The Use of Speech Retrieval Systems: A Study Design

Jinmook Kim and Douglas W. Oard
College of Information Studies and
Institute for Advanced Computer Studies
University of Maryland
College Park, MD 20742-4345
1-301-405-2033
{jinmook, oard}@glue.umd.edu

ABSTRACT
What relevance criteria do users apply when selecting a speech recording? What attributes of the recording do they rely on for each criterion? This paper proposes a qualitative research study design to explore those questions. A conceptual framework is presented, research questions are introduced, and the study design is described. The paper concludes with some observations on how the results of the study might inform the design of future systems.

1. INTRODUCTION
As the networking and storage infrastructure of the Internet becomes more robust, the potential for physical access to speech recordings is increasing dramatically. Intellectual access is another matter, however, about which less is presently known. We know quite a lot about how people search written text, but the characteristics of recorded speech are sufficiently different that we may ultimately find that users behave differently when searching for speech recordings [4, 9, 10, 12, 14]. Studying search behavior poses a bit of a chicken-and-egg dilemma, however – we learn what kind of support people need by observing their search behavior, but we cannot observe their behavior until we have built a system that they can search. Internet-based speech retrieval systems are now starting to appear, so we believe this is a propitious time to begin to explore how they are being used. In this paper, we propose the design of a study to examine that question.

2. CONCEPTUAL FRAMEWORK
The concept of relevance is widely used as a basis for evaluating the effectiveness of information retrieval systems [6, 8]. Researchers have sought to define relevance from two perspectives that are often referred to as system-oriented and user-oriented. The system-oriented perspective focuses on topical relevance and concerns finding documents that address a concept -based information need. Recall and precision are typically used as measures of effectiveness with this view of relevance. The user-oriented perspective on relevance is somewhat broader, seeking to characterize the relationship between information and the user's problem situation and attempting to account for the various aspects of human cognitive processes used in making relevance judgments. In this view, common terms that refer to relevance are *utility*, *pertinence*, *satisfaction*, and *situational relevance* [7, 13]. The user-oriented view does not reject topical relevance – rather it sees it as one of many factors that affect the behavior of searchers [11]. Because we seek to understand search behavior from the broadest possible perspective, we have chosen to adopt a user-oriented view of relevance for our study.

Table 1. Some bases for selecting journal articles

Criteria	Associated Attributes
Topicality	Title, abstract, keyword
Novelty	Title, abstract, journal, publication date
Authority	Author, affiliation, journal, publisher
Recency	Publication date
Reading time	Number of pages
Availability	Owning library
Accessibility	Language, media

The cognitive processes underlying human relevance judgments have been widely studied, often with the goal of identifying factors that influence relevance judgments [1, 5, 7, 11]. Table 1 shows some of the most commonly cited factors that have been identified by previous studies of searchers seeking journal articles. In general, searchers seem to base their assessment of relevance on criteria that they are able to articulate; for example, they may balance the novelty of a document – how new the ideas are to them – with the authority of the source. Criteria such as novelty and authority are abstract concepts, however, so searchers must ground their interpretation of each criterion in some set of document attributes. For example, a searcher might assess the novelty of articles in a journal that they read regularly based solely on the journal name

and publication date. For articles in an unfamiliar journal, however, they may need to examine the abstract of each article.

Some of the relevance criteria and attributes identified in Table 1 may be directly applicable to speech recordings, but others may not. Furthermore, the characteristics of specific genres of recorded speech may affect these factors. For example, the distinction between a news program and a news story would be important in news broadcasts, while in recorded classroom lectures the distinctions among course, section (in multi-section courses), and session would be more useful. Table 2 summarizes some relevance criteria and associated attributes that may be applicable to recorded news broadcasts.

Table 2. Possible bases for selecting news broadcasts

Criteria	Associated Attributes
Topicality	Program title, story, summary, speaker
Novelty	Story, summary, program title, date
Authority	Speaker, affiliation, program title
Recency	Date
Listening time	Story length
Accessibility	Language, file type

3. RESEARCH DESIGN

The goal of our proposed study is to characterize the relevance criteria that searchers apply when searching a collection of recorded radio programs and the attributes of the recordings on which those criteria are based. We seek to focus on understanding how users perceive relevance; at this stage, we are interested in the cognitive process that results in a relevance judgment, not merely in the outcome of that judgment process. Qualitative research methods are well suited to a study of this type [2, 3], so we have adopted a research design based on case studies, one of the most widely used qualitative methods.

3.1 Research Questions

The central issue that we wish to understand is how searchers decide which recordings are relevant to their needs when using an interactive retrieval system. In order to explore this issue, we have adopted the flowing "foreshadowing questions" to focus our inquiry:

a. What criteria do searchers rely on when choosing a recording?
b. How do searchers integrate multiple criteria when deciding whether to select a document?
c. What attributes of the recordings do searchers use as a basis for assessing each relevance criterion?
d. How do searchers integrate evidence from multiple attributes when assessing the relevance of a document?
e. What presentation strategies best convey useful attribute information to searchers?

