Look or Listen: Discovering Effective Techniques for Accessing Speech Data

Steve Whittaker and Julia Hirschberg

University of Sheffield, Information Studies, 211 Portobello, Sheffield, S1 4DP, U.K.
Tel: *+44 114 222 6340*
Fax: *+44 114 278 0300*
Email: *s.whittaker,julia@shef.ac.uk*
URL: *www.shef.ac.uk/~is/people/whittake/index.html*

Commercial interfaces for accessing digital speech data are based on 'tape recorder' metaphors. However, such interfaces make it highly laborious to access complex speech data. The absence of effective interfaces is a major obstacle to exploiting the increasing number of speech archives now available online. More novel research interfaces provide potentially more effective access by presenting *visual* or *textual* indices into the underlying speech data. The current experimental study evaluates the utility of these newer techniques compared with a 'tape recorder' interface. We compare: (a) High-level Visual Overviews showing the distribution and density of user query terms; (b) Textual Transcripts generated using state of the art ASR; (c) a tape recorder baseline. Laboratory tests showed that, contrary to our expectations, high-level visual information proved more useful than textual information, although both perform better than a tape-recorder baseline. Visual overviews enable users to quickly identify relevant regions to be played. In contrast, Textual transcripts can mislead users who try to extract detailed information solely by reading the transcript, without listening to the underlying speech.

Keywords: Speech browsing, speech recognition, transcripts, visualisation, laboratory study, evaluation, speech-as-data, multimedia access.

1 Introduction

Recently, there have been major increases in the amounts of data stored in digital speech archives. Broadcasting companies have made radio programmes 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. Furthermore, better tools could exploit other types of speech record, such as recordings of meetings [Arons,1994], or telephone calls [Hindus et

al., 1993, Wilcox et al., 1994]. However, these archives are currently under-utilised, in large part due to the absence of effective user-centered techniques for speech access.

Browsing speech is difficult because of limitations in human ability to extract information from sequential media. In contrast, experiments on extracting information from texts show that people do not read these sequentially or in their entirety. Instead they quickly visually scan text, focusing on content as opposed to function words, exploiting structural information (such as headings) and looking for key words to find regions that merit detailed further processing [Askwall,1985, Oakhill and Garnham, 1988].

Such a strategy is not possible with speech, which has to be processed sequentially. The absence of explicit speech structure makes it hard to control information processing to proactively focus on important, and ignore irrelevant information. Furthermore, when users try to scan ahead in speech, they often experience difficulties in remembering what they have already listened to. For example, a study of information extraction from a 4-minute corpus of short voicemail messages using a 'tape-recorder' interface showed that users repeatedly replayed material they had already heard [Whittaker et al., 1998a]. This suggests people have problems with remembering both the structure of the message and also the details of what they have just heard. Clearly these access problems will be exacerbated when accessing large databases containing hundreds of hours of speech.

However, a number of recent speech browsing tools potentially overcome these limitations of sequential access. These tools rely on constructing different types of external indices into the speech record. These indices enable users to focus on relevant speech regions for detailed processing, while ignoring less relevant information. Two main types of index have been developed: high-level visual and textual indices. High-level visual representations include information about: speakers [Hindus et al., 1993, Oard et al., 1997, Kazman et al., 1996], emphasis [Arons, 1994, Wilcox et al., 1994], or visual events such as key frames in accompanying video [Christel et al., 1998, Drucker et al., 2002, Kazman et al., 1996]. The second technique of constructing textual indices works in the following way: textual indices are derived by applying automatic speech recognition (ASR) to recorded speech. The textual transcript is time-aligned with the corresponding speech. Even though the ASR transcript contains errors where the original speech is misrecognised, it can still be effective for accessing the underlying speech. Users can read or search the transcript, select relevant regions and play the related speech [Kazman et al., 1996, Whittaker et al., 1999, 2002].

There have been several evaluations of combined textual and high-level visual indices [Kazman et al., 1996, Whittaker et al., 1999, 2002]. These found objective benefits to combining both high-level visual and textual indices when compared with a 'tape-recorder' style interface. Of course, however, the design of these

studies makes it hard to determine which type of index - textual or high-level – was more useful in supporting information extraction from speech.

Our study attempts to compare the utility of textual and high-level visual indices, to determine which technique is more useful for accessing speech. As a control, we evaluate each of these against a baseline 'tape-recorder' UI that supports sequential access. Our expectation was that textual indices will be more effective than visual ones, because the information presented in visual interfaces is a *subset* of that provided by textual ones. Contrary to our predictions, we find that high-level visual indices are more effective than automatic textual indices, although both are better than a 'tape-recorder' UI. Both supported browsing better than the tape-recorder interface. We conclude by discussing the design and theory implications of our findings, and provide an explanation for these unexpected results.

