatic speech summarization might provide a starting point for human summary creation, although, again, it is not clear how 'good' this must be, or how people would subjectively judge its quality for this purpose [14]. It may be that, given the laborious nature of speech access, even poor automatic summaries are preferable to playing entire stories. Third, skimming techniques [1] that use acoustic information to identify areas of high relevance might provide shortcuts to summarization similar to automatic summarization, with comparable potential weaknesses. Lastly, we might explore UI techniques that control playback by allowing speeded up playback, or access using structural properties (e.g. speaker or topic shifts).

Finally our data have implications for basic measures and evaluations in information retrieval. The problem of local navigation arises from the fact that supporting relevance at the "speech document" level is insufficient to address retrieval

problems with speech. We need to move away from a purely document-level view to successfully address speech access. Our studies showed generally better performance with the SCAN UI, which provides *within-story* relevance information. This indicates that we need to generate relevance metrics that operate within stories, to support notions of local relevance. We also need to devise new evaluation tasks, such as the ones we used here, that draw on the requirements for local information. Finally for speech retrieval, our data suggest that traditional document level relevance judgments may be affected by document length. Users stated a preference for shorter documents, because of the tedium of accessing speech. If future experiments support this observation, then future document level metrics may need to modify relevance metrics depending on the retrieval medium, with speech retrieval relevance showing greater weighting for shorter documents.

6. REFERENCES

[1] Arons, B. Interactively skimming speech. Unpublished PhD thesis, MIT Media Lab, 1994.

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

[3] Chapanis, A., Ochsman, R., Parrish, R. and Weeks, G. Studies in interactive communication: I. The effects of four communication modes on the behavior of teams during cooperative problem-solving. Human Factors, 14:487-509, 1972.

[4] Choi, J., Hindle, D., Hirschberg, J., Magrin-Chagnolleau, I., Nakatani, C., Pereira, F., Singhal, A., Whittaker, S., SCAN - speech content audio navigator: a system overview, Proceedings of the International Conference on Speech and Language Processing, 1998, (forthcoming).

[5] Gould, J. Human factors challenges – The speech filing system approach. In ACM Transactions on Office Information Systems, 1(4), October 1983.

[6] Haas, C., and Hayes, J. What did I just say? Reading problems in writing with the machine. In Research in the teaching of English, 20 (1), February 1986.

[7] Hauptmann, A. and Witbrock, M. Informedia: News-on-Demand Multimedia Information Acquisition and Retrieval. In Maybury, M., ed. Intelligent Multimedia Information Retrieval, AAAI Press, 1997.

[8] Hearst, M. Tilebars: Visualization of term distribution in full text information access. In *Proceedings of CHI'95 Human Factors in Computing Systems*, ACM Press, New York, 1995.

[9] Hirschberg, J., and Nakatani, C., Acoustic indicators of Topic segmentation, Proceedings of the International Conference on Speech and Language Processing, 1998, (forthcoming).

[10] Jones, G., Foote, J. T., Sparck Jones, K. and Young, S. J. Retrieving spoken documents by combining multiple index sources. In Proceedings of SIGIR-96, Zurich, 1996.

[11] Kazman, R., Al-Halimi, R., Hunt, W., and Mantei, M. Four paradigms for indexing video conferences. In IEEE Multimedia, 3(1),63-73, 1996.

[12] Ladd, A. The structure of intonational meaning. Indiana University Press, Bloomington, Ind: 1980.

[13] Magrin-Chagnolleau, I., Parthasanathy, S., and Rosenberg, A., Automatic labeling of broadcast news into different sound classes using gaussian mixture models, manuscript in preparation.

[14] Mitra, M., Singhal, A., and Buckley, C., Automatic text summarization by paragraph extraction, Proceedings of the ACL Workshop on Intelligent Scalable Text Summarization, 39-46, 1997.

[15] Moran, T. P., Palen, L., Harrison, S., Chiu, P., Kimber, D., Minneman, S. Van Melle, W. and Zellweger, P. "I'll get that off the audio". In Proceedings of CHI-97, pp. 202-209, 1997.

[16] Oard, D. Speech based information retrieval for digital libraries. In Proceedings of AAAI Spring Symposium On Cross Language Text and Speech, 1997.

[17] Salton, G., (ed.), The SMART Retrieval System - Experiments in Automatic Document Retrieval. Prentice Hall, Inc., Englewood Cliffs, NJ, 1971.

[18] Shahraray, B., and Gibbon, D. C. Automated authoring of hypermedia documents of video programs. In Proceedings of the Third ACM Conference on Multimedia, 401-409, San Francisco, 1995

[19] Stifelman, L. Augmenting real-world objects: a paper-based audio notebook. In Proceedings of CHI-96, 199-200, 1996.

[20] Voorhees, E. M., and Harman, D. K., Overview of the seventh Text Retrieval Conference (TREC-7), in Voorhees, E. M., and Harman, D. K. (eds.), Proceedings of the Sixth Text Retrieval Conference (TREC-7), 1998, forthcoming.

[21] Whittaker, S., Frohlich., D., and Daly-Jones, O. Informal workplace communication: what is it like and how might we support it? In Proceedings of CHI'94 Human Factors in Computing Systems, 130-137, ACM Press, New York, 1994.

[22] Whittaker, S., Hirschberg, J. & Nakatani, C. All talk and all action: Strategies for managing voicemail data. In Proceedings of CHI'98 Human Factors in Computing Systems, ACM Press, New York, 1998.

[23] Whittaker, S., Hirschberg, J. & Nakatani, C. Play it again: a study of the factors underlying speech browsing behaviour. In Proceedings of CHI'98 Human Factors in Computing Systems, ACM Press, New York, 1998.

[24] Whittaker, S., Hyland, P, and Wiley, M. Filochat: In Proceedings of CHI'94 Human Factors in Computing Systems, 271-277, ACM Press, New York, 1994.