3.2 Search Systems

Two broad classes of audio search technology have emerged on the Web. National Public Radio's NPR Online (http://www.npr.org/archives) is an example of a site that supports searching based on manually prepared transcripts, summaries, and/or metadata. Compaq's SpeechBot (http://speechbot.com) is an example of the other approach, in which speech recognition technology is used in conjunction with a limited amount of automatically obtained metadata such as date and source to support searches. Fortuitously, NPR Online and SpeechBot index some of the same programs, including American RadioWorks, Car Talk, The Diane Rehm Show, Fresh Air, Marketplace, and Public Interest. Each system accepts a text query, optionally with some desired values for metadata (e.g., program date), and returns a list of hits in order of decreasing likelihood that the recording will satisfy the query based on the information that is indexed. The systems differ in terms of what is indexed, the information displayed for the search results, and other user interface design details. Each system allows searchers to replay part or all of each program that is presented in the search results (in each case, using RealPlayer). We plan to use both NPR online and SpeechBot, focusing on some of the programs that they have in common, in order to explore a broader range of issues than would be possible with either system alone.

3.3 Participants

Ideally, we would like to explore the behavior of experienced searchers using a system with which they are familiar. Web-based audio searching is still relatively new, however, so it would be difficult to identify participants for our study that have these characteristics. We have therefore chosen to recruit from an accessible group of potential participants that approximates these desirable characteristics to some degree. We plan to recruit students enrolled in a Fall 2001 graduate-level seminar on visual and sound materials, in which students study acquisition, preservation, access, and management issues. All the students have completed a prerequisite course on information access, but few are likely to have experience with audio searching. We have arranged with the instructor to help design an assignment that will expose the students to Web-based audio search technology using the systems that we have chosen for our study. Participation in the study will be voluntary – students wishing to participate may do so by agreeing to allow us to observe them as they complete the assignment; other students may choose to complete the assignment on their own. There are sixteen students registered for the class, and we expect that a substantial fraction of that number will volunteer.

4. DATA COLLECTION

Our data collection plan includes observation and think-aloud during each search, and one semi-structured interview at the end of the session. Each participant will be asked to perform a series of searches. Some searches will be based on narrow questions that can be answered by listening a single recording; others will reflect a broader information need. At least one search will be based on an information need that is developed independently by the searcher. During each search, an observer will make notes that capture their impression of the searcher's behavior using our criteria/attribute framework. While searching, each participant will be asked to think aloud, explaining in their own words why they formulated a query in a certain way, selected a specific recording, or took some other action. Finally, a semi-structured

interview will be conducted at the end of the session to obtain additional insight into the relevance criteria and attributes that each participant used to find relevant recordings. We expect that a typical session will take about 90 minutes, but each participant will be allowed to continue his or her search as long as desired. The anonymity of participants will be protected by coding all records with a participant number rather than a name and by limiting access to information that might tend to reveal the identity of an individual participant.

4.1 Observational Protocol

The focus of the observer's activity will be on understanding how the searcher chooses to select or not to select a recording, not on how well the results actually meet their needs. Any unexpected behavior may also be noted and used to guide clarification questions during the interview. The observer will not interrupt the searcher during a search, and searchers will be asked not to consult the observer as a source of expert advice during the session. The observer may, however, help the user start the Web browser and reach the appropriate search page, since those are not tasks that we seek to study.

4.2 Think-aloud Protocol

Think-aloud methods have both advantages and disadvantages as a component of qualitative study designs. One important concern is that verbalizing thoughts can inspire introspection, which in turn might alter the behavior that we wish to study. On the positive side, however, think-aloud can provide insights into the cognitive processes of a searcher that may not be available in any other way. We have considered these factors and chosen to ask participants to think aloud. Some brief training on the think-aloud will be given immediately before the actual session. We will inform the participants of our interest in the way they formulate queries and select recordings, but we do not plan to provide them with specific guidelines on what we want them to talk about or how they should express their thoughts. With the consent of each participant, the think-aloud will be audiotaped and subsequently transcribed.

4.3 Interview Protocol

A semi-structured interview will be conducted immediately following each participant's last search. The goal of the interview is to obtain additional information about the process by which a participant made relevance judgments. Figure 1 identifies the topics that will be explored and a suggested question that can be used to initiate discussion on each topic. With the consent of the participant, the interviews will be audiotaped and subsequently transcribed.

