# A Graphical Interface for Speech-Based Retrieval

**Laura Slaughter, Douglas W. Oard, Vernon L. Warnick, Julie L. Harding and Galen J. Wilkerson**

Digital Library Research Group, College of Library and Information Services
University of Maryland, College Park, MD 20742 USA
{lauras, oard, skip, jharding, galn}@glue.umd.edu

**ABSTRACT**

This paper describes preliminary usability testing for a graphical interface designed to facilitate rapid browsing of speech records. Expert interviews and focus group discussions were used to assess the alignment between browsing behaviors employed by members of the intended user population and an early mockup of the interface. The results provide guidelines for the next iteration of prototype development and suggest that graphical representations offer a viable method for browsing audio and multimedia recordings.

**Keywords**

Speech-based retrieval, graphical interfaces, digital library

**INTRODUCTION**

Imagine searching 100,000 hours of speech-based recordings at the National Archives without recourse to hand-annotated metadata. If present trends continue, it will soon be possible to transcribe 10,000 hours of speech per year on a single computer, however, automated systems do not operate in a vacuum. We know from text retrieval experience that users exploit information such as title and source in a selection interface to identify promising documents. Speech recordings can provide additional cues, such as the number of speakers and their turn-taking behavior. In this paper we present the results of an initial investigation into the potential of such cues to browse speech recordings.

Present experimental speech-based retrieval systems have fairly simple graphical displays for browsing speech recordings. The Cambridge Video Mail Retrieval (VMR) system presents a subject line and a representation of the predicted degree of relevance of the recording to the user's query [1]. When a specific recording is selected, the VMR interface facilitates browsing with a horizontal timeline on which each occurrence of a query term is indicated. The CMU Informedia system uses a similar strategy, showing some content words instead of a subject line [2]. No present speech-based retrieval systems depict turn-taking behavior, but such interfaces have been used for analysis of meetings. Kimber, et al developed an interface in which speech segments - contiguous periods of speech by a single speaker - were depicted using horizontal line segments [3]. The segments extend across the horizontal axis and are separated vertically indicating each individual speaker. The Jabber system developed at the University of Waterloo used a similar display [4].

VoiceGraph, shown in Figure 1, is our first prototype of a graphical selection interface for speech-based retrieval [5]. Search controls in the upper left allow specification of query terms and, where known, speaker identity. The selection interface in the upper right can display as many as 20 alternation patterns, with the line segments colored to indicate speaker category (e.g., male or female). Once the user selects a specific recording, a portion of the recognized text is displayed in the lower right. Metadata such as date and duration are displayed in the lower left along with the controls for audio replay. We have focused this initial study on understanding how well our selection interface design will support the information-seeking behavior of the intended user population because that component represents the most significant departure from present practice.

**METHODS AND RESULTS**

We chose to focus on two intended user groups for the VoiceGraph system, journalists and professional librarians. Static screen shots and mock-ups were used to elicit responses from the participants since we conducted the study before the search functions were integrated. Our principal goal was to better understand how users search

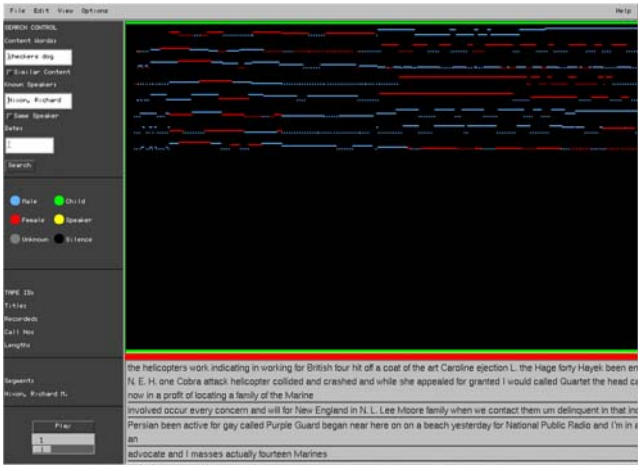

Figure 1. The VoiceGraph preliminary interface design.

for speech data and how a browsing interface could support that search. Two domain experts were interviewed, one faculty member from journalism and one from the library school. Two focus groups were conducted, the first consisted of eight journalism students, and the second consisted of two library school students. All participants attended the University of Maryland, were enrolled in graduate-level courses, and none had prior exposure to the interface.

