Searching Large Collections of Recorded Speech: A Preliminary Study

Jinmook Kim
College of Information Studies, University of Maryland, College Park, MD 20742-4345
Email: jinmook@glue.umd.edu

Douglas W. Oard
College of Information Studies and Institute for Advanced Computer Studies, University of Maryland
Email: oard@glue.umd.edu

Dagobert Soergel
College of Information Studies, University of Maryland
Email: ds52@umail.umd.edu

This paper reports on an exploratory study of the criteria searchers use when judging the relevance of recorded speech from radio programs and the attributes of a recording on which those judgments are based. Five volunteers each performed three searches using two systems (NPR Online and SpeechBot) for three questions and judged the relevance of the results. Data were collected through observation and screen capture, think aloud, and interviews; coded; and analyzed by looking for patterns. Criteria used as a basis for selection were found to be similar to those observed in relevance studies with printed materials, but the attributes used as a basis for assessing those criteria were found to exhibit modality-specific characteristics. For example, audio replay was often found to be necessary when assessing *story genre* (e.g., report, interview, commentary) because of limitations in presently available metadata. Participants reported a strong preference for manually prepared summaries over passages extracted from automatic speech recognition transcripts, and consequential differences in search behavior were observed between the two conditions. Some important implications for interface and component design are drawn, such as the utility of summaries at multiple levels of detail in view of the difficulty of skimming imperfect transcripts and the potential utility of automatic speaker identification to support authority judgments in systems.

Introduction

Far more is spoken each day than is written, and technology to acquire, store, and replay spoken content is now ubiquitous. Searching spoken word content is therefore a challenge of significant importance. We know quite a lot about how people interact with text retrieval systems to specify their needs and select relevant documents, but we do not yet understand well how those behaviors carry over to searching collections of speech recordings. Our goal in this study is to explore the selection behavior of users of interactive speech retrieval systems in order to provide insight into interface and system design issues.

Much of the research on speech retrieval to date has focused on development of algorithms for automating the search process based on emerging speech technology such as automatic transcription (speech recognition) and topic segmentation (Allan, 2002; Voorhees & Harman, 2000). As this technology has matured, end-to-end systems that support interactive searching have started to appear. Research systems such as Informedia (Christel et al., 1995), the Audio Notebook (Stifelman et al., 2001), and SCAN (Whittaker et al., 2002) have explored access issues to spoken word collections of television news, lectures, meetings, interviews, and voicemail. Commercial systems (e.g., Virage and Fast-Talk) are now starting to appear. The emergence of complete and scalable systems has in turn made it possible for researchers to augment earlier component-oriented user studies (e.g., [Arons, 1997]) with studies of situated users performing typical tasks (e.g., [Whittaker et al., 1998]).

Systems designed to support interactive searching typically offer two interaction opportunities: (1) formulation (and reformulation) of an information need statement, and (2) selection of the most useful information from among a set of promising candidates that are identified by the system. Observational studies have led to

well developed theoretical frameworks for each task (e.g., [Taylor, 1962] for query formulation and [Saracevic, 1976] for document selection). Consequently, a great deal is now known about application of such frameworks to study the process by which retrieval systems are used to search large collections of written text (e.g., [Wang & Soergel, 1998]). The study reported here is the first that we are aware of to apply a similar methodology to study the use of systems that search the spoken word.

For our initial study, we have chosen to focus on the selection task. Our immediate goal is to understand how existing speech retrieval systems support the cognitive processes involved in relevance judgment in order to inform the design of future interactive systems for searching recorded speech. The next section describes the conceptual framework that guided our study. We then present our research questions, describe the application environment in which we explored those questions, and introduce our study design. Findings and a brief discussion of the limitations and implications of the study follow, along with some ideas for future work that have been inspired by this study.

Conceptual Framework

The concept of relevance is widely used as a basis for evaluating the effectiveness of information retrieval systems (Saracevic, 1976; Schamber, 1994). However, terminology for relevance studies is far from standardized. Some researchers use *topicality*, *utility*, *pertinence*, *satisfaction*, or *situational relevance* to refer to similar concepts (Wilson, 1973; Schamber, 1994). In this study, we define relevance as the degree to which a speech recording (or a part of a recording) meets a searcher' needs. This broad statement is meant to be inclusive, subsuming *topical relevance*, situational factors (e.g., previously seen documents), and other factors related to the nature of the document (e.g., *authority*, the concept that underlies Google's PageRank score).