1. What relevance criteria were applied in some specific cases? (Suggested question: Why did you choose to listen to [some specific recording]?)
2. In those cases, how were the criteria used together to reach a decision? (Suggested question: Were some factors more important than others?)
3. In those cases, how were attributes of the recordings used as a basis for assessing each relevance criterion? (Suggested question: How did you determine that the [topic was appropriate, source was authoritative, etc.])
4. What aspects of the design of each system were beneficial? (Suggested question: What features of each system were most helpful?)
5. What capabilities were not present in either system that would have been desirable? (Suggested question: Were there any features that you had expected to see in an audio search system that were not present in either system?)

Figure 1. Semi-structured interview questions

5. PILOT STUDY

The time constraints of the course will require that all sessions be completed within two weeks. We therefore plan to validate our data collection process in advance using a small pilot study with one or two graduate students who have backgrounds similar to those of the study population but who do not plan to take the course.

6. DATA ANALYSIS

Data analysis will begin as soon as the first search session concludes. Observational notes, think-aloud transcripts, and semi-structured interview transcripts will be categorized based on a conceptual framework evolved from the relevance criteria and attributes identified in Table 2. New categories will be created if the analysis reveals additional criteria and/or attributes. The QSR NUD*IST system provides extensive support for qualitative analysis of coded datasets, so we plan to code our categories for use with that software. We will seek to confirm indications obtained from one data source using another, a process known as triangulation, in order to gain confidence in the reliability of our interpretations. As we learn more about the cognitive processes of early participants, we will use that understanding to guide our design of probing questions in subsequent semi-structured interviews with other participants. We plan to employ transition diagrams to depict the interaction between query formulation and document selection, and matrices resembling Table 2 will be used to explore patterns and trends in the application of relevance criteria and associated attributes to make decisions.

7. VALIDITY ISSUES

Our study design includes purposive sampling of the possible searcher population, but the limited experience of our participants does pose a threat to the validity of our study that we will need to recognize when reporting our findings. We will take three steps to enhance the validity of our analysis: triangulation (described above), member checks, and peer debriefing. The participants in our study (the "members") are certainly in the best position to assess whether we have interpreted their actions and statements correctly, so we plan to check our results with them in two ways. First, we plan to meet privately with some of our participants to discuss our findings and the manner in which we reached those conclusions. Second, we will offer to present

our research results to the class near the end of the semester and solicit their comments.

Qualitative study designs rely heavily on subjective interpretation, so the validity of the analytic process is also an important concern. Member checks offer excellent insight into the validity of our interpretations, but they cannot ensure that our analysis process is applied in an appropriate way. For this reason, we will ask some of our colleagues at the University of Maryland with experience in qualitative research to review our categorization, coding and analysis processes.

8. CONCLUSION

Speech retrieval systems are now beginning to appear on the Internet, but we do not yet understand well how these systems will be used. We believe that the question of how people will use such systems is best explored using qualitative research methods, and in this paper we have proposed the design of a case study. Although our usage scenario, a class assignment, is necessarily somewhat artificial, we believe that our proposed study can offer a useful degree of insight into the issues that we have raised. Any user study requires a substantial investment of time and effort, so we would welcome comments on our proposed design that could enhance the potential payoff from this investment. In addition, it is our hope that the insights that we gain will be of benefit to the designers of future speech retrieval systems. With that in mind, we feel that it is important to help advance the dialog between those who build systems and those who study their use. Perhaps our discussion of this study design can be one step in that direction.

9. ACKNOWLEDGMENTS

The authors are grateful to Delia Neuman and the reviewers for their thoughtful comments on an earlier draft of this paper and to Thomas Connors for inviting us to work with his class.

REFERENCES

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

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

[3] Maxwell, J.A. (1996) *Qualitative research design: An interactive approach*. Thousand Oaks, CA: Sage Publications, Inc.

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

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

[6] Saracevic, T. (1976) Relevance: A review of a framework for the thinking on the notion in information science. *Advances in Librarianship*, 6, 79-138.

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

[8] Schamber, L., Eisenberg, M.B, and Nilan, M.S. (1990) A re-examination of relevance: Toward a dynamic, situational definition. *Information Processing & Management*, 26(6), 755-776.

[9] Slaughter, L., Oard, D.W., Warnick, V.L., Harding, J.L., and Wilkerson, G.J. (1998) A graphical interface for speech-based retrieval. *The Third ACM Conference on Digital Libraries*, Pittsburgh, PA.

[10] Tombros, T., & Crestani F. (2000) Users' perception of relevance of spoken documents. JASIS, 51(10): 929-939.

[11] Wang, P. & 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.

[12] Whittaker, S., Hirschberg, J., & Nakatani C.H. (1998) Play it again: A study of the factors underlying speech browsing behavior. In *Proceedings of ACM CHI 98*, April 18-23.

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

[14] Yankelovich, N., & Lai, J. (1998) Designing speech user interfaces. In *Proceedings of ACM CHI 98*, April 18-23.