2. Overall System

We compare three different types of interfaces for speech browsing in the context of a system that allows querying and retrieval of a large archive of speech "documents" such as news stories [Garofolo et al., 2000]. Specifically, we contrast the effects on browsing of providing: (a) high level visual structural information relevant to a user's query; (b) textual information provided by an ASR generated transcript of the speech; (c) a tape recorder player.

Prior research shows that ASR accuracy is an important determinant of speech browsing performance using transcripts [Stark et al., 2000]. Obviously, with perfect ASR transcripts users can extract information directly from the transcripts without needing to play them. And if transcript quality is poor then users have to rely on playing the entire story. In this study, we used state of the art ASR. Our mean word error rate (28%) matched that for this class of data in the latest public evaluation [Garofolo et al., 2000]. Other recent evaluations show that this level of ASR quality is state of the art for other classes of spontaneous speech data, e.g. voicemail and human-human dialogues [NIST, 2003].

To make our experimental comparisons, we built a modular system that allowed us to enable and disable various UI components (see Fig. 1). We first describe the overall system and then present the different experimental systems that people used in the three conditions.

2.1 System Data, Segmentation and Querying

The system provides access to broadcast news from the NIST/DARPA test set [Garofolo et al., 2000]. This dataset consists of recorded radio and TV news. It is made up of programmes such as current affairs discussions, breaking news and headlines. The stations and programmes include: NPR: All Things Considered, ABC: World News Tonight, CNN: Early Primetime News, NPR Market Place.

Our system first transcribes the speech in order to generate the high level visual and textual representations. To do this, we segment the speech into "audio paragraphs", using acoustic information, classify the recording conditions for every audio paragraph (narrowband or other) and apply relevant acoustic and language models to each. Our recogniser uses a standard time-synchronous beam search algorithm operating on a weighted function transducer representing the context-dependency, lexical and language model constraints and statistics of the recognition task. Context-dependent phones are modelled with continuous density, three state, left to right hidden Markov models. State densities are modelled by mixtures of up to 12 diagonal covariance Gaussians over 39-dimensional vectors [Bacchiani et al., 2001].

