Access to Recorded Interviews: A Research Agenda

FRANCISKA DE JONG
University of Twente
DOUGLAS W. OARD
University of Maryland
and
WILLEMIJN HEEREN and ROELAND ORDELMAN
University of Twente

Recorded interviews form a rich basis for scholarly inquiry. Examples include oral histories, community memory projects, and interviews conducted for broadcast media. Emerging technologies offer the potential to radically transform the way in which recorded interviews are made accessible, but this vision will demand substantial investments from a broad range of research communities. This article reviews the present state of practice for making recorded interviews available and the state-of-the-art for key component technologies. A large number of important research issues are identified, and from that set of issues, a coherent research agenda is proposed.

Categories and Subject Descriptors: H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing—*Indexing methods, linguistic processing, thesauruses*; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Search process*; I.2.7 [**Artificial Intelligence**]: Natural Language Processing—*Speech recognition and sythesis, text analysis*

General Terms: Algorithms

Additional Key Words and Phrases: Interviews, access technology, spoken word archives, oral history, speech indexing

ACM Reference Format:
de Jong, F., Oard, D. W., Heeren, W., and Ordelman, R. 2008. Access to recorded interviews: A research agenda. ACM J. Comput. Cultur. Heritage 1, 1, Article 3 (June 2008), 27 pages. DOI = 10.1145/ 1367080.1367083 http://doi.acm.org/ 10.1145/1367080.1367083

1. INTRODUCTION

Historical scholarship has traditionally placed greater emphasis on documentary evidence than on memory when seeking to interpret the human experience. Three characteristics of written documents have been important in this regard: they exhibit a degree of immutability that adds some authority

The contribution of D. Oard was supported in part by NSF award IIS-0122466. The contributions of F. de Jong, W. Heeren and R. Ordelman were in part supported by the Dutch bsik-programme MultimediaN (`http://www.multimedian.nl/`), the NWO programme CATCH (`http://www.nwo.nl/catch/`) and the EU IST-FP6 project MESH (`http://www.mesh-ip.eu/`).
Authors' addresses: F. de Jong, Department of Computer Science, Human Media Interaction Group, University of Twente, Enschede, The Netherlands; email: f.m.g.dejong@utwente.nl; D. W. Oard, College of Information Studies and Institute for Advanced Computer Studies, University of Maryland, College Park, MD 20742; W. Heeren and R. Ordelman, Department of Computer Science, Human Media Interaction Group, University of Twente, Enschede, The Netherlands.

to the information they contain, they possess a degree of permanence that brings the information forward in time to where it can be interpreted, and they are accessible because knowledge organization techniques have been developed that make it possible for scholars to obtain the documents they need. For several thousand years, these factors gave primacy to the written word. In recent years, however, new technologies have emerged that are transforming the role of the spoken word in scholarship. The first factor, immutability, was addressed convincingly by the widespread use of audio tape recording in the second half of the twentieth century. The result was explosive growth in the collection of recorded interviews for scholarly, journalistic, and personal reasons. The second, permanence, is now being addressed by digital preservation projects. In this article, we focus on the third great challenge: providing access to the important part of our cultural heritage that recorded interviews provide.

Interviews recorded in the course of inquiry by scholars are often referred to as *oral history*. These interviews form a basis for further research and analysis of individual and collective memories, often of groups whose points of view are less prominent in written history, such as laborers and women. Moreover, recorded sources have an added dimension that may be of interest to scholars of various disciplines, the speakers' voices. These not only convey opinions, but also intonation that can help with interpretation of those opinions and that can sometimes better convey emotion. There are also interviews from other sources that scholars use. Chief among these are interviews conducted for use in radio or television programs, and community memory projects in which individuals contribute their recorded recollections. While distinctions between sources are important to scholars, the technical issues involved with providing access to the content of those interviews are largely common across the genre. In this article, we focus on those technical issues.

Recorded interviews are, of course, just one type of spoken-word cultural heritage collection. An incomplete list of other important genres includes speeches, meetings, courtroom and legislative debates, and telephone calls. Each of these genres raise unique challenges, but they also have much in common. In this article, we focus on recorded interviews, although we note when specific techniques have the potential for broader utility.

