# Speech-Based Information Retrieval for Digital Libraries*

**Douglas W. Oard**
Digital Library Research Group
College of Library and Information Services
University of Maryland, College Park, MD 20742
oard@glue.umd.edu

### Abstract

Libraries and archives collect recorded speech and multimedia objects that contain recorded speech, and such material may comprise a substantial portion of the collection in future digital libraries. Presently, access to most of this material is provided using a combination of manually annotated metadata and linear search. Recent advances in speech processing technology have produced a number of techniques for extracting features from recorded speech that could provide a useful basis for the retrieval of speech or multimedia objects in large digital library collections. Among these features are the semantic content of the speech, the identity of the speaker, and the language in which the speech was spoken. We propose to develop a graphical and auditory user interface for speech-based information retrieval that exploits these features to facilitate selection of recorded speech and multimedia information objects that include recorded speech. We plan to use that interface to evaluate the effectiveness and usability of alternative ways of exploiting those features and as a testbed for the evaluation of advanced retrieval techniques such as cross-language speech retrieval.

## Introduction

Future digital libraries are expected to contain vast holdings of material in electronic form. Although recorded speech and multimedia objects which include recorded speech comprise only a small portion of most library collections, enormous quantities of such material are being produced on a daily basis. As the cost of digital storage continues to fall, it will become increasingly practical to collect and store such material. Access to such collections poses a serious challenge, however, because present search techniques based on manually annotated metadata and linear replay of material selected by the user do not scale effectively or efficiently to large collections. In this technical report we propose a comprehensive approach for discovering information objects in large digital collections based on analysis of recorded speech in those objects.

Speech-based information retrieval is a special case of the information retrieval problem in which the information content of the objects in the collection is determined through analysis of recorded speech contained in those objects. Concurrent increases in processing power and the aggressive development of increasingly sophisticated algorithms for processing spoken dialogs recorded in natural environments have combined to make it practical to use recorded speech to help users locate speech or multimedia objects which satisfy their information needs. We refer to this process as speech-based information retrieval because the goal of the process is to retrieve information objects (that may include multiple modalities such as speech and video) using speech, and it is not limited to retrieval of the speech itself. The research proposed here seeks to identify a set of speech features that users find valuable when seeking information objects, and to develop effective ways of exploiting those features to enhance both the effectiveness and usability of a speech-based information retrieval system.

Existing technology makes it possible to identify three broad types of information in recordings of spoken language. "Speech recognition" seeks to determine the information content of the speech. Transcription to written text and keyword spotting within (otherwise ignored) speech are examples of speech recognition. "Speaker identification" seeks to determine which speaker generated a particular speech segment. Speaker verification, in which the system attempts to determine whether the a specific speech segment was generated by a specific speaker, is one example of speaker identification. Finally, "language identification" seeks to determine the natural language (English, French, Chinese, ...) being used by a speaker. Dialect determination is a special case of language identification, and accent identification is a closely related task.

*This work was supported, in part, by Army Research Institute contract DAAL01-97-C-0042 through MicroAnalysis and Design.

In the next section we describe the conceptual design of a prototype speech-based information retrieval system that combines these capabilities in meaningful ways that may contribute to enhanced effectiveness or usability. That description is integrated with a discussion of the speech processing technology on which such a system could be based. We then describe the characteristics of the available collections of recorded speech that could be used to evaluate that system. With that as background, we then describe the initial design of some experiments to determine which aspects of the user interface make the greatest contribution to effectiveness and usability. We conclude with some remarks on the potential significance of the proposed research.

## User Interface Design

Our user interface design is based on a ranked retrieval paradigm in which available information objects are assigned positions in a single ranked list in a way that seeks to place the most useful objects (in the user's judgment) near the top of the list. Real user needs typically extend beyond simple topical relevance, incorporating additional factors such as whether the user has already developed the conceptual framework that would be needed to interpret the information content of an object, and whether the object contains information which extends the user's present understanding of the topic rather than simply duplicating it (Soergel 1994). Ranked retrieval has proven to be a powerful basis for user interface design in other information text applications because it permits a synergistic combination of the machine's ability to apply relatively simple techniques to large amounts of information with human abilities to apply sophisticated selection strategies to limited sets of objects.

The ranked retrieval paradigm is particularly well suited to the technical characteristics of existing speech recognition and speaker identification technology because many of the available techniques for those tasks produce confidence measures that can serve as a basis for constructing a ranked list. The remaining challenge is to depict information about the content of each information object in a way that allows users to recognize useful information objects in a ranked list. In text retrieval, document titles are often used used for this purpose. In video retrieval, miniature representations of key frames extracted from each scene can be used. Both of these ideas can be used in speech-based information retrieval as well if the required information (topically informative text annotations or synchronized video) is associated with the recorded speech. But the most challenging scenario for speech-based information retrieval is presented when only recorded speech is available and thus the document surrogate displayed in the ranked list must be based on speech information alone. We defer for the moment further discussion of how the ranked list is formed in order to first describe how representations of recorded speech can be presented to a user in a way that facilitates the selection of useful information objects.