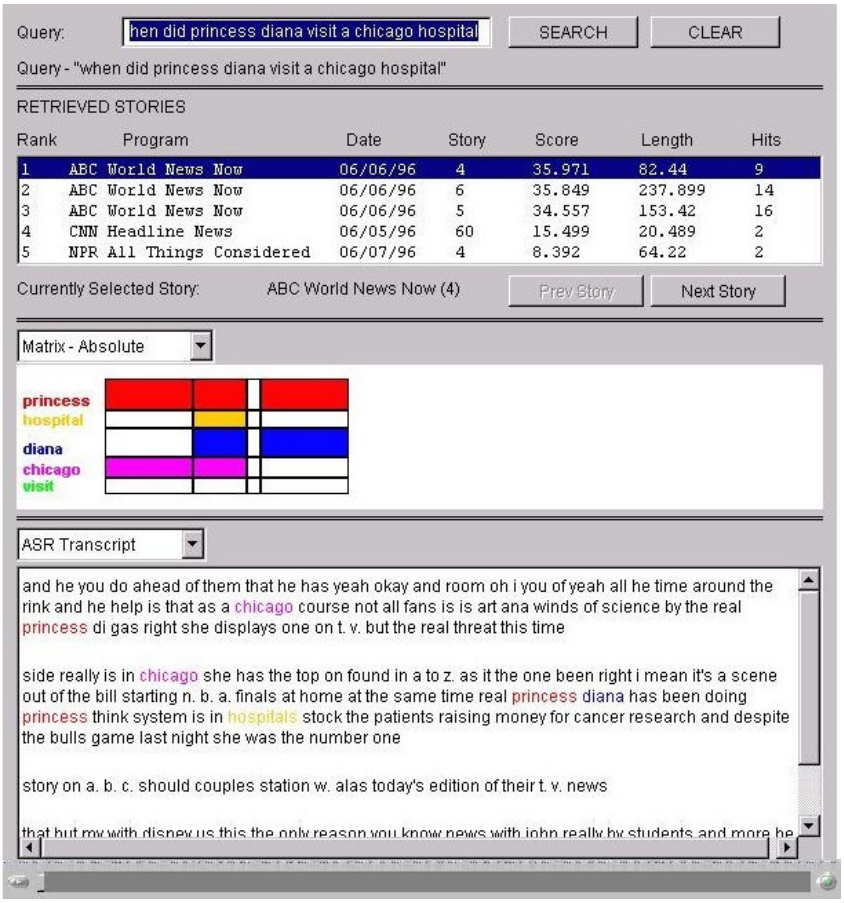

Figure 1 – Modular User Interface

We concatenate ASR results for each audio paragraph so that for every "speech document" we have a corresponding (errorful) ASR transcript. As mentioned above word error rates averaged 28% in this experiment. When the speech recogniser makes errors, they are deletions, insertions and substitutions of the recogniser's vocabulary, rather than the types of non-word errors that are generated by OCR. So, if the target speech contains large numbers of words that are not in the recogniser's vocabulary, 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 misrecognised more than content words.

Terms in each transcript are indexed for retrieval by the SMART IR engine [Salton, 1971]. When the user types a query ("When did Princess Diana visit a Chicago hospital?") into the Search box at the top of the browser, the system searches the errorful transcripts for relevant documents (see Fig.1). Search results are depicted in the panel immediately below, as a relevance-ranked list of 5 "speech documents", corresponding to the 5 most relevant news stories. The user selects a story by clicking on it.

2.2 Visual Overview

The Visual Overview component is intended to provide high-level visual information about individual "speech documents". Users can rapidly scan this to locate potentially relevant audio regions within a story. It displays which query terms appear in each audio paragraph of the story. Each query word is colour-coded, and each audio paragraph is represented by a vertical column in a visual matrix. A similar technique is used for textual documents in [Hearst, 1995]. Thus the word "chicago" occurs in the first and second audio paragraph and hence in the first and second matrix columns. The width of the matrix columns represents the relative length of each audio paragraph. The occurrences of different query term within a given audio paragraph are shown as blocks within the same column. For example, column 2 in the Overview in Fig. 1 indicates that audio paragraph 2 contains instances of each of the words, "princess", "diana", "chicago", and "hospital". Such co-occurrence of several query terms suggests a potentially highly relevant region within the "document".

Users can also examine the distribution of specific query terms by examining colour distributions across audio paragraphs. Most importantly, the visual index can be used to access speech. Users can directly access the speech for any audio paragraph by double clicking on the corresponding column. Selecting a column initiates play from the start of the corresponding audio paragraph. The Overview also supports global comparison between "speech documents". Visually comparing Overviews for multiple documents can suggest which have a greater density of query terms and hence contain potentially more relevant regions.

It is important to note that the success of the visualisation depends on speech recognition quality. If key terms are incorrectly recognised, they will not be displayed in the Overview.

2.3 Transcript

The ASR Transcript is intended to provide detailed, if sometimes inaccurate, textual information about the contents of a story. It uses the same ASR transcripts that are used to support search. The Transcript panel displays a transcription of the selected story. Because the transcript has been generated automatically, it usually contains errors (e.g. in paragraph 2 of the transcript in Fig. 1, the first word "Di" (as in Princess Di) is transcribed as "side").

Query terms in the Transcript are highlighted and colour-coded, using the same coding scheme used in the Overview (e.g. the words "chicago" and "princess" are highlighted in Fig. 1, paragraph 1 of the Transcript). Users can play a given audio paragraph by double-clicking on the corresponding paragraph in the Transcript. Alternatively they can select a paragraph and click on the player to play that paragraph.

The Transcript has several potential functions. First, in regions where it is 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 audio paragraph 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 "story on abc should couples station…." (beginning of paragraph 3, Fig. 1) indicate the Transcript is inaccurate suggesting to users that they should rely more on the audio rather than the Transcript.

2.4 Player

The player is shown at the bottom of Fig. 1. The UI is analogous to a tape-recorder. Like most state of the art commercial speech players, it presents a simple play bar that in this case represents a single story. Users can insert the cursor at any point in the bar to indicate the position to begin playing. Start, fast-forward, rewind and stop operations are available to control play. The buttons shown in the player are context sensitive and change state in response to user's actions. For example the 'play' button on the right hand side of the player becomes a 'stop' button after play is started. The player may be used in isolation, or in combination with the Overview and Transcript once a paragraph or story has been selected.