Most of the investment to date in supporting access to recorded interviews has been focused on straightforward techniques that typically involve manual segmentation and description. While these techniques are accurate, they are not easily or affordably scaled to very large collections. For example, manual transcription at a cost of perhaps $35 per recorded hour would be an excellent choice for a 100-hour collection, but a 100,000-hour collection (several of which exist) would require very large investments. As a result, few of the collections presently available over the Web make even 100 hours of material available. Fortunately for us, massive investments over the past twenty years in the development of techniques for automated transcription of speech have yielded an excellent starting point for developing a new generation of scalable techniques for access to spoken word collections. That is our central focus.

While automated transcription provides an important core capability, it is just one part of the complex interdependent set of technologies and processes that people will use to discover and make sense of what has been recorded. Expertise from a broad range of disciplines will be needed, including audio engineering, speech processing, information retrieval, natural language processing, library science, human-computer interaction, and image processing. Notably, none of these disciplines can plausibly address these problems in isolation one hallmark of application-centered research is that the application often draws forth challenges that span disciplinary boundaries. We therefore focus on issues that lie at the intersection of research and practice, at the intersection of system design and the processes by which that system will be used, and the intersections between disciplines.

The remainder of this article is organized as follows: Section 2 describes the present state of practice for providing online access to recorded interviews. Section 3 complements this with a state-of-the-art

review for some key component technologies. Section 4 sketches out a research agenda that could transform scholarly access to recorded interviews, drawing together requirements and opportunities to identify specific research issues that merit attention. Section 5 concludes this article with some suggestions for where to put priorities.

2. STATE-OF-THE-PRACTICE

This section describes how interview collections have been acquired and how they are currently being maintained in archives. Section 2.1 treats the technicalities related to the acquisition of interview collections. In Section 2.2, the creation of digital objects is presented, which is highly important for the future preservation of and online access to the collections. Section 2.3 deals with the current practice in documentation, and Section 2.4 discusses issues related to collection access.

2.1 Acquisition

The massive acquisition of recorded interviews by researchers was prompted in part by the introduction of the compact cassette in the 1960's. Cassette tapes (essentially, a stereo analog format on 1/4 inch tape) is therefore now the most common format for legacy materials that were recorded in the field, although the use of video tape with analog audio also became common starting in the 1990's. Reel-to-reel audio recorders generally yielded audio quality superior to that of cassette tape in this time frame so it is common to find legacy studio recordings on either tape reels or on video tape. A relatively small proportion of interviews were recorded in other formats (e.g., wire reels or wax cylinders). Some recordings made in less common formats have great historical significance, however, so a limited number of replay devices for those formats are still maintained. Portable digital voice recorders have recently emerged as the preferred device for recording field interviews, and digital voice recording has been widely used in studio settings for some time.

The use of a single far-field microphone has been the most common practice for field interviews. Often this microphone is desk-mounted, but use of a microphone built into the recording device has also been common. Lapel microphones mounted on a speaker's clothing are commonly used in radio stations, and head-mounted microphones are sometimes used for studio interviews recorded in radio stations. Interviews are sometimes recorded in stereo, often with substantial cross-channel effects due to microphone placement.

2.2 Creation of the Digital Object

Digitization of archives involves digitization of the recordings (2.2.1) and digitization of their descriptions (2.2.2). Both processes are an essential prerequisite to online access and future preservation.

2.2.1 *Digitization of Recordings.* The European IST-FP6 project PRESTO[1] has performed a survey for the content maintained at ten of the larger broadcast archives. According to their estimate, in 2000, there were some 200 million hours of recorded audio in Europe broadcast archives alone, which in general was not yet available even for serious scholarly uses. Of this content, 80% was kept on analog carriers [Wright and Williams 2001]. Recorded interviews were a substantial part of that amount. For future preservation of these materials, digitization seems inevitable as analog media decays with time, and both the media and the equipment needed to reproduce audio from that media are becoming obsolete. Hence, the International Association of Sound and Audiovisual Archives (IASA) provides recommendations for digital formats and storage.

[1]URLs for many projects, standards and examples mentioned here and in subsequent sections can be found in the Appendix.