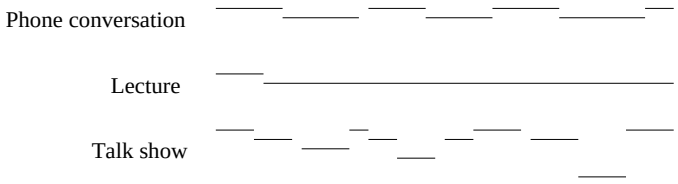


Figure 1: Speech alternation pattern examples.

### Speech-Based Selection Interface

Document titles and thumbnail images provide visual cues about the content of an information object that exploit human perceptual and cognitive capabilities to facilitate selection of promising information objects for more detailed examination. Human visual perception is characterized by high bandwidth, and cognitive abilities permit the application of sophisticated heuristics which readily accommodate complex interactions between the nature of the query that was specified and the characteristics of the objects that are displayed. It is those characteristics that we seek to exploit when designing a selection interface for speech-based information retrieval.

Our basic approach is to display a color-coded alternation pattern of the spoken dialog as a graphic aid to selection, and to augment that information with additional metadata in text form when useful metadata can be obtained. Figure 1 shows examples of simple alternation patterns for some common types of recorded speech. In those alternation patterns, the horizontal axis represents time and the vertical axis within a single alternation pattern represents speaker identity. The first example shows a telephone conversation, in which two speakers alternate talking. The second example is a lecture in which the host briefly introduces the lecturer and then the lecturer talks for an extended period. The third example depicts a portion of a radio talk program in which two speakers dominate the discussion, but several other speakers participate episodically.

Kazman, *et al.* have used automated analysis of such alternation patterns as a basis for automatic categorization of dialog types in an electronic meeting support system(Kazman *et al.* 1996). Our application differs in that we propose to present the alternation

patterns directly to the user as one source of information that can facilitate selection. We are not aware of prior use of alternation patterns in the selection interface component a speech-based information retrieval system. Such alternation patterns can be constructed automatically by first recognizing boundaries between speech segments that are associated with a change of speakers and then solving a set of open set speaker identification problems to determine which subsequent speech segments are associated with the same speakers. Wilcox, *et al.* have developed an iterative algorithm for this task which achieved 99.5% accuracy on a small test collection with three speakers (Wilcox, Kimber, & Chen 1994). Zissman and Weinstein have also described useful techniques for recognizing periods in which both speakers are talking simultaneously, a common occurrence for some sources of recorded speech (Zissman & Weinstein 1990).

Speaker identification can also provide a basis for coloring the lines that represent speech segments. Gales has demonstrated techniques for determining whether a speaker is male or female, and has speculated that the technique could be extended to include a third category of speech by children if training data such as that recently collected by Miller, *et al.* is available (Galles 1993; Miller *et al.* 1996). The three primary colors could be used to label speech by adult males (blue), adult females (red), and children (green). Another color (yellow, perhaps) can be used to indicate speech by a specific known speaker when such an indication is appropriate to the type of search being conducted. Speaker-specific searching is discussed below. White can be used to represent segments containing music. Saunders has developed techniques for music detection (Saunders 1996), and Wold, *et al.* have described techniques for retrieval based on non-speech audio that could serve as a future enhancement to the speech-based search capabilities described below (Wold *et al.* 1996). A menu selection should be provided to allow the user to enable or disable the depiction of music, since music detection may introduce undesirable display clutter for some types of collections. When music is depicted, it could be shown as a white region extending across all of the speech lines in an alternation pattern. No lines at all would be displayed for periods when no speech (or music, if selected) is present in the recorded audio. A menu function will be needed to change the default color assignments in order to accommodate user preferences and perceptual limitations.

Another basis for marking the lines that represent speech segments is automatic language identification. For a language-specific search (also discussed below), solid lines can be used to represent speech in the specified language. If some information objects contain speech segments in other language as well, dashed lines can be used to represent those segments. In this way it is possible to encode both speaker category (adult male, adult female, child, known speaker, non-speech audio) and language (specified, other) simultaneously. When no language is specified for the retrieval operation, solid lines can be used to depict every speech segment.

If the number of depicted speakers is limited to five, a total of 20 information objects can be represented in a 600x720 pixel selection interface window on a 17-inch SVGA (768x1024 pixel) display. If more than five speakers are recognized in an information object, the first five speakers could be represented using solid or dashed lines as described above and a sixth line, always dotted (regardless of language) could be used to depict the presence of an additional speaker. This approach facilitates accurate construction of the alternation patterns by limiting the number of distinct speakers that must be recognized while simultaneously minimizing display clutter. As with other lines, dotted lines indicating undistinguished speakers could still be colored to indicate speaker category. A menu function allowing user selection of the desired maximum number of depicted speakers would likely be useful, since the best choice for this parameter may depend on user preferences and the characteristics of the collection.

A consistent time scale should be used for each information object in order to facilitate comparisons between simultaneously displayed objects, but we may consider introducing a horizontal "fisheye" on the time scale in order to simultaneously provide a comprehensive overview of long information objects and a useful level of detail in the region designated by the user. In our initial implementation we are considering a fixed temporal resolution of five seconds per pixel which would allow a one hour section of an information object to be depicted in a 600x720 pixel selection interface display area. Users should, however, be provided with a menu function to adjust the temporal resolution. In order to allow for display and perceptual limitations, a minimum of two consecutive pixels (10 seconds at the default temporal resolution) should be required before a line representing a speech segment is displayed.