2.5 The Three Experimental Systems

The goal of this study was to compare the utility of high level visual, textual and tape-recorder components. We therefore modified the modular system in three ways. In one condition, users had access to high-level visual, but no textual information (Fig. 2). In the second, they had only textual information from the errorful Transcript (Fig. 3). In both conditions they also had access to the Player. We compared these two experimental conditions with a control condition in which they only had access to the Player. The control represents the current state of the art for browsing tools, and we wanted to determine whether our two experimental UIs outperformed this baseline. In all conditions users were also able to change stories using the Header Panel. This panel was present in all experimental conditions but is not shown in Figs. 2 or 3. It can be seen in Fig. 1, however.

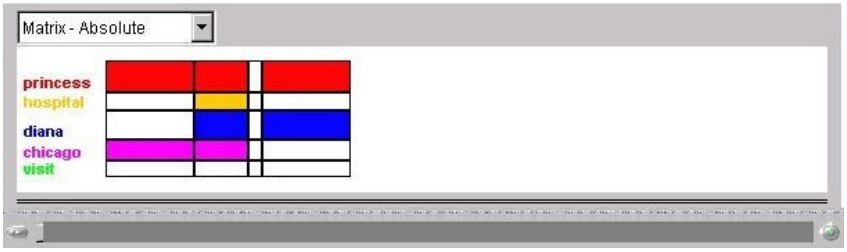

Figure 2: Overview and Player Interface providing high-level visual access

The current experiment was concerned with browsing and not search. We therefore disabled the search engine, giving users pregenerated queries so that they all accessed consistent sets of documents for each experimental task. We also removed all identifying information from the header panel about the stories (Programme Name, Duration, Relevance Score, Hits), as we did not want this external identifying information to influence users' browsing behaviour in the experiment.

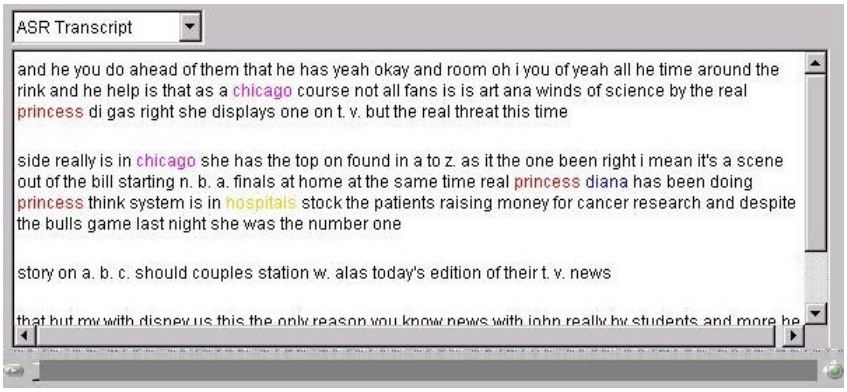

Figure 3: Transcript and Player Interface providing textual access

3. Method

3.1 Tasks

We wanted to compare retrieval situations along several different dimensions, including: making global judgments about sets of "speech documents", and locating specific information from within a "document". We therefore compared the three interfaces on the following two tasks that ethnographic studies [Whittaker et al., 1998b] indicated are representative of real-world user tasks with this type of data:

- Relevance Judgments - compare five news stories to determine which is most relevant to a given topic (e.g. "how good was Valujet's safety record prior to the Florida accident?");
- Fact-Finding - extract detailed factual information from a story to answer a specific question (e.g. "who starred in the Broadway musical 'Maggie Flynn'?");

The study was controlled by a sequence of webpages, one for each task. At the start of each task, we automatically generated the prespecified query and users were asked to carry out a given task associated with the query, using whichever of the three browser types (Overview, Transcript, Player-only) was presented on the page. So for example, users might be presented with the query: "how good was Valujet's safety record prior to the Florida accident?"), along with the Transcript browser, and a short description of the relevance judgment task, i.e. to identify which of five stories shown in the Header Panel was most relevant to the query. For all tasks, users could potentially access five stories, and our prespecified queries meant that these were common across all users for a specific task. We attempted as far as possible to normalise story length across the tasks.

3.2 Procedure and Measures

The experimental design was randomised within subjects. Nineteen users (8 women and 11 men, ranging in age from 18-42) took part. Users were volunteers consisting of researchers, administrative staff and marketers at a large corporate research lab. They had no prior knowledge of the project or experimental hypotheses, but all were experienced users of standard PC software. The entire procedure took about an hour and users were given a small food reward for participating. They were first given a self-paced tutorial about how to use the overall system, followed by three trial tasks - with the Transcript, Overview and Player, to familiarise themselves with the technologies and the procedure. This normally took about 10 minutes.