The number of large-scale digitization initiatives for cultural heritage collections has recently increased substantially. The project IST-FP6 PrestoSpace, for example, aims at the development of facilities for digital preservation of Europe's audiovisual heritage. Large broadcast archives such as the BBC, INA, RAI and Instituto Luce take part. As an example of a national initiative, the massive retrospective digitization at the Institute for Sound and Vision (the Netherlands broadcast archive, also known as: Beeld en Geluid) could be mentioned. It will start in 2007 and will lead to an archive of approximately 600,000 hours of digital video for a wide range of genres and topics, including many subcollections that are relevant from a historical perspective.

2.2.2 *Metadata Standards.* As is common practice in archiving content collections in general, curated digital interview collections are usually enriched with description that help to support interpretation of their contents and context. The term *metadata* is commonly used to refer to the complete set of description pertaining to a information object. Metadata for digital content is typically also stored in digital format both for entire collections and for individual items. The management of and accessibility to collections of recorded interviews calls for metadata standards that support the unique content and contextual characteristics of audio objects and that facilitate syntactic and semantic interoperability.

Best practices regarding the selection of standards for audiovisual collections generally are to some degree still in flux. As a consequence, multiple standards are presently in use. Moreover, standardization efforts typically focus more on what can be encoded than on what must be encoded so selective implementation can also lead so incompatibilities. Recent overviews of current practices in the cultural heritage domain can be generated in the context of IST-FP6 project MultiMatch (Multilingual/Multimedia Access To Cultural Heritage see Oomen and Smulders [2006] and Goldman et al. [2003]) development and use of metadata models is also one of the focuses of the IST-FP6 Coordination Action CHORUS for multimedia content search engines.

Metadata models (i.e., the application of metadata standards to specific cases) that can be effectively deployed for encoding audio content documentation typically follow what has been proven to work well for other content types. For example, distinctions are typically drawn between different types of description such as administrative features and technical characteristics. Of course, metadata fields specific to recorded interviews (e.g., time-coding for specific points in the transcription or the name of the interviewer) are also needed. Metadata models specific to oral history collections have therefore been proposed that associate descriptions with specific points in the information life cycle, for instance, production and digitization. (See Hunter and James [2000]). As interview collections will often be just a subcollection in more heterogeneous archives, interoperability of metadata across format types is also getting a lot of attention as well as compatibility with so-called Semantic Web standards (e.g., RDF and OWL) as advocated by W3C. The International Federation of Television Archives FIAT/IFTA promotes MPEG-7 as the metadata standard that guides the description of multimedia content, including audio. MPEG-7 uses XML to store metadata. As it allows the incorporation of time codes to tag particular fragments, the descriptions can be used by all kinds of applications for searching and browsing of recorded interviews.

2.3 Manual Content Description

Generally speaking, oral history collections are often relatively rich in manually created metadata. Automated counterparts to many of these techniques are described in Section 3, but our focus in this section is on present practice and automated techniques for content description are not yet widely deployed. Depending on the metadata model that is employed, manually assigned descriptive metadata typically consists at least of a title, a summary or (remarkably often) a full transcript, and background

information on the speakers. In addition, sound archives usually also store technical and administrative metadata.

2.3.1 *Transcription and Summarization.* Metadata for oral history interviews often consists of detailed interview transcripts on the one hand, and information on the interviewees and their social environment on the other. The generation of full transcripts is an extremely time-consuming undertaking: it takes around 10 times real time. Some examples of oral history initiatives for which full transcripts are available include the University of Michigan-Dearborn Voice/Vision Holocaust Survivor Archive and the Oral Histories of the American South project. Because of the heavy time investment needed for full description, manually created summaries are often used instead of accurate interview transcripts. Some examples of historic audio collections with relatively rich summaries include the Memories of The East collection [Steijlen 2002], the Kilbirnie-Lyall Bay Community Centre project, and the Oral History of British Photography project.

2.3.2 *Controlled Vocabulary Indexing.* As was common practice for indexing text-based collections in an era before search engines could automatically read the text, indexing untranscribed audio can be accomplished by having indexers manually select appropriate descriptors from a thesaurus. The full set of thesaurus descriptors thus essentially amounts to a shared vocabulary between indexers and users. The use of a controlled vocabulary yields standardized description, which minimizes ambiguity and thus (with suitable training) can increase search accuracy. Moreover, relations between concepts that are encoded in the thesaurus can serve as a basis for navigation, and they can help new users learn more about the structure of knowledge in a domain. Thesauri are available for many domains; an example developed specifically for audiovisual archiving in the Netherlands is the GTAA. Perhaps the largest scale example of thesaurus use for the description of oral history interviews is the collection of more than 50,000 interviews assembled and indexed by the USC Shoah Foundation Institute for Visual History and Education, [Gustman et al. 2002].