Each type of search described below results in automatic identification of one or more speech segments in every retrieved object that satisfy the search specification. The speech segment which best satisfies the search specification should be centered horizontally in the selection interface display area in order to help the user quickly locate interesting speech segments. A hor-

izontally centered vertical line passing through each alternation pattern can be provided to help the user recognize those segments, and other segments in the same object which also satisfy the search specification could be marked with additional vertical lines that are limited to that particular object's alternation pattern. Alternation patterns which would extend beyond the edge of the selection interface display window could be displayed with fuzzy edges near the window border to indicate to the user that the pointing device (e.g., a mouse) can be used to scroll that alternation pattern horizontally in order to see the remainder of the pattern. When scrolling an alternation pattern, all of the vertical lines marking potentially interesting segments (including the appropriate portion of the main vertical line) should scroll horizontally with the alternation pattern itself.

**Multimedia Selection Interfaces** When other modalities such as video are present, the selection interface could be enhanced using material from those modalities. In this section we present techniques for integrating information from a video stream that is synchronized with the recorded speech. Integration of synchronized closed-caption text is described in the next section.

Information from the video stream can be integrated directly into the alternation pattern display. Wolf has developed a motion-based technique for identifying one or more key frames in a video segment (Wolf 1996). The speech alternation patterns provide a basis for video segmentation, and a thumbnail representation of one key frame for each sufficiently long speech segment could be overlaid on the line that represents that segment. Selection of this image-augmented mode would require a significant change in both temporal resolution and the number of displayed information objects in order to accommodate reasonably-sized thumbnail images. If 50x50 pixel thumbnail images are desired, for example, then ten five-minute alternation patterns could be accommodated in a 720x600 pixel selection interface. In this configuration would be possible to associate a key frame with any speech segment that has a duration of at least 30 seconds.

## Content Display

A separate display area will be provided for automatically recognized content from a speech segment. Placing the pointing device within the vertical limits of an alternation pattern and clicking with the left button will cause the text representation of the speech content of that information object to be displayed in the content display window. The display area below the selection interface on a 768x1024 pixel display should be adequate to display approximately 80 columns and 10 lines of text using a reasonable font size. This corresponds to approximately one minute of recorded speech. Muted horizontal and vertical bars in an unused color (beige, perhaps) can be used to indicate the selected object and the temporal extent of the displayed speech content respectively. The text associated with the segment at the center of the vertical bar should be centered in the display area, with any text from preceding and following segments that will fit in the text display window rendered in a slightly smaller and noticeably lighter font. A scroll bar should be provided in the text display window to allow the user to smoothly scroll through the recognized text, and the horizontal position of the vertical bar should be adjusted so that the user will be able to easily identify the segment which is the source of the displayed text. "Page up" and "page down" keyboard functions could also be provided to allow the user to move in window-sized increments through the recognized text. Longer jumps are supported by this design by allowing the user to select another segment for examination by clicking on it using the left button and the pointing device.

Figure 2 provides an idea of how this information will be displayed when only audio sources are available.[1] If closed caption text is available, the user may specify that closed caption text should be displayed in place of or in addition to the speech recognition results using a menu selection. If simultaneous display of recognized speech and closed caption text is selected, the content display window should be split horizontally to facilitate comparison between the two sources of content information. The user should be allowed to adjust the vertical extent of the content window (at the expense of the selection interface display) if additional content information is desired.

Although recognition errors may limit the value of the automatically recognized speech, we expect that users will find that even error-filled recognized speech is a useful adjunct to the other information that is available in the other display fields. Resnik has developed a gisting technique for cross-language text retrieval in which multiple translations are displayed if unresolvable ambiguity is encountered in the translation process, and he reports that users were able to read past the multiple alternatives and gain the gist of the material presented without undue difficulty (Resnik 1997). It is not clear whether this technique would be equally

[1] Since the figure is smaller than the depicted display, the number of speakers and the number of alternations with a single pattern have been limited. A white background is used for clarity in this monochrome presentation — the default color pattern is intended for use with a black background.

effective with recognized speech because recognition errors may not be limited to isolated words, but we plan to investigate this and to incorporate Resnik's technique in the content display if it proves to be worthwhile.

**Auditory Content Display** One important goal of our user interface is to provide access to multimedia objects that include recorded speech, so direct presentation of the audio content is an important aspect of the user interface. The auditory display capabilities we plan to implement are also important for information selection, since humans are able to extract far more useful information from recorded audio than any machine. Although a graphical user interface can display a substantial amount of information about speech content, the speaker, and the language used, augmenting the user interface with an effective auditory display capability will enable the user to base selection and query reformulation decisions on a broader range of useful characteristics.

Users may reproduce the recorded audio for any speech segment by clicking on the "play" button in the content display window using the left mouse button and the pointing device. Playback will begin at the start of the selected segment, and it will continue until a the playback function is deselected by clicking on the "play" button again. The "play" button should be highlighted during playback in order to make this alternate action functionality clear to the user. A slider bar should be provided to allow users to adjust the degree of acceleration used when reproducing the recorded audio.

Three accelerated playback techniques can be provided — pause elimination, Synchronized Overlap and Add (SOLA) time compression, and dichotic time compression — and menu selections should be offered to allow the user to enable or disable each technique (Arons 1994a). Dichotic time compression should not be selected by default because users unskilled in its use report that the unusual sensory effect associated receiving slightly different audio content in each ear is disturbing, but including it as a menu option will offer experienced users the opportunity to achieve faster playback for a given level of intelligibility. A headset can be used in environments where external sound sources would pose a problem or when aural reproduction must be limited to a single user, and their use is mandatory if dichotic time compression is selected.