The first part of the interviews and focus groups was aimed at eliciting the functionality desired by the selected user groups. Participants were asked to identify attributes they use when searching recorded speech. They named keywords, speaker name, historical period, topic, location, occasion/event, title, date, genre, affiliation, duration, language, tone, cadence, source, and file type (e.g., .au or .wav). They were then asked to draw a visual representation of a speech file before viewing the VoiceGraph interface. This proved to be difficult for most participants, who typically drew icons depicting topics, pictures of events and photos of speakers. Participants tried to generate generic icons that communicated content. The only exception was the library school faculty member who attempted to draw a representation of "discernable conversation patterns."

After viewing the VoiceGraph interface, discussions focused on the feasibility of using a graphical display to browse speech, suggestions for ways to represent speech files, and potential improvements to the system. All participants felt that a graphical representation of speech could be a useful "high level" tool for browsing sets of speech files. The library school focus group stated that without text or sound output, alternation patterns alone would not support effective selection of relevant speech files. On the other hand, the library school faculty member stated that graphical alternation patterns "might be a good medium for providing relevance feedback to the user" and that they appeared to be an efficient method for distinguishing between speech records. The journalism faculty member and the library school students discussed that the intended use for the recordings would change the search strategy. Browsing for generic sounds, such as two children talking, was described as different from browsing for content-related speech.

A modified version of a subjective satisfaction questionnaire (QUIS) was used to assess the interface. [6] The overall mean for all questions was a 5.3 on a 1-9 likert scale. The scale is arranged so that positive adjectives anchor toward 9, negative toward 1. Comments and question ratings on the QUIS indicated areas for interface improvements. The quantitative questionnaire results supported the focus groups and interview results.

## LESSONS LEARNED AND FUTURE PLANS

Users suggested several enhancements such as incorporating additional search criteria (e.g., source) and clustering the retrieval results. We also learned that the idea of using alternation patterns to identify promising files may not be easy for novice users to grasp. We ultimately plan to add informative keywords to each alternation pattern. Another enhancement is to permit searches based on criteria such as the number of speakers and/or the alternation pattern structure (e.g., monologue or dialog).

From this study we are able to conclude that graphical representation of speech is a potential method for browsing sound files. The results found from this preliminary usability testing will be combined with testing that is in progress using a functioning prototype with search capabilities. With a comprehensive usability report, we intend to add search characteristics that support user identified search tasks and create a user-centered design.

## ACKNOWLEDGMENTS

This work was supported in part by Army Research Institute contract DAAL01-97-C-0042 through Micro-Analysis and Design, and by IBM through a SUR equipment grant. The authors would like to thank Tony Tse and Gary Marchionini for their helpful comments.

## REFERENCES

1. Young, S.J., Foote, J.T., Jones, G.J.F., Sparck Jones, K., and Brown, M.G., "Acoustic Indexing for Multimedia Retrieval and Browsing", in *Proc. of ICASSP 97,* April 1997.
2. Informedia Project. http://www.informedia.cs.cmu.edu/
3. Kimber, D.G., Wilcox, L.D., Chen, F.R. and Moran, T.P. "Speaker segmentation for browsing recorded audio" in *Proc. Human Factors in Computing Systems*, 1995, pp. 212 – 213.
4. Kazman, R., Al-Halimi, R., Hunt, W., and Mantei, M., "Four Paradigms for Indexing Video Conferences", *IEEE Multimedia*, Spring 1996, pp. 63-73.
5. Oard, D.W., "Speech-Based Information Retrieval for Digital Libraries" (Technical Report CS-TR-3778). College Park, MD: University of Maryland, 1997.
6. Chin, J. P., Diehl, V.A., and Norman, K.L., "Development of an instrument measuring user satisfaction of the human-computer interface", in *Proc. of Human Factors in Comp. Sys.*, pp. 213-218, 1988.