2.3.3 *Alignment and Topic Segmentation.* Oral history interviews can be quite long (sometimes a dozen hours over multiple sessions), covering many topics. For some ways in which interviews might ultimately be used, it is therefore useful to either subdivide an interview into several segments, or (more generally) to index the interview content in a manner that aligns the indexed content (e.g., the transcript) with the time(s) in the interview at which that content is present. Without this kind of segmentation or alignment, the time-consuming shuttle search process of moving backwards and forwards in a tape to look for a particular point would be needed before the audio could be heard. Content segmentation to support fully automated search of transcripts can be done automatically, but this approach typically yields replay start points that are poorly aligned with topic changes and thus poorly suited to interactive applications. An alternative is to manually segment the transcripts into topically coherent segments. Although manual content segmentation yields an improved user experience, it can be a relatively expensive process. It is therefore most often employed in relatively small-scale applications such as when samples from a larger collection are posted on the Web (in this case, segmentation also helps limit transfer times for users with slow Web connections). Another challenge with topic segmentation is that additional segment-scale metadata is required if metadata-based access to individual segments is to be supported. One good example of the potential of these this approach is the Voices of the Colorado Plateau project Web site in which a list of topics leads to segments from different interviews in which that topic was addressed.

2.3.4 *Linking.* While metadata is undeniably useful, it is also relatively expensive. One way to reduce that cost is to include provisions in the metadata standard for linking to existing resources that have descriptive value. For example, interviews can be linked to one or more photos of the interviewees

as is done in the Community Memory of Veterans History project. Alternatively, photos can be used as access points to reach the audio, as in the German Migration Audio Archiv project. Geographic sources such as maps can also be linked with interviews as in the Voice/Vision Holocaust Survivor Archive and in the Stories of the Dreaming Web site. More generally, there are a wide range of Web sites that integrate information from multiple sources about events or periods in the history of the 20th century. While Wikipedia may be the most obvious of these, opportunities for this kind of linking abound. For example, links that guide users from interviews (or interview segments) to digitized newspaper articles from that time period can be provided.

2.4 Access Support Methods

This section discusses the access modes that are currently offered to the users of interview collections. Digital archives can support online access, while nondigitized collections must rely on other procedures.

2.4.1 *Content Delivery.* Many archives now support online access to their catalogs. But while metadata descriptions can often be searched remotely, access to the actual collection content is often less easy. Since the vast majority of sound archives have not yet been digitized, users have to actually visit the archive and request the tapes in which they are interested. Even fully digitized collections may not be accessible online because (1) streaming audio transfer imposes greater demands on technical infrastructure than Web-based access to text, and/or (2) copyright and/or privacy issues may preclude widespread public access. Nevertheless, the number of sound archives that provide substantial online content is growing as these challenges are addressed.

As a compromise between no access and full access to audio collections, audio excerpts are in some cases given as an illustration of the collection's content. Examples of this approach include the collections of the USC Shoah Foundation Institute for Visual History and Education and the Australian ABC Archives. In the case where digital audio can be listened to via a Web server, the most basic presentation form is one in which the audio is given with regular playback options (play-pause, etc.) that allow the listener to navigate through the data (e.g., the StoryCorps Community Memory project). Usually, some metadata is supplied, specifying the title and the duration of the segment or interview. Rich description and richer control over the replay than the simple play/pause/fast-forward controls (e.g., a clickable replay timeline) can help users find what they wish to listen to more rapidly. This can be far more important for audio than it would be for text. The time investment for browsing an audio segment can be substantial so substituting metadata-based navigation for direct examination of content to the extent possible is typically beneficial.

2.4.2 *Visualization.* At present, search results most often present the associated metadata as text, although photographs, timelines, and geographic visualizations can also be provided. The prominent role of relationships between individuals in the content of oral history collections also suggests other opportunities. For example, in the Digital Monument to the Jewish Community in the Netherlands, over 8,000 blocks link to the record of individual families. Although that project does not involve audio materials, a similar organizing principle could be used with oral history collections in which members of the same family were either interviewed or mentioned.