**Multimedia Content Display** When synchronized video content is available, a portion of the content display window can be reserved for video. Before "play" is selected, this area should contain a larger version of the key frame associated with the selected speech segment. Once replay is initiated, the key frame should be replaced with video. Video replay must be synchronized with the audio content for both the normal-speed and accelerated replay modes, and an alternative highly accelerated replay mode (without audio) should be available as an additional menu selection.

Video scene analysis can provide additional cues for the detection of significant shifts in the content of an information object (Kobla, Doermann, & Rosenfeld 1996). When scenes and speakers change nearly simultaneously, the change may be more meaningful to the user than when either changes in isolation. For this reason we plan to offer an additional menu selection to allow the user to define the start point for the "play" action based on either speaker shifts or on (presumably less frequent) speaker shifts that occur at very nearly the same time as a scene shift. We expect that this option may offer users a replay functionality that is more closely linked to their perception of thematic shifts in the recorded material.

## Metadata Display

The third major display area will contain data about the information object and about the particular segment of that information object that the user has selected using the pointing device. Some of these "metadata" elements, such as the date a recording was made, can only be obtained from external sources that will vary from collection to collection. For this reason, we plan to identify a standard set of external metadata that will be presented in the metadata display window if that metadata is available for the information object indicated by the user. Although the metadata we choose may not be appropriate for every possible collection of recorded speech to which our system might eventually be applied, we expect to be able to identify a sufficiently rich set of external metadata to support our evaluation goals fairly easily. We intend to accommodate the eventual need to add additional types of metadata by designing fairly flexible display management software using object oriented techniques.

A second type of object-level metadata is automatically extracted information such as the number of speakers and the temporal extent of the information object. Segment-level metadata such as the language spoken and (if known to the system) the identity of the speaker must all be automatically extracted since recorded speech is not typically annotated with external metadata about individual speech segments. Automatically extracted object-level metadata should be integrated with the external metadata in one part of the metadata display window, and segment-level meta-

data should be grouped in a separate area. Menu selections could be provided to enable or suppress the display of specific metadata elements if display clutter becomes a concern. Figure 2 illustrates the metadata display area in the lower left corner of the display.

## Query Interface

The final display field provides facilities for the user to issue queries against the collection. The query itself will remain displayed upon completion of the ranked retrieval operation in order to help the user interpret the rank ordered information objects that are represented in the selection interface window. Queries can be based on four types of information: speech content, speaker identification, language identification, and metadata.

**Content-Based Queries** Content-based queries can be initiated either by explicitly typing query keywords or through a query-by-example technique (known as "relevance feedback") in which the user designates an object or a portion of an object as representative of the information need. A button on the pointing device can be depressed as the user performs a sweeping motion to highlight the portion of an information object that should be used for this purpose. The highlighted sections will be indicated on the display using a light background of an otherwise unassigned color. For keyword-based searches, the words may be freely chosen by the user, and the user may specify that they be searched in either the recognized speech, the synchronized closed caption text (if available), or both. Keyword and relevance feedback queries can be combined to create a relevance feedback query in which certain words receive additional emphasis.

Ng has surveyed content-based speech retrieval techniques and has identified two basic approaches: large-vocabulary continuous speech recognition and word spotting (Ng 1996). The goal of Large-Vocabulary Continuous Speech Recognition (LVCSR) is to transcribe as much of the the spoken information to written text as is possible with reasonable accuracy. Standard ranked text retrieval techniques (which have some inherent tolerance for erroneous input) can then be used to identify the most useful information objects based on either an explicit query or relevance feedback. The same techniques can also be applied to synchronized closed caption text if it is available. LVCSR-based speech retrieval is presently a fairly dynamic research area, and it is possible that effective techniques which achieve improved performance by exploiting information about uncertain recognition will be developed (e.g., by better handling out-of-vocabulary terms or by using the internal representation of a best-n recognizer). The interface that we propose to develop can be used to evaluate the impact of such techniques on both retrieval effectiveness and usability as they become available.

Word spotting offers an alternative to LVCSR-based speech retrieval. Rather than seek to recognize every word, the goal in word spotting is to recognize the words which appear in the query and to reject as "garbage" all other words. Word spotting may be able to achieve a similar level of retrieval effectiveness to LVCSR-based speech retrieval at a lower computational cost when applied to raw speech, but such an approach would preclude real-time retrieval because the full representation of each information object would need to be examined. LVCSR, on the other hand, produces a fairly compact text representation for which rapid and fairly effective retrieval techniques are already well known (Frakes & Baeza-Yates 1992).

We are not yet ready to commit to a single technical approach for content-based speech retrieval because we wish to explore the potential for compact lattice representations that could support rapid and fairly accurate word spotting. Our ultimate choice may be influenced as strongly by the size of such a representation as by the speed and accuracy of the matching process. Another consideration dictated by the design of our interface is that regardless of the matching technique that we select, (LVCSR-based text retrieval or word spotting), LVCSR will still be needed to support the display requirements of the content window.