Once users were ready to proceed with the experiment itself, we gave them a total of 12 tasks each (6 relevance and 6 fact finding). For one third of the tasks they

used the Overview browser, for one third the Transcript browser and one third the Player only. The order of tasks was fixed, but we randomised the sequence in which different browsers were presented. Our prior studies showed that there is enormous variability between different users' performance and strategy in accessing audio data [Whittaker et al., 1998a, 1999, 2002]. Using a within-subjects experimental design allowed us to control for this variance by directly comparing users with their own baseline performance.

For each question we measured outcome information: time to solution and solution quality (as assessed by two independent judges). To determine solution quality, the judges listened to the relevant programmes in their entirety while reading perfect human generated transcripts of the original recordings that we provided them. They then agreed correct answers to all questions, independently scored users' responses (0-100 for each question) before verbally resolving scoring disagreements.

Previous research has demonstrated that users differ in the amount of time and effort they dedicate to optimise performance on these types of experimental tasks [Whittaker et al., 1999, 2002]. If users are prepared to spend a long time listening to broadcasts they have a much greater chance of success, whatever interface they are using. To control for this, we used a compound success measure in which success scores are normalised in terms of the time the user took to complete the task, i.e. score/(unit time). Our predictions below will focus on normalised success. For the sake of completeness, we will report independent time and quality scores, although we make no explicit predictions about these as independent measures.

We also collected information about the processes by which people answered each task. We logged each browser operation along with information about how much speech (if any) was played for that operation.

We also collected subjective data, by administering a short web-based post-test questionnaire after each task. We asked users to judge the difficulty of the task they had just completed (on a 5 point Likert scale). If they had used the Transcript or Overview for that task, then we also asked them whether they would prefer to use the Transcript or Browser if they were to do that task again. Because we were interested in browsing strategies and processes, we also encouraged users to "think aloud" as they carried out the tasks, and we tape-recorded their statements.

3.3 Hypotheses

We made predictions about Outcome, Task Effects, Processes (i.e. playing behaviour) and Subjective Judgments. We first present motivation for our experimental hypotheses, then our results for each hypothesis.

Outcome. Overall we expected users to perform better with the Transcript than the Overview. The Transcript provides richer information than the Overview. Both afford a way to quickly see the distribution of query terms across audio paragraphs, allowing access to relevant areas of underlying speech. However the Transcript

also provides additional detailed information. It shows non-query terms that can be used to aid browsing in two ways. If the Transcript is accurate then information can be read directly from it. If the Transcript is inaccurate, these other terms may still provide a finer grained index to access and play relevant regions of the underlying speech. These led us to predict the superiority of the Transcript compared with the Overview.

We also expected both Overview and Transcript to perform better than the Player. The additional information that each provides about the distribution of query hits should provide greater control over what gets played.

Tasks. We also expected task differences. We expected the Transcript to better support Relevance judgments than the Overview, again because it provides richer information. Although the Transcript may contain errors, it should still enable users to determine the gist of a story without playing it. In contrast, the Overview provides only rudimentary information about a story, necessitating playing parts of the story to judge its Relevance. The other task, Fact-finding, requires detailed and accurate analysis of stories. As neither the Overview nor Transcript provides this, we expected performance to be equivalent. We expected both Overview and Transcript to perform better than Player for both tasks, because of the more detailed information each provides.

Amount of Speech Played and Number of Play Operations. We expected more play operations, and more speech to be played overall in the Overview than the Transcript condition. This is because the Transcript should enable people to extract information by reading rather than playing. Even when the Transcript is inaccurate, the additional lexical information it provides should still allow users to determine the overall gist of the story by reading. We therefore also expect Transcripts to reduce both the number of play operations and the amount played more in the Relevance-judgment than the Fact-finding task We expected the smallest number of play operations in the Player condition, because we anticipated that the lack of structural information in the Player would lead users to simply play messages from beginning to end. This uninterrupted play strategy should mean that users have fewer play operations but play more speech overall with the Player, than the Overview and Transcript.

Subjective Judgments. We expected these to be closely related to outcome and play operations, so that the Transcript should be perceived as making tasks easier than the Overview. Both should be judged to make tasks easier than the Player. Overall users should prefer the Transcript to the Overview when asked to choose between these.

4. Results

We tested the predictions using Analysis of Variance (ANOVA). In all analyses, we used the following independent variables: Browser (Transcript vs. Overview vs.