Table 1. Examples of bases for selecting journal articles

Relevance Criteria	Associated Attributes
Topicality	Title, abstract, descriptors
Novelty	Title and author
Quality	Journal, author, citation status
Availability	Journal and type, owning library
Accessibility	Language, media
Recency	Publication date
Authority	Author and author's affiliation
Reading time	Number of pages
Relation/Origin	Author

The cognitive processes underlying judgments of relevance have been widely studied, often with the goal of identifying criteria that influence relevance judgments (Park, 1993; Barry, 1994; Wang & Soergel, 1998; Tang & Solomon, 2001). Table 1 summarizes the most widely observed criteria from those studies, all of which focused on selection of journal articles using bibliographic databases. Relevance criteria are, however, abstract concepts, and searchers must ground their interpretation of each criterion in some set of observable attributes (examples of which are also shown in Table 1). For instance, a searcher might assess the topicality of a journal article based on the title and abstract of the article and any thesaurus descriptors that have been assigned to the article.

Table 2. Possible bases for selecting radio programs

Relevance Criteria	Possibly Associated Attributes*
Topicality	Program title, story title, brief story summary, detailed story summary, short extract from automatic transcript, longer extract from automatic transcript, highlighted terms in transcript, speaker name, position in ranked list, query similarity score, full story audio, full program audio, audio from user-selected passage start time
Novelty	Program title, story title, brief story summary, detailed story summary, short extract from automatic transcript, longer extract from automatic transcript, audio from user-selected passage start time, full story audio, full program audio
Authority	Speaker name, speaker's affiliation, program title
Recency	Date
Listening time	Story length

* Attributes available in at least one system in our study.

We expect the set of criteria and their mapping to associated attributes to be different for searching recorded speech, both because of different characteristics speech has and because of differences in the set of available attributes. For example, skimming, which involves rapid browsing, looking ahead, and looking back, is much more difficult for speech than for text. Units of retrieval, such as fixed-length chunks or meaningful segments isolated through automatic segmentation, may lack a meaningful title. Genre is also an important factor. For example, in recorded classroom lectures, we might expect associated

visual materials (e.g., PowerPoint slides) to be useful attributes when assessing *topical relevance*. Table 2 shows the relevance criteria and associated attributes that we envisioned for radio programs at the outset of our study (Kim & Oard, 2002).

Study Design

Research Questions

The goal of our study was to characterize the relevance criteria searchers apply when searching a collection of recorded radio programs and the observable attributes of those recordings on which those assessment were based. We formalized that interest with two research questions that guided our data collection and analysis:

a. What relevance criteria do searchers apply when choosing a recording or a passage of a recording?

b. What attributes of the recordings do searchers use as a basis for assessing each criterion?

Search Systems

We used National Public Radio's "NPR Online" (available at http://npr.org/archives) and Compaq's "SpeechBot" (available at http://speechbot.com) (Thong et al., 2002). NPR Online supported searching based on human-prepared metadata (e.g., titles, written summaries, program name, date, etc.). SpeechBot, by contrast, relied entirely on automatic processing, using automatic speech recognition as a basis for automatic indexing.

NPR Online offers access to many National Public Radio programs, and SpeechBot indexes preselected audio and video programs from several sites, including some programs from NPR Online. On the broadest level, the broadcast material can be divided into *programs*, such as *American RadioWorks*, *Car Talk*, *The Diane Rehm Show*, *Fresh Air*, *The Connection*, and *Public Interest*. Each program consists of *episodes*, the airing of the program on a given day. An episode, then, may have *stories*, defined broadly as a topically coherent segment (such as the answer to a specific car repair question in *CarTalk*). An episode can also be divided into arbitrary chunks of a given length. In NPR Online, the retrieved units are stories whose boundaries have been determined manually. In SpeechBot, they are chunks of 200 words (called extracts) from an ASR transcript. Both systems accept a text query and return a list of search results in order of decreasing degree of topical match. Searchers can replay part or all of any story or extract that appears in the search results (in each case, using RealPlayer).

The systems differ principally in the basis for search and the way in which search results are displayed. NPR Online users can specify search terms that will be compared with terms in the human-prepared titles and summaries, and the search can optionally be limited to stories from a single program. SpeechBot users can specify search terms that will be compared with terms in automatically generated (imperfect) transcripts, and they can optionally limit their search to either episodes of a single program or to all programs within a genre (e.g., news programs). With either system, searchers can choose to limit their search to a recent period (e.g., the past day, week, month, or year).