**Speaker-Based Queries** Speaker-based queries are, by contrast, fairly straightforward to deal with. For some collections it will be relatively easy to isolate training samples of well-known or otherwise important speakers and associate those samples with manually-entered identity information. In those cases, the user will be provided with a dynamic query interface to select from a list of known speakers. Since other speakers may also be represented in the collection, this is an open set speaker identification problem. Open set speaker identification techniques typically apply a threshold on similarity in a feature space to determine whether a specific speech instance is sufficiently similar to the training instances to declare that a match has been found. Such techniques are easily adapted to provide ranked output by building the ranked list in the order that would be produced by using increasingly relaxed thresholds.

Speaker-based queries may also be specified using query-by-example. In this case, the user need only designate a specific segment as the initial example. Because speaker identification accuracy can be improved if multiple training instances are available, additional

instances of speech by that speaker will first be sought within the same information object using the results of the analysis already performed to generate the alternation pattern display. Because a relatively small number of speakers will speak in any individual information object, this technique leverages the success of the easier within-object speaker identification task to potentially improve the accuracy of collection-wide speaker identification.

We also plan to investigate techniques for combining content-based and speaker-based queries to produce a single ranked list which could be used to locate instances of specific speakers talking about specific topics. Two techniques for combining these features are under consideration. The first approach is to allow users to specify the number of objects to select using one feature (either speaker recognition or topic recognition), and then rank order the selected objects using the other feature. The alternative approach is to provide a menu selection or a slider bar which allows the user to specify the relative importance of each type of information for forming the ranked list. This value would then be used to form a linear combination of the rank assigned to each object in ranked lists based on each technique. The resulting "merged rank" values would then be sorted into a list that is ranked based on the combination of features.

**Language-Based Queries** Identification of language, dialect and accent are currently active areas of research (c.f., (Lund, Ma, & Gish 1996; Zissman *et al.* 1996; Hansen & Arslan 1995)). We plan to initially implement only queries based on the natural language being spoken, both because we feel that language information will be useful in the greatest number of applications and because sufficiently high accuracy has been reported for language identification to allow us to consider a design based on unranked selection. Language selection will thus be implemented initially as a constraint on the set of information objects to be ranked using one or both of the other features (content and speaker), so boolean selection buttons will be used for this purpose. Several languages may be selected simultaneously if desired. Although fairly accurate, language identification is not perfect. Furthermore, speech in a language for which the system has not been trained may be present in a collection. For these reasons, we will also make an "unknown" language selection available. When designing future enhancements we will consider the addition of dialect and accent identification if language-based queries prove useful. Such enhancements would involve a redesign of the interface to allow ranking based on the likelihood that a desired dialect or accent is present if sharp identification of these features proves difficult.

## Experimental Evaluation

The goal of our experimental investigation is to identify those features that users find most useful when searching for information objects. For this reason, our system will seek to maximize effectiveness and usability. We have chosen to value these factors over efficiency at this stage of our research because we expect that the results of these experiments will allow us to focus subsequent efforts more narrowly. Using this strategy we seek to avoid unproductive optimization of techniques which, although effective under laboratory conditions, may ultimately prove to be of little benefit to situated users.

Our initial experiments will emphasize exploitation of speech information because it is the display of speech alternation patterns in our user interface that represents the most radical departure from present practice in the field. Once the effectiveness and usability of our speech-based techniques have been demonstrated, we will turn our attention to evaluation of integrated speech, video and closed caption text information. In this section we describe the goals of our planned evaluation and the speech collections that we plan to use to conduct those experiments. The information presented here is intended as an overview. Detailed experiment design is expected to proceed in parallel with system implementation.

For our experiments we will adopt both comprehensive and component-level effectiveness measures. Our comprehensive effectiveness measures will seek to evaluate the user's ability to identify useful information objects (in our initial experiments, sound recordings) based on the system's selection and rank ordering performance and the user's ability to interpret the displayed information and make informed choices. This evaluates both the performance of the retrieval techniques and the utility of the information that is displayed to the user. System performance will be maximized when the two components (the retrieval technique and the user interface) have complementary strengths.

In information retrieval, the typical strategy is to design retrieval techniques which concentrate the desirable information objects near the top of a ranked list because humans do not handle large quantities of information as well as small quantities. Thus it is useful to study the performance of the retrieval component separately from that of the user interface component as well. The performance of a retrieval technique can be characterized using traditional information retrieval measures such as recall, precision and fallout.

We plan to perform those measurements on collections of recorded speech for which a standard set of text queries are available and for which impartial judging of relevance to those queries can be performed. The spoken document retrieval (SDR) track in TREC 6 provides a useful venue for such an evaluation, and we have applied to participate in TREC 6 for that reason. In that evaluation, 50 hours of recorded speech will be available along with a standard set of transcripts produced using LVCSR. The availability of this test collection will allow us to conduct some initial experiments even before LVCSR and/or word spotting capabilities are incorporated in our system. The evaluation will be based on known-item searching over approximately 1,000 information objects.

Although it is possible to characterize (albeit imperfectly) the performance of the retrieval technique in isolation, it is more difficult to construct meaningful effectiveness experiments for only the user interface component. This has led us to select comprehensive testing as our second set of effectiveness experiments. We plan to augment our TREC 6 experiments with a "manual run" in which the user interacts with the system to identify the "known item" which satisfies the TREC topic description as quickly as possible. Future TREC SDR evaluations may adopt a more traditional evaluation technique in which sets of information objects that are topically relevant to a collection of more general queries must be identified. The "pooled assessment methodology" used to identify the largest possible number of relevant documents would benefit from the improvement in comprehensive effectiveness that a manual run of this type can produce when more than one relevant document is sought.