2.4.3 *Access Technology Projects.* As acquisition and storage costs have declined over the past several decades, it has become increasingly apparent that fully manual techniques for providing access to oral history collections can affordably meet just a small fraction of the potential need. In 2002, the DELOS/NSF working group in Spoken Word Audio Collections met to discuss an agenda in the area of spoken word archives and collaborative research projects [Goldman et al. 2005]. A number of items from that agenda have since been taken up in research projects.

One initiative is the MALACH (Multilingual Access to Large spoken ArCHives) project, a US NSF project (2001–2007), that investigated access to a vast collection of testimonies from Holocaust survivors, witnesses and rescuers [Oard et al. 2002]. The goal of that project is to advance automatic speech recognition (ASR) for the oral history domain and to study how recognition can be best incorporated in further processing and retrieval steps [Gustman et al. 2002]. Another project that did, however, not specifically aim at access to interview collections but that contributed to advancing spoken document retrieval in the cultural heritage domain was the National Gallery of the Spoken Word project. In that project, the SpeechFind spoken-document retrieval system was developed. It automatically generates metadata for audio documents by segmenting the audio and generating ASR transcripts and also makes the audio searchable though a Web interface [Hansen et al. 2005].

There are a number of initiatives in spoken-document retrieval in the cultural heritage domain for languages other than English. For example, the IST-FP5 project ECHO (European CHronicles Online) aimed at the realization of a searchable multilingual collection of video documentaries deploying speech recognition as one of the core technologies. CHoral [Ordelman et al. 2006] is one of the ten projects within the Dutch CATCH program (Continuous Access To Cultural Heritage), funded by the Dutch Research Council (NWO), in which computer science groups and cultural heritage institutes work together to develop access technologies for cultural heritage. CHoral aims at the development of audio indexing technology. In addition to advances in speech recognition for spontaneous Dutch speech, the extraction and usage of additional features that can be derived from the audio, such as speaker and bandwidth information, are being investigated.

Interview collections that are fully transcribed allow fine-grained, automatic access with relatively little effort. Through alignment of the audio and the corresponding text, a word-level index can be generated that is useful for online search, access and browsing as illustrated in the Radio Oranje search engine which gives access to historical speeches [Heeren et al. 2007]. Another approach was taken by Klemmer et al. [2003] who added barcodes to paper transcripts to create a way of linking interview transcripts to direct video access on a PDA.

However, these projects have yet to achieve broad impact in practice, precisely because their focus was on research. If we are to leverage these and similar investments to the benefit of practice, we must foster a stronger connection between the state-of-the-art in practice and the state-of-the-art in the development of potentially useful technical capabilities. In many cases, these techniques have been developed for other applications and have not yet been applied to oral histories. That is where we next turn our attention.

3. STATE-OF-THE-TECHNIQUE

There is a considerable range of tools and technologies available that could be deployed to improve access to spoken word collections [Goldman et al. 2005]. Besides the substantial effort that has been devoted to development of technology for content-based access to audiovisual archives in general, several initiatives have successfully been undertaken to apply some of this technology to cultural heritage collections. In this section, we review the state-of-the-art with regard to key technologies that can complement the manual content descriptions of Section 2.3. We start with an overview of spoken-word audio analysis technology in Section 3.1. The direct results of audio analysis are relatively low-level features that, when coupled with other forms of semantic annotations, can be used to generate more high-level representations that are suitable for direct use in information access applications. Technology aimed at the generation of these high-level representations is discussed in Section 3.2. This is followed by an overview of techniques for supporting interactive access in Section 3.3 that can enhance the techniques described in Section 2.4.

3.1 Audio Analysis

Techniques for extraction of information from the audio signal can broadly be divided into those focusing on speech-related features and those focusing on sound in general. The latter includes technology for music retrieval and sound classification. Audio analysis technology with a focus on speech is particularly relevant to cultural heritage collections, although there are cases in which more general audio processing can be important (e.g., for distinguishing between speech and non-speech segments such as music, jingles, and other sources of noise in order to limit indexing to the spoken content).