Both systems display a ranked list of search results that includes brief summary information for each item. NPR Online treats each retrieved story as an independent unit. SpeechBot groups retrieved extracts by episode and presents an ASR transcript of the extract underneath a timeline of the entire episode with tick marks identifying points in the episode where query terms were found. The tick mark for the retrieved extract is highlighted so that searchers can easily select other possibly relevant points to see an extract and/or to begin audio replay at an appropriate point. Table 3 summarizes the types of data and metadata available from NPR Online and SpeechBot.

Table 3. Available attributes

NPR Online	SpeechBot
Metadata	**Metadata**
Story title	(No title for extracts)
Program title	Program title
Episode date	Episode date
Brief summary	Short extract from transcript
Detailed summary (only for a few stories)	Longer extracts from transcript
	Highlighted terms in transcript
	Keyword-cued episode timeline
Speaker name	
Speaker affiliation	
Story length	
Query similarity score	
Audio Replay	**Audio Replay**
Full story audio	Audio from passage start time
Full episode audio	Full episode audio

Participants

We adopted a case study methodology, which is well suited for exploratory research in which the goal is hypothesis generation rather than hypothesis testing (Creswell, 1994; Maxwell, 1996). An optimal case study design would be based on observations of experienced users performing a real task using a speech retrieval system. However, the population of such users is presently so limited (and in this case, so highly tasked) that we were unsuccessful in obtaining volunteers from within the news organization that we were working with at the outset of

this study. Five volunteers from a graduate-level seminar on visual and sound materials for library and archives students were therefore recruited to participate in our study. Students in that course studied acquisition, preservation, access, and management issues. All five had completed a prerequisite course on information access techniques, but none had any experience with audio searching. Participants P2 and P4 reported that they were frequent NPR listeners, while participants P1, P3, and P5 reported rarely listening to NPR.

Procedure

Each participant was asked to perform three searches: two based on topics that we provided, and one developed independently based on their own interests.[1] Participants were encouraged to try both systems on their own prior to participation in the study. At the outset of each session, we provided 20 minutes of training on how to search stories using each system, on how to browse search results, and on the think-aloud procedures that we would ask them to use. After the training, participants performed their searches, three starting with NPR Online (P2, P4 and P5) and two with SpeechBot (P1 and P3). Participants used the assigned system for both of the topics that we provided, and were encouraged to (and did) use both systems for the topic that they chose independently.

Participants were encouraged to reformulate their queries as often as necessary, but we limited the first two searches to about 15 minutes each in order to minimize potential fatigue effects. We placed no limit on the duration of the third search. On average, participants spent a total of about one hour actually searching during the study. We used observation and think aloud during each search, and we conducted semi-structured interviews immediately following each three-search session. Obtaining multiple perspectives in this way helped to establish greater confidence in inferences that were supported by more than one source of evidence, a process known as triangulation.

Observation Protocol

Each participant was scheduled for an individual session in order to permit close observation. The observer (the first author of this paper) noted the searcher's choice of query terms and the actions they took while examining the results of each search iteration. These observations were guided by our research questions, so we focused on understanding how the searcher appeared to be selecting recordings. Any unexpected behavior was also noted and used to guide clarification questions during the semi-structured interview. The observer minimized interruptions during a search, and searchers were asked not to consult the observer as a source of expert advice during the session. A low-resolution (NTSC) videotape of on-screen activity was also made using a video capture card.

Think-aloud Protocol

We sought to augment the observational notes by asking our participants to contemporaneously describe the reasons for their actions. Concurrent verbal reporting can be problematic, since it may alter the very behavior that we wish to study. On the other hand, think-aloud can yield insights that would be difficult to obtain by other means, particularly when prior knowledge is an important basis for an activity (Wilson, 1994). On balance we felt that the advantages outweighed the disadvantages in this case. We told each participant of our interest in the way they determine which recordings meet their needs, but we did not provide specific guidelines on what to talk about or how to express their thoughts. The think-aloud was audiotaped and subsequently transcribed.

Interview Protocol

A 30-minute semi-structured interview was conducted by the first author immediately following each participant's last search. Figure 1 shows the topics that we explored and suggested questions (Q) that illustrate the type of questions that we would use to initiate the discussion of each topic. As we learned from early participants, we refined the questions that we posed in interviews with subsequent participants. The interviews were audiotaped and subsequently transcribed.