Regardless of the evaluation approach used at TREC, the experiments we conduct in that venue will directly evaluate only the effectiveness of our system when using content-based queries. In order to overcome this limitation, we plan to conduct additional experiments using both the TREC speech retrieval collection and collections of recorded speech from the Nonprint Media Collection and the Public Broadcasting Archives at the University of Maryland College Park Library. The library already has plans to a substantial amount of the collection in order to improve the accessibility of the material, since much of it is presently stored on the original media which is not suitable for repeated use. We have identified three portions of the collection which have particularly useful characteristics.

The first collection is a complete set of video tapes of the congressional Iran-Contra hearings. These are high quality recordings, made with a relatively small set of speakers who engaged in a protracted dialog, and they are recorded with a fairly consistent set of microphones. These characteristics make it far easier to obtain good speech recognition performance than is presently possible in less well controlled conditions. Furthermore, the material is in the public domain, so exchange of this information with our research collaborators will be greatly facilitated. Furthermore, a complete manually produced transcript is available, although it is presently available only in printed form. If an accurate electronic version of that transcript can be produced automatically, the resulting collection can be used to produce speaker-based and content-based queries that can serve as the basis for practical component-level and comprehensive known-item evaluations.

A second collection that is of particular interest for these experiments is a set of approximately 5,000 hours of audio tape reels from the National Association of Educational Broadcasters (NAEB) network that predated National Public Radio. It may be possible to arrange unlimited copyright clearance for large parts of this collection fairly easily because all of that material at least 25 years old and and copyright for the majority of the collection is held by a relatively small group of public institutions. The initial portions of this collection are presently being digitized in order to make the material available on media which can be easily handled and stored. The collection is quite diverse, and it is considerably larger than any of the other collections that we are considering. Thus it should prove a greater challenge to the speech recognition and speaker identification algorithms that we plan to incorporate in our system. It is extremely important to evaluate the performance of our system under adverse conditions, since many speech recognition algorithms demonstrate markedly worse performance with conditions that are even slightly degraded.

Neither of these collections contain a substantial amount of material in languages other than English, so we plan to use a collection of recorded audio material from the Voice of America for this purpose. Again, copyright clearance for this material should be fairly easy to obtain. We have not yet determined the full range of languages in the collection, but based on the date and location of the recordings we expect that it contains at least English, French and Vietnamese. The Voice of America collection also contains a substantial amount of music. We expect this to facilitate our evaluation of the music identification component that allows us to label music in the selection interface, and it should make it possible to determine whether this feature materially aids users seeking to recognize de-

sirable information objects.

Each of our comprehensive evaluations will include a usability evaluation as well. Usability evaluation will be the most challenging aspect of our experiment design because usability measures which are both easily determined and insightful are difficult to construct. At present we are considering measuring the time required to perform specific tasks and additionally collecting information about user preferences using questionnaires, structured interviews and direct observation of user behavior. We expect to refine these ideas as we develop detailed experiment designs.

## Possible Enhancements

In establishing the scope of this project we have sought to design a system with an interrelated set of capabilities that can be used to answer important questions about the effect that providing access to specific information in the selection interface has on the effectiveness and usability of a speech-based information retrieval system. The system we have described will also be useful as a testbed for the integration of more sophisticated search capabilities, and in this section we describe three examples of such technologies that could easily be integrated into such a system.

A number of techniques are known for cross-language text retrieval, and recently Sheridan, *et al.* have demonstrated the first cross-language speech retrieval system (Sheridan, Wechsler, & Schäuble 1997). Cross-language retrieval techniques can be use to produce systems capable of multilingual searching, making it possible to locate recorded speech in any language using query terms formulated in the users preferred language. The best known cross-language text retrieval techniques achieve about 75% of the retrieval effectiveness of within-language text retrieval, and there is every reason to believe that similar performance will eventually be achieved for cross-language speech retrieval. Monolingual users may then require speech translation services such as those being developed by Waibel (Waibel 1996), but users performing speech-based retrieval of multimedia objects may be able to use other modalities (e.g., video) regardless of the language in which the associated speech was expressed.

Another enhancement which could be incorporated is similarity-based ranking of non-speech audio using a query-by-example strategy. Musical cues are sometimes used to identify transitions from one topic to another in programs that are otherwise consist entirely of recorded speech, and users may find search facilities which exploit such cues to be valuable. Incorporation of such capabilities would also expand the range of applications for such a system by providing some useful functionality for collections of recorded audio that contain no speech at all. Wold, *et al.* have described a system for computing the similarity of recorded audio segments based on pitch, loudness, brightness, bandwidth and harmonicity (Wold *et al.* 1996). A California company, Muscle Fish, has incorporated this technology into an audio-based retrieval module for the Informix database management system.

Another intriguing possibility is the addition of a capability to find alternation patterns that are similar to an example that is designated by the user or selected from a set of prototypical alternations patterns for specific types of dialog. Kazman, *et al.* have proposed the use of augmented transition networks based on speaker identity, speech duration, and overlap between speakers for this purpose. In preliminary experiments with this technique they have discovered that such transition graphs can be used to recognize decision points in electronic meeting records.