Player), and Task (Fact-finding vs. Relevance). The respective dependent variables were: normalised success score, amount of speech played, number of play operations, and perceived task difficulty. The results are shown in Tables 1 and 2. Table 1 shows overall data independent of task, while Table 2 shows task interactions. The results for each hypothesis were as follows.

- *Outcome differences*. We expected people to generate better solutions more quickly with the Transcript than the Overview. Both should be better than the Player.

Our predictions were not confirmed (as Table 1 shows). Overall, we found that there were differences between browsers for the normalised success measure ($F(2,234)=13.4$, $p<0.001$). Contrary to our prediction, however, planned comparisons showed that users performed better with the Overview than the Transcript browser ($p<0.05$). As predicted planned comparisons showed that users performed better with the Overview and Transcript than the Player ($p<0.001$ and $p<0.05$ respectively).

Time to solution results were identical ($F(2,234)=14.3$, $p<0.001$), with post-hoc differences between Overview and Transcript and between Transcript and Player. However success scores showed a slightly different pattern: browsers were different ($F(2,234)=6.4$, $p<0.005$), and Overviews scores were better post hoc than Transcript and Player, but there were no post hoc differences between Transcript and Player.

- *Task differences*. We expected that users would perform better with the Transcript than the Overview in the Relevance-judgment task. We expected equivalent performance between browsers for the Fact-finding task. We expected that both would be better than the Player for both tasks.

There was an interaction between task and browser ($F(2,234)=6.5$, $p<0.005$). Again, however, as Table 2 shows, our comparisons between Overview and Transcript contradicted the Task hypothesis. Planned comparisons showed that users did not perform better with the Transcript for Relevance judgment tasks ($p>0.05$). Instead users performed better with the Overview for Fact-finding tasks ($p<0.05$). As predicted, planned comparisons showed that Overview and Transcript performed better than Player for each task (all $p<0.05$), except when the Transcript was used for Fact-finding ($p>0.05$).

	Overview	Transcript	Player	Statistical Significance?
Mean Normalised Performance	0.57	0.36	0.24	O>T>P
Mean Time to Solution (secs.)	227.1	262.3	397.3	O<T<P

Mean Success Score (%)	86.1	76.8	76.6	O>T=P
Mean Amount of Speech Played (secs.)	123.9	37.6	237.8	T<O<P
Mean # of Play Operations	15.3	11.0	7.23	O=T>P
Mean Perceived Task Difficulty (1=very hard, 5=very easy)	3.5	3.1	2.6	O=T>P

Table 1: Effects of Browser on Outcome, Process and Subjective Measures.

- *Amount of Speech Played and Number of Play Operations*. The Transcript should enable users to play less speech and use fewer play operations overall than the Overview browser. We also expected Transcripts to reduce the amount played and number of play operations more than the Overview for the Relevance-judgment task. There should be no differences for the Fact-finding task. Both Overview and Transcript should have more total play operations than the Player, and less overall speech played.

Table 1 shows our predictions were confirmed about the effects of the browsers on the amount of speech people played. People played different amounts of speech with the different browsers ($F(2,234)=42.7$, $p<0.001$). Planned comparisons showed as predicted, that users played less speech when using the Transcript than the Overview ($p<0.001$). Both Transcript and Overview had less playing than the Player (both $p<0.001$).

	Relevance Judgment			Fact finding		
	Overview	Transcript	Player	Overview	Transcript	Player
Mean Normalised Performance	0.29	0.29	0.18	0.84	0.42	0.31
Mean Amount of Speech Played (secs.)	193.4	45.1	295.8	54.4	30.0	179.7
Mean # of Play Operations	21.9	12.1	8.08	8.6	9.8	6.4
Mean Perceived Task Difficulty (5=very easy, 1=very hard)	2.9	2.9	2.33	4.1	3.2	2.8

Table 2: Effects of Task on Performance with Transcript, Overview and Player Browsers

There was an expected interaction between browser and task ($F(2,234)=3.7$, $p<0.05$). As predicted (and as shown in Table 2), users played considerably more speech using the Overview than Transcript for the Relevance-judgment task ($p<0.001$). Planned comparisons between Overview and Transcript showed no differences between the amounts of speech users played on the Fact-finding task ($p>0.05$).

Play Operations. The hypotheses are also partially confirmed for play operations. Table 1 shows there were different numbers of operations between the three browsers ($F(2,234)=9.4$, $p<0.001$). Contrary to predictions, however, planned comparisons show there were no differences between the Overview and Transcript ($p>0.05$). Consistent with our predictions that there were fewer operations using the Player than either Overview or Transcript (both $p<0.01$).