---

a. What relevance criteria did searchers apply to select or discard a recording or a passage? (Q: Why did you select or discard [some specific] recording?)

b. How did searchers combine the criteria to reach a decision? (Q: Were some reasons more important than others?)

c. What attributes did searchers use as a basis for assessing each relevance criterion? (Q: How did you usually determine [some relevance criterion that was mentioned]?)

d. What actions did searchers take during a search when making relevance judgments and why? (Q: Why did you do [some specific behavior]?)

---

[1] The topics we specified were:
(1) Find stories about organic food standards; and
(2) Find stories that discuss alternative approaches to the problem of combating terrorism.
The study was conducted about one month after the September 11, 2001 terrorist attacks. The topics chosen by the participants were:
(P1) Interviews with the new poet laureate, Billy Collins;
(P2) International reaction to Bush's recent speech to Congress;
(P3) Music in Moulin Rouge;
(P4) Drug testing in sports; and
(P5) Voucher funding for private schools.

e. What attributes were not present in either system that would have been desirable? (Q: Were there cases you had to actually listen to a recording to determine whether it was useful?)

f. What capabilities were not present in either system that would have been desirable? (Q: Were there any features that you had expected to see that were not present in either system?)

Figure 1: Semi-structured interview topics

Data Analysis

We used the QSR NVivo data analysis system[2] to organize and analyze the observational notes, think-aloud transcripts, and semi-structured interview transcripts. We developed an initial coding scheme based on the conceptual framework described above, adding new categories as our analysis revealed additional relevance criteria and associated attributes. Upon completion of the coding, we examined the resulting cross-referenced data to identify patterns and trends in the application of relevance criteria and attributes.

We sought to maximize the validity of our analysis by submitting a draft of our findings to two study participants (selected based on availability) for "member checks," since study participants (the "members") are in an excellent position to assess whether we interpreted their actions and statements correctly. The comments that we received proved to be beneficial, improving our understanding on several points.

Research Findings

Qualitative research is best thought of as inductive rather than deductive, building concepts and theories from details of phenomena (Creswell, 1994). We therefore present our findings principally in narrative form.

Table 4. Mentions of relevance criteria by users

NPR Online (176)	**SpeechBot** (86)
Topicality (119; 68%)	Topicality (58; 67%)
Story genre (22; 13%)	Time frame (12; 14%)
Time frame (18; 10%)	Story genre (9; 10%)
Recency (8; 4.5%)	Recency (3; 3.5%)
Listening time (6; 3.5%)	Listening time (2; 2.3%)
Authority (3; 2%)	Authority (2; 2.3%)

Our first research question addressed the relevance criteria searchers apply when selecting recorded radio programs. Table 4 lists the criteria searchers were observed to use, listed in order of decreasing number of mentions in think-aloud and semi-structured interviews (mention counts and percentages are shown in parentheses). Table 5 provides a more detailed view of the same data in which criteria are paired with associated attributes (in that case, with mention counts on each association).

Topicality

Topicality was mentioned most frequently by far for both systems (119 of 176 mentions for NPR Online and 58 of 86 mentions for SpeechBot) by all five participants, which comports with previous findings for document retrieval applications. Searchers using NPR Online were able to judge the topical relevance of a recording by examining the *story title* (79 of 195 mentions; 41%), *brief summary* (51 of 195 mentions; 26%), *audio replay* (25 of 195 mentions; 13%), *detailed summary* (23 of 195 mentions; 12%), *speaker name(s)* (14 of 195 mentions; 7%), and/or *speaker's affiliation* (3 of 195 mentions; 2%). The corresponding attributes for SpeechBot were *longer extract from transcript* (26 of 69 mentions; 38%), *short extract from transcript* (22 of 69 mentions; 32%), *audio replay* (13 of 69 mentions; 19%), and *highlighted terms in transcript* (8 of 69 mentions; 12%).

All five participants stated that the availability of titles and brief summaries in NPR Online helped them quickly select potentially relevant stories for further examination. For example, participant P2 typically first glanced at all the titles in a retrieved set and then selected specific items for further examination. This resembles the common practice in text retrieval systems, where a title and some form of summary are typically provided. NPR Online was preferred over SpeechBot by all five participants; the similarity to systems with which they were already familiar may have been an important factor in establishing that preference.