These three examples are intended to illustrate how the system we have designed could be used to explore the utility of advanced search techniques. Undoubtedly other ideas will emerge as we proceed with the development, and which ideas we choose to implement will be guided by the applications we envision and the search strategies that we think may be effective in those applications.

## Existing Speech-Based Retrieval Interfaces

This research is certainly not the first work on speech-based information retrieval. We are not, however, aware of any existing systems or research projects which include a selection interface that exploits such a wide range of information that can be automatically extracted from recorded speech. Several existing systems do provide selection interfaces for recorded speech or non-speech audio, however, and we discuss their contributions in this section.

Arons developed a system known as SpeechSkimmer at the Massachusetts Institute of Technology (Arons 1994b). SpeechSkimmer was designed to explore the level of functionality that could be provided using a strictly auditory display. It incorporated sophisticated compression and abstracting techniques, some of which are already included in the design we have presented. Other techniques from SpeechSkimmer such as abstracting based on pitch-based detection of topical emphasis may be beneficial as future enhancements to our system.

Graphical selection interface that exploit the ranked retrieval paradigm have been implemented by four re-

search groups. The Cambridge University video mail retrieval system displays a ranked list of manually entered metadata associated with the information objects. (Brown *et al.* 1995). Selection interfaces which include automatically extracted information have been implemented at Muscle Fish, at the Swiss Federal Institute of Technology (ETH Zurich), and at Carnegie Mellon University. In the Muscle Fish non-speech audio retrieval interface, numerical values for duration, pitch, loudness, brightness, and other parameters are displayed along with manually annotated metadata (Wold *et al.* 1996). The ETH system uses a similar approach for speech-based retrieval, displaying the recognized phonemes information objects selected by the user (Wechsler & Schaüble 1995). Carnegie Mellon University's News-on-Demand system adopts a different strategy, displaying a small number of keywords that are recognized in the recorded speech (Hauptmann & Witbrock 1997). In future experiments it would be interesting to compare the effectiveness and usability of this approach with the complete display of error-filled recognized that we propose here. Perhaps the most best strategy will turn out to be a combination of the two approaches.

## Conclusion

We have described an ambitious effort that has the potential to significantly advance the state of the art in the design of speech-based information retrieval systems. Usable techniques for each aspect of our design are presently known, and the required technology is rapidly maturing to a point that will permit robust application on widely deployed computing systems. It is clearly the right time to develop such a prototype, demonstrate the potential of our approach, learn which aspects of the interface real users find most helpful, and begin to experiment with advanced techniques such as cross-language speech retrieval. We believe that there are immediate commercial prospects for this technology, and that there is a clear need for it in the digital libraries of the future.

### Acknowledgments

The author would like to thank Wei Ding, Julie Harding, David James, Gary Marchionini, Allan Rough, Tony Tse, Martin Wechsler and Galen Wilkerson for their insightful comments on early versions of the ideas presented here.

## References

Arons, B. 1994a. Efficient listening with two ears: Dichotic time compression and spatialization. In Kramer, G., ed., *Proceeding of the International Conference on Auditory Display*, volume 18, 171–177. Sante Fe Institute.

Arons, B. M. 1994b. *Interactively Skimming Recorded Speech*. Ph.D. Dissertation, MIT. http://barons.www.media.mit.edu/people/barons/.

Brown, M. G.; Foote, J. T.; Jones, G. J. F.; Jones, K. S.; and Young, S. J. 1995. Automatic Content-Based Retrieval of Broadcast News. In *Proceedings of ACM International Conference on Multimedia*, 35–43. San Francisco: ACM.

Frakes, W. B., and Baeza-Yates, R., eds. 1992. *Information Retrieval: Data Structures and Algorithms*. Englewood Cliffs, NJ: Prentice Hall.

Galles, C. F. 1993. Automatic gender recognition from speech. Master's thesis, University of Maryland, College Park.

Hansen, J. H. L., and Arslan, L. M. 1995. Foreign accent classification using source generator based prosodic features. In *1995 International Conference on Acoustics, Speech, and Signal Processing*, volume 1, 836–839. IEEE.

Hauptmann, A. G., and Witbrock, M. 1997. Informedia: News-on-demand – multimedia information acquisition and retrieval. In Maybury, M., ed., *Intelligent Multimedia Information Retrieval*. To appear. http://www.cs.cmu.edu/afs/cs/user/alex/www/.

Kazman, R.; Al-Halimi, R.; Hunt, W.; and Mantei, M. 1996. Four paradigms for indexing video conferences. *IEEE Multimedia* 3(1):63–73. http://www.cgl.uwaterloo.ca/Jabber/ieee-mm4.ps.

Kobla, V.; Doermann, D.; and Rosenfeld, A. 1996. Compressed domain video segmentation. Technical Report CS-TR-3688, University of Maryland, College Park.

Lund, M.; Ma, K.; and Gish, H. 1996. Statistical language identification based on untranscribed training. In *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 2, 793–796. IEEE.

Miller, J. D.; Lee, S.; Uchanski, R. M.; Heidbreder, A. F.; and Richman, B. B. 1996. Creation of Two Children's Speech Databases. In *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 2, 849–852. IEEE.