Table 2 shows there was an expected interaction between browser and task for play operations ($F(2,234)=5.31$, $p<0.01$). Planned comparisons between Overview and Transcript showed no differences between the number of play operations in the Fact-finding task ($p>0.05$). As predicted, users had considerably more play operations with the Overview than Transcript for the Relevance-judgment task ($p<0.001$).

- *Subjective judgments*. We expected users to perceive tasks to be easier when using the Transcript compared with the Overview browser. When given a choice between the two browsers, we also expected them to express a preference for the Transcript browser. We expected tasks to be perceived as harder when using the Player, than both Overview and Transcript.

There were overall perceived differences between the browsers in their effects on task difficulty ($F(2,234)=7.5$, $p<0.001$). Contrary to our predictions, however, planned comparisons showed there were no overall differences between the Overview and Transcript browsers ($p>0.05$). However when we asked users which of the two browsers they would choose for the task they had just completed, they chose the Overview 71% of the time. This is significant on a one-sample t test ($t(155)=3.5$, $p<0.001$). As we predicted both Overview and Transcript were judged to make browsing tasks easier than the Player (both $p<0.005$).

User comments: Our objective findings suggest that the Overview was useful for directing play operations for both tasks. User comments are consistent with this: *"I used the Overview to direct my playing to regions where I thought there would be important information."* In contrast, users reported problems with the Transcript in the Fact-finding task, when they tried to access particular facts directly from the Transcript. For example one user tried to scan the Transcript for a specific word: *"I thought that I could find the word 'found' in the Transcript, but it was like trying to find a needle in a haystack, because [the Transcript] was so bad."*

5. Conclusions

What are the implications of this work? First we confirm previous studies showing the utility of ASR generated textual indices for speech browsing when compared with commercial UIs [Whittaker et al., 1999, 2002]. Our data show that Transcripts were more effective than the Player; they reduce the amount of speech users played -- allowing users to identify specific regions of speech for detailed processing, or read information directly from Transcripts. However there are limitations to Transcripts that we discuss below. We also found good evidence for the utility of another interface technique, namely Overviews. Again these reduced the amount of speech that users played, by allowing them to focus on relevant regions.

Contrary to our expectations, however, high-level visual information provided by Overviews proved more useful than Transcripts. This is a significant result because it disconfirms our initial hypothesis, motivated by our belief that Transcripts provided more information than Overviews. Why then did users perform better with Overviews than Transcripts? One significant clue is offered by the fact that users played much more speech with Overviews. It seems that users may have relied too much on the errorful Transcripts, by trying to extract information from them directly by reading. Alternatively they may have spent too much time trying to decipher the meaning of the transcripts in order to better direct their playing.

The notion that transcripts can sometimes be misleading is consistent with the task data. We found that Overviews were better than Transcripts for fact-finding. Indeed Transcripts were no better than the Player for this task. Users played much more speech when using the Overview and Player than the Transcript for this task. This offers strong support for the view that users were sometimes misled into trying to extract detailed information directly - by reading errorful Transcripts. In contrast, people used the Overview to focus on important regions of speech, which they then played to extract information. On the Relevance Judgment task, there were no differences between the browsers. Both were better than the Player for this task. This indicates that Transcripts are reasonably effective for gisting information. Here, users spent much more time playing speech with the Overview than with the Transcript, even though overall performance was equivalent in both tasks. This suggests that the costs that Overview users incurred in having to access more speech were balanced by the effort needed to gist information from poor Transcripts.

Of course, as we noted earlier, our findings about Transcripts are dependent on the quality of the actual Transcript. Perfect transcripts would enable users to reliably answer all questions by reading without having to play any of the underlying story. And other research confirms that transcript quality has a large effect on speech browsing [Stark et al., 2000]. However the data we report is for currently state of the art levels of ASR [Garofolo et al., 2000]. It also seems that ASR performance

levels will not improve drastically in the next few years [NIST, 2003], making our findings important for the design of speech browsing systems in the near future.

We now turn to the design implications of our study. Our findings about the utility of Overviews suggest future research should focus on the development of novel visual representations of speech data. We also partially confirm earlier research in showing benefits for textual indices. However, our findings also suggest that users may spend time trying to directly decipher poor quality transcripts when they might be better employed playing the speech associated with these. What then might we do to preserve the demonstrated benefits of Transcripts, while preventing these maladaptive behaviours? One possibility is that we might use ASR confidence scores to present visual information about the quality of a Transcript, e.g. by greying out regions of the Transcript, where the ASR had low recognition confidence. This would signal to users when they can trust the transcript, and hence read information from it directly, and when they need to switch strategies and play the speech associated with it. Further research might be directed at other automatic methods for determining transcript quality, and hence indicating to users when the transcript can be read directly.

Finally, our unexpected and counterintuitive results, together with other observations of task-specificity, suggest that we need to combine the development of new techniques for speech access and browsing, with careful empirical work evaluating these effects.

6. Acknowledgements

Thanks to the users for donating their time to do the experiment.

References

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

2.Askwall, S. [1985]. Computer supported reading vs. reading text on paper: a comparison of two reading situations, International Journal of Man Machine Studies, 22, 425-439.

3.Bacchiani, M., Hirschberg, J., Rosenberg, A., Whittaker, S., Hindle, D., Isenhour, P., Jones, M., Stark, L., & Zamchick, G. [2001]. SCANMail: Audio Navigation in the Voicemail Domain. In Proceedings of the Workshop on Human Language Technology, 105-111, IEEE Press.

4.Christel, M. Smith, M. Taylor, C., & Winkler, D. [1998]. Evolving Video Skims into Useful Multimedia Abstractions. In Proceedings of Conference on Computer Human Interaction, 171-178, New York, ACM Press.

5. Drucker, S. [2002]. Consumer Level Browsing and Skipping of Digital Video Content. In Proceedings of Conference on Computer Human Interaction, New York, ACM Press.

6.Emnett, K., & Schmandt, C. [2000]. Synthetic News Radio. IBM Systems Journal 2000, 39(3), 646 – 659.

7.Garofolo, J., Lard, J., & Voorhees, E. [2000]. TREC-9 Spoken Document Retrieval Track http://www.nist.gov/speech/sdr2000.

8. Hearst, M. [1995]. Tilebars: Visualization of term distribution in full text information access. In Proceedings of Conference on Computer Human Interaction, New York, ACM Press.

9.Hindus, D., Schmandt, C., & Horner, C. [1993]. Capturing, structuring and representing ubiquitous audio. ACM Transactions on Information Systems, 11, 34-51.

10.Kazman, R., Al-Halimi, R., Hunt, W., & Mantei, M. [1996]. Four paradigms for indexing videoconferences. In IEEE Multimedia, 3(1), 63-73.

11.NIST. [2003]. National Institute of Standards. http://www.nist.gov/speech/

12.Oard, D. [1997]. Speech based information retrieval for digital libraries. In Proceedings of AAAI Spring Symposium On Cross Language Text and Speech, 34-42, Menlo Park, CA, AAAI Press.

13.Oakhill, J., & Garnham, A. [1998]. Becoming a skilled reader. Oxford: Blackwell.

14.Salton, G. [1971]. The SMART Retrieval System, Prentice-Hall, Englewood Cliffs, NJ.

15.Stark, L., Whittaker, S., & Hirschberg, J. [2000]. ASR satisficing: the effects of ASR accuracy on speech retrieval. In Proceedings of International Conference on Spoken Language Processing, 34-38, IEEE Press.

16.Whittaker, S., Hirschberg, J., Choi, J., Hindle, D., Pereira, F., & Singhal, A. [1999]. SCAN: designing and evaluating user interfaces to support retrieval from speech archives. In Proceedings of SIGIR99 Conference on Information Retrieval, 26-33, New York: ACM Press.

17.Whittaker, S., Hirschberg, J. & Nakatani, C. [1998a] Play it again: a study of the factors underlying speech browsing behaviour. In Proceedings of Conference on Computer Human Interaction, 247-248, New York, ACM Press.

18.Whittaker, S., Hirschberg, J. & Nakatani, C. [1998b]. All talk and all action: strategies for managing voicemail messages. In Proceedings of Conference on Computer Human Interaction, 249-250, New York, ACM Press.

19.Whittaker, S., Hirschberg, J., Amento, B., Stark, L., Bacchiani, M., Isenhour, P., Stead, L., Zamchick G., & Rosenberg, A. [2002] SCANMail: a voicemail interface that makes speech browsable, readable and searchable. In CHI2002 Conference on Human Computer Interaction, 275-282, NY: ACM Press.

20.Wilcox, L. Chen, F., Kimber D. & Balasubramanian, V. [1994]. Segmentation of Speech Using Speaker Identification. Proceedings of International Conference on Computer Speech and Signal Processing, 234-238, IEEE Press.