Thanks to impressive advances in speech recognition technology made in the last decades (Section 3.1.2) and the proven effectiveness of this technology for indexing some types of spoken word collections (Section 3.1.1), speech recognition is emerging as a key technology for supporting information access. In this context, it can be used either as a replacement for manually created metadata or as an enhancement when manually created metadata does exist. The results of automated transcription can be searched directly, and transcripts can also serve as a basis for many NLP techniques that can help to characterize the semantic content of speech. In the context of the kind of access technology that is the focus of this paper, the role of NLP coincides more or less with automated metadata generation. In this section, we discuss the state-of-the-art in search for spoken word content as well as the techniques that enable this functionality: the automated creation and enhancement of speech transcripts.

3.1.1 *Spoken Document Retrieval.* In this section, we describe the application of ASR techniques to support automated indexing of spoken word collections.

Research on searching spoken word collections using automated transcription dates to 1997 with the inception of the Spoken Document Retrieval track at the Text Retrieval Conference (TREC). Over the four years of that evaluation, it was shown that automatically transcribed English news broadcasts could be reliably searched using natural language queries [Garofolo et al. 2000]. The Topic Detection and Tracking (TDT) evaluation conducted additional annual evaluations of broadcast news retrieval between 1998 and 2004 that yielded similar results for Arabic and Chinese and that demonstrated effective techniques for automatically dividing news broadcasts into individual stories [Allan 2002].

The TRECVID video retrieval benchmark is in some sense a successor to the TREC SDR track. The TRECVID test collections contain not just video, but also ASR-generated transcripts of segments that contain speech. TRECVID results for search tasks confirm that exploitation of spoken content for generation of a time-coded index for audiovisual content can help to bridge the semantic gap between image-derived features and search needs. Systems that exploit these transcripts typically achieve better retrieval effectiveness measures than systems that do not incorporate speech-derived features [Smeaton et al. 2006].

The National Gallery of the Spoken Word, an NSF project (1998–2005), was the first to explore the utility of automated transcription for searching historical spoken-word collections (e.g., speeches and interviews). Results from that project indicate that the acoustic conditions in which a recording was made can have a far greater effect on transcription accuracy than the age of the materials (and hence, presumably, the original medium) [Hansen et al. 2005].

In the first years of this millennium, the EU-funded ECHO project investigated the value of automatic speech transcription as a basis for automatic metadata extraction for a multilingual collection of documentaries for which there was historical interest. The diversity in the collection was large both with respect to the topics and the recording dates (spanning half a century from the early years of film onwards). Results indicated that integration of manual indexing with automated indexing is feasible but that the diversity within the collection made it impractical to produce sufficiently accurate ASR output for the entire collection. The main bottlenecks were relatively poor audio quality, old-fashioned vocabulary and speaking styles, and a lack of representative training materials, which generally were

not available in digital form. An important finding of the project was that the availability of these types of collateral sources of training data should be explored as a way of improving ASR performance when planning cultural heritage projects. Scanning and OCR is increasingly being applied to historically significant text, so the the amount of available and representative text may well grow in the future, which would help with tuning vocabularies and language models.

3.1.2 *Large Vocabulary Continuous Speech Recognition* . In this section, we outline a number of issues and techniques for the automated creation and enhancement of speech transcripts using ASR.

3.1.2.1 *ASR performance.* Substantial investments have been made in automatic transcription of spontaneous conversational speech since 1992 [Godfrey et al. 1992]. This has resulted in progressively more accurate transcriptions. Word error rates for ASR on conversational speech are, however, still substantially higher than those for ASR on the planned speech that is common in broadcast news collections. For state-of-the-art broadcast news transcription, word error rates below 10% have been reported for widely studied languages such as English; for spontaneous speech, word error rates between 40% and 60% are by no means exceptional. (See Huijbregts et al. [2007] for recent ASR performance figures on a subcollection of the TRECVID dataset with surprise data.) Variations across recordings are, however, often far greater than variations across words. An important question is therefore what fraction of the content can be processed well enough to support specific tasks. While experiment results for the TREC SDR task have proven that usable retrieval is feasible even with word error rates as high as 50%, transcription errors severely complicate the other types of automated analysis. Examples of tasks that become more difficult are automated summarization and the detection of entities, relations, and