Ng, K. 1996. Survey of approaches to information retrieval of speech messages. http://sls-www.lcs.mit.edu/~kng/papers/IRsurvey-working.ps.

Resnik, P. 1997. Evaluating multilingual gisting of web pages. In *AAAI Symposium on Cross-Language Text and Speech Retrieval*. American Association for Artificial Intelligence. http://www.clis.umd.edu/dlrg/filter/sss/papers/.

Saunders, J. 1996. Real-time Discrimination of Broadcast Speech/Music. In *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 2, 993–996. IEEE.

Sheridan, P.; Wechsler, M.; and Schäuble, P. 1997. Cross-language speech retrieval: Establishing a baseline performance. In *Proceedings of the 20th International ACM SIGIR Conference on Research and Development in Information Retrieval*. http://www-ir.inf.ethz.ch/Public-Web-Pages/sheridan/.

Soergel, D. 1994. Indexing and retrieval performance: The logical evidence. *Journal of the American Society for Information Science* 45(8):589–599.

Waibel, A. 1996. Interactive translation of conversational speech. *Computer* 29(7):41–48.

Wechsler, M., and Schaüble, P. 1995. Speech retrieval based on automatic indexing. In *Final Workshop on Multimedia Information Retrieval (MIRO '95)*. http://www-ir.inf.ethz.ch/ISIR-Papers.html.

Wilcox, L.; Kimber, D.; and Chen, F. 1994. Audio indexing using speaker identification. In Mammone, R. J., and J. David Murley, J., eds., *Automatic Systems for the Identification and Inspection of Humans–SPIE 1994*, volume 2277, 149–156. SPIE.

Wold, E.; Blum, T.; Keislar, D.; and Wheaton, J. 1996. Content-based classification, search and retrieval of audio. *IEEE Multimedia Magazine*.

Wolf, W. 1996. Key Frame Selection by Motion Analysis. In *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 2, 1228–1231. IEEE.

Zissman, M. A., and Weinstein, C. J. 1990. Automatic talker activity for co-channel talker interference suppression. In *1990 International Conference on Acoustics, Speech, and Signal Processing*, volume 2, 813–816. IEEE.

Zissman, M.; Gleason, T.; Rekart, D.; and Losiewicz, B. 1996. Automatic dialect identification of extemporaneous, conversational, Latin American Spanish speech. In *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 2, 777–780. IEEE.

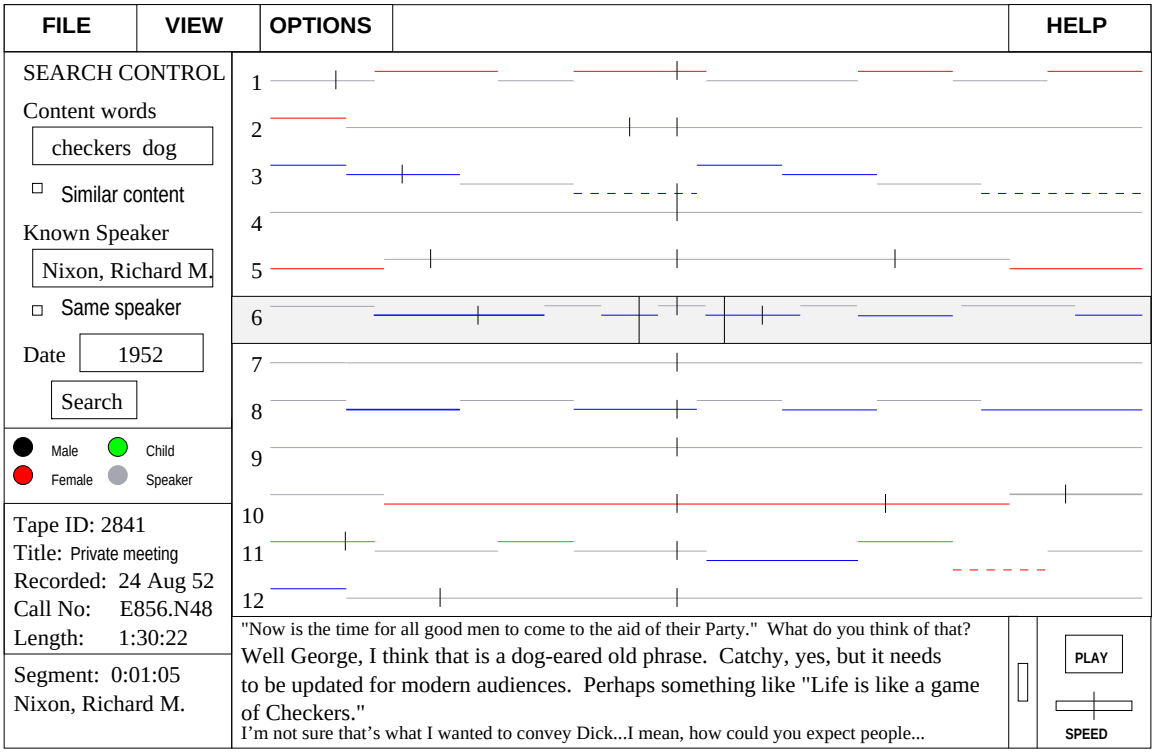


Figure 2: User interface example for recorded speech without video or closed captions.