Neither story titles nor human-prepared summaries were available in SpeechBot, but a short extract from the (imperfect) transcript served a similar purpose. However, we observed that searchers often found it difficult to assess topical relevance based solely on a short (20 word) extract from the speech recognition transcript in the absence of other contextual cues. Users could obtain longer (50-200 word) extracts from the same transcript, and our participants used that feature more often than they relied on short summaries. This pattern was evident with all five participants, and particularly pronounced for the two who used SpeechBot for their initial searches. For example, P3 stated on one occasion that:

> *"... Um, I am going to look at more because I am not getting enough from that little initial transcript..."*

[2] http://www.qsr.com.au/products/nvivo_details.html

Table 5. Associated Attributes: NPR Online and SpeechBot

Relevance Criteria	Relevance Attribute	
	NPR Online	**SpeechBot**
Topicality	Story title (79) Brief summary (51) Audio replay (25) Detailed summary (23) Speaker name (14) Speaker affiliation (3)	Longer extract from transcript (26) Short extract from transcript (22) Audio replay (13) Highlighted terms in transcript (8)
Story genre	Detailed summary (9) Brief summary (7) Audio replay (6) Story title (4)	Audio replay (9)
Time frame	Broadcast date (16) Brief summary (5) Audio replay (4)	Broadcast date (12)
Recency	Broadcast date (8)	Broadcast date (3)
Listening time	Story length (6)	Program title (2)
Authority	Speaker name (3) Speaker affiliation (1)	Program title (2)

Although NPR Online now provides detailed summaries for many stories, that feature was available only for a small percentage of the stories at the time we conducted our study. We observed that users of NPR Online relied heavily on these detailed summaries when they were available. Similar behavior is not typically seen in full text retrieval systems because the structure and layout of written text make it relative easy to skim. When considered in conjunction with the proclivity of viewing longer extracts, we believe it provides convincing evidence regarding the importance of written representations with more detail than the very compact summaries that are typically displayed as part of a ranked list.

Four of our five participants often listened to the audio (with both systems). The most common reason given for listening to audio was to confirm a tentative relevance judgment, although sometimes (particularly during their third search) people listened to a story just out of personal interest. Those four participants sometimes changed their relevance judgments after listening to the audio, which seemed to subsequently increase their proclivity to listen to audio even when relatively sure of their initial judgment.

The remaining participant (P5, who used NPR Online initially) rarely listened to any audio with either system, noting (during the interview), that they felt they were able to make an accurate decision in most cases without listening. The four other users found that with SpeechBot there were some cases in which even the longer extract failed to provide enough information to make their decision, as illustrated by this quote from P1:

"... In some case (sic.) it was because I was interested in listening to it. But in most cases it was because there wasn't enough information..."

In some cases, this seemed to be caused by disfluencies and inaccuracies in the automatically generated transcript, and that confounding factor precluded a detailed analysis of other possible limitations of transcript-based access to recorded speech. All five participants did, however, mention other factors (e.g., word order, duplication, inconsistent word usage, and the lack of punctuation). These limitations of present speech recognition systems help to explain why all five participants expressed a preference for NPR Online; as P4 stated:

"… I liked NPR a lot better… On SpeechBot, a lot of the text wasn't actually accurate or didn't make

sense…like you couldn't read it in a sentence… I could see the words but the context around the words didn't make any sense…"

These remarks offer particular insight when viewed in contrast to the absence of even a single mention of the effect of speech recognition errors on the quality of the retrieved set. Evaluations of the automated component of speech retrieval systems that build ranked lists for display to the user have repeatedly shown relatively small effects from words missed through recognition errors. But, it appears from our observations that this type of error can have a serious effect on the usability of complete interactive speech retrieval systems.

With NPR Online, four of the five participants (P1, P2, P4 and P5) sometimes found that the speaker's name and/or affiliation helped them to determine *topicality*. This ability was clearly closely coupled with individual factors such as prior exposure to a speaker or their organization. However, both P1 and P2 remarked they were more interested in finding out what the speaker was going to talk about than who the speaker was. P2 noted that this could often be determined by listening to the first instance in which each speaker spoke during a story. This suggests that displaying speaker turns, a capability not provided by either of the systems we used, might be useful in some cases. Finally, four participants (P1, P3, P4, and P5) mentioned that highlighting search terms in transcripts (a capability provided by SpeechBot) is helpful. In addition to helping to focus the eye on salient parts of a transcript, highlighting search terms might have the additional benefit of making the operation of the search system more transparent, which perhaps might improve the ability of searchers to formulate (or reformulate) more effective queries.

Story Genre

The *story genre* (22 of 176 mentions for NPR Online and 9 of 86 mentions for SpeechBot) of a story was an important criterion in some cases. Examples of story genre that we observed were interview, special report, commentary, debate, announcement, and call-in program. Participants P1 and P2 chose to focus on interviews and reactions, respectively, for their third search. As a result, they mentioned *story genre* as a criterion more often than the others. It therefore seems that the importance of *story genre* depends somewhat on the topic. The attributes associated with *story genre* in NPR Online were *detailed summary* (9 of 26 mentions; 35%), *brief summary* (7 of 26 mentions; 27%), *audio replay* (6 of 26 mentions; 23%), and *story title* (4 of 26 mentions; 15%). For example, participant P1 said:

"…I'm not sure that this is necessarily an interview so I might want to listen to this…looks like it might just more like a general announcement that he was Poet Laureate."

With SpeechBot, *audio replay* (9 mentions) was the only attribute mentioned as informing the judgment of *type*. Since playing several passages can be time consuming even if each passage is relatively short, it appears that some form of automatic *story genre* classification technique would be useful (e.g., classification based on turn-taking behavior, as suggested by [Oard, 2000]).

Time Frame

Time frame (18 of 176 mentions for NPR Online and 12 of 86 mentions for SpeechBot), which refers to a span of dates associated with an event, was mentioned by three participants (P2 and P4, who started with NPR Online, and P3, who started with SpeechBot). When searching for something associated with a specific event, searchers expressed a desire to limit their search to a particular date or a range of dates (once the date of the event was known). For example, participant P2 learned from an early search that the first organic food standard was set in December, 2000, and subsequently looked only for stories aired after that date. Both systems allowed searchers to specify a period extending backwards from the present date (e.g., "search the past 7 days"), but neither allowed specific dates or spans of dates to be specified. Participant P2 remarked:

"… I would appreciate it if I were able to retrieve all stories aired on this date on the topic I am searching for…"

As with *story genre*, the importance of *time frame* seems to depend on the topic. The difference in the relative predominance of those criteria between NPR Online and SpeechBot thus probably says more about topic selection than it does about the relative importance of those criteria. In addition to the obvious attribute *broadcast date* (16 of 25 mentions; 64%), our participants in NPR Online discerned evidence about time frame from *brief summary* (5 of 25 mentions; 20%) and *audio replay* (4 of 25 mentions; 16%). With SpeechBot, *broadcast date* (12 mentions) was the only attribute mentioned by our participants.

Recency

Recency (8 of 176 mentions for NPR Online and 3 of 86 mentions for SpeechBot) is a special case of a *time frame* in which the *broadcast date* is sufficiently recent to suggest that the information provided has not been superseded. *Recency* was mentioned less often than *topicality*, *type*, and *time frame*, but four participants did mention it and one (P2, who chose a current news topic for their third search) mentioned it six times.

Listening Time

Our participants generally seemed willing to listen to the audio in cases where it was clear that what they were looking for could be found there. In cases where the relevance of a story was less clear, three participants (P2, P3, and P4) mentioned *listening time* (6 of 176 mentions for NPR Online and 2 of 86 mentions for SpeechBot) as a factor in their decision. *Listening time* was associated with the *story length* attribute (6 mentions) in NPR Online. The comparable attribute in SpeechBot would have been the timeline interval over which query terms were found, but no participant mentioned that attribute. Two participants did, however, mistakenly treat program length (which was provided by SpeechBot) as if it were story length. *Program title* (2 mentions) was also used by one participant (P4, a frequent NPR listener) to infer listening time, stating:

> *"... It's on Morning Edition, so I know the pieces are going to be shorter. So, I may listen... It's a three-minute segment... This is on Talk of the Nation, so that's going to be an hour..."*

Authority

Finally, three participants (including P2 and P4, who were frequent NPR listeners) sometimes based their selections on the *authority* (3 of 176 mentions for NPR Online and 2 of 86 mentions for SpeechBot) of a source, as expressed in the following quote from P2:

> *"… The commentator is Robert Siegel, and I usually like his segments, which makes me more likely to stick with the whole clip because I think he's very thoughtful…I think he couches his questions well…"*

Speaker name (3 of 4 mentions) and *speaker affiliation* (1 of 4 mentions) were used in judging the authority of a story with NPR Online, but neither of those attributes was available in SpeechBot. One participant (P4) also used *program title* (2 mentions) as the basis for determining authority, saying:

> *"... Gosh this is just kind of confusing. Sightings on the Radio with Jeff Rense... I don't think that's going to be a good one... I'll try Public Interest just to figure out what's going on here..."*

Limitations

It might be tempting to attempt a more detailed analysis based on the figures in Tables 4 and 5, but it is important to bear in mind the limitations of the case study method that we used for this study. Our goals in this study were exploratory rather than comparative, and we found that interest to be well served by rich data collection and extensive analysis, factors that necessarily limited the number of searchers and search sessions that we could accommodate. Reliable quantitative analysis would require many more searchers and searches. However, as we have demonstrated, considerable insight can be gained from a limited number of search sessions using qualitative analysis methods.

A second important limitation of our design is that our inferences were drawn from a relatively homogenous user population. Selection of appropriate study participants is particularly challenging with emerging technologies, since any of the possible searcher populations (e.g., early adopters, professional searchers experienced with a single system, etc.) would likely exhibit some characteristics that would not be representative of the others. Our choice to study trained searchers who had only recently been exposed to search technology for spoken word collections in a laboratory setting turned out to be both feasible and insightful, but the field would clearly benefit from additional studies with other searcher populations.

Finally, our choice of domain and systems was opportunistic; much of the early work on automating access to spoken word collections has focused on broadcast materials because those materials are both technically tractable (people tend to speak clearly) and widely available. Technique effects (manual vs. automatic) and system effects (NPR vs. SpeechBot) are unavoidably confounded in our observations, and this limitation will persist until a broader range of systems become available. In the mean time, self-report data (from think aloud and semi-structured interviews) can provide at least some insight into the thought process that led to observed actions.

Implications

This study has illuminated several issues that we believe have consequences for the design of future interactive speech retrieval systems, including:

- In SpeechBot, recognition errors appeared to have a far greater effect on human performance than on automatic components of retrieval systems. If accuracy cannot be improved, summarization techniques designed to enhance selection performance (e.g., the phrase extraction ideas of [Merlino & Maybury, 1999]) may help.
- Detailed text summaries provided important information that searchers seemed to rely on when judging *topicality*. This is a marked contrast with the comparable condition in retrieval of written text, where the possibility of rapid browsing makes full-text display more common.
- The *genre* of story (interview, report, etc.) was important to searchers in some cases. This suggests that automatic determination of story genre might be useful.

- The *broadcast date* of a story was an important factor when the date of an event was known. This suggests that it might be valuable to highlight date information in story summaries. Moreover, our findings support a recommendation that some form of easily set two-sided date range (e.g., begin and end sliders on a timeline) be incorporated in the query interface.
- *Story length* was considered by searchers when deciding whether to listen to a story whose relevance could not be determined in any other way. When story length is not available as metadata, this information might be automatically determined using topic boundary detection based on vocabulary shift (Allan, 2002).
- The identity of well-known speakers provided a useful basis for assessing *authority* (and, in some cases, *topicality*). This suggests that automatic speaker identification would be helpful, at least for the most frequent speakers in a collection.

These observations have clear implications for the development of component technologies, the design of systems to search spoken word collections, and the need for further user studies. Perhaps the most urgent need is for additional studies to examine the behavior of other searcher populations, searching different types of materials, using search systems based on a broader range of techniques. The study design presented in this paper should be a useful point of departure, perhaps with the addition of a focus on iterative query formulation and refinement.

Among component technologies, it seems that the most urgent needs are for improvements in the readability (or, more specifically, the "skimability" of transcripts produced by automatic speech recognition, and for development of automatic summarization techniques that are able to effectively exploit the characteristics of presently available speech recognition systems. Our finding that searchers rely heavily on detailed summaries when they are available suggests that system developers could make good use of such technology.

Conclusion

Searching the spoken word imposes demands on both system and searcher that differ from those involved in searching the written word. The selection criteria that we identified, *topicality*, *story genre*, *time frame*, *recency*, *listening time*, and *authority*, have clear counterparts in the retrieval of written text, but the attributes used to assess those criteria differ in important ways. Moreover, the linear nature of audio places a high premium on the quality of the available written descriptions of the spoken content. Some other emerging technologies (e.g., speaker identification and story type detection) also appear to be promising, at least for the broadcast domain that we studied in this paper. Much more is spoken each day than is written, so systems that help access that vast trove of information will likely assume increasing importance as the search technology improves. Through studies such as the one reported in this paper, we can help to ensure that the investments we make in the development of systems and component technologies will be well aligned with the needs of real searchers.

ACKNOWLEDGMENTS

The authors are grateful to Bill Byrne, Allison Druin, and Anton Leuski for their thoughtful comments on earlier versions of this paper, to Tom Connors for inviting us to work with his class, and to the participants in our study. This material is based on work supported by the National Science Foundation under grant IIS-0122466. Any opinions, findings, conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF.

REFERENCES

Allan, James (ed). (2002). Topic Detection and Tracking: Event-Based Information Organization. Klewer Academic, Boston.

Arons, B. (1997). SpeechSkimmer: a system for interactively skimming recorded speech. ACM Transactions on Computer-Human Interaction, 4 (1), 12-19.

Barry, C.L. (1994). User-defined relevance criteria: an exploratory study. Journal of the American Society for Information Science, 45 (3), 149-159.

Christel, M., Wactlar, H., Steven, S., Sirbu, M., Reddy, R., Mauldin, M., and Kanade, T. (1995). Informedia digital video library. Communications of the ACM, 38 (4), 57-58.

Creswell, J.W. (1994). Research design: Qualitative and quantitative approaches. Thousand Oaks, CA: Sage Publications Inc.

Kim, J. and Oard, D.W. (2002). The use of speech retrieval systems: a study design. Coden, A.R., Brown, E.W., and Srinivasan, S. (Eds.), Information Retrieval Techniques for Speech Applications (pp. 86-93). LNCS 2273, Springer-Verlag.

Maxwell, J.A. (1996). Qualitative research design: an interactive approach. Thousand Oaks, CA, Sage Publications Inc.

Merlino, A., and Maybury, M. (1999). An empirical study of the optimal presentation of multimedia summaries of broadcast news. Mani, I., and Maybury, M. (Eds.), Automated Text Summarization (pp. 391-402). MIT Press.

Oard, D.W. (2000). User interface design for speech-based retrieval. Bulletin of the American Society for Information Science, 26 (5), 20-22.

Park, T.K. (1993). The nature of relevance in information retrieval: an empirical study. Library Quarterly, 63 (3), 318-351.

Saracevic, T. (1976). Relevance: a review of the literature and a framework for thinking on the notion in information science. Voigt, M.J., and Harris, M.H. (Eds.), Advances in Librarianship 6 (pp. 81-139).

Schamber, L. (1994). Relevance and information behavior. Annual Review of Information Science and Technology, 29, 1-48.

Stifelman, L., Arons, B., and Schmandt, C. (2001). The Audio Notebook: paper and pen interaction with structured speech. Proceedings of CHI 01 (pp. 182-189). Seattle, WA, ACM Press.

Tang, R., and Solomon, P. (2001). Use of relevance criteria across stages of document evaluation: on the complementarity of experimental and naturalistic studies. Journal of the American Society for Information Science & Technology, 52 (8), 666-685.

Taylor, R. (1962). The process of asking questions. American Documentation, 391-396.

Thong, JM.V., Moreno, P., Logan, B., Fidler, B., Maffey, K., and Moores, M. (2002). SpeechBot: an experimental speech-based search engine for multimedia content in the Web. IEEE Transactions Multimedia Journal, 4 (1).

Voorhees, E., and Harman, D. (2000). Overview of the Ninth Text Retrieval Conference. The Ninth Text REtrieval Conference (TREC-9) (pp. 1-14). Gaithersburg, MD, NIST.

Wang, P. and Soergel, D. (1998). A cognitive model of document use during a research project, Study I: document selection. Journal of the American Society for Information Science, 49 (2), 115-133.

Whittaker, S., Hirschberg, J., Amento, B., Stark, L., Bacchiani, M., Isenhour, Pl, Stead, L., Zamchick, G., and Rosenberg, A. (2002). SCANMail: a voicemail interface that makes speech browsable, readable and searchable. Proceedings of CHI `02 (pp. 275-282). Minneapolis, MN, ACM Press.

Whittaker, S., Hirschberg, J., and Nakatani C.H. (1998). Play it again: a study of the factors underlying speech browsing behavior. Proceedings of ACM CHI ’98. Los Angeles, CA, ACM Press.

Wilson, P. (1973). Situational relevance. Information Storage and Retrieval, 9, 457-471.

Wilson, T.D. (1994). The proper protocol: Validity and completeness of verbal report. American Psychological Society, 5 (5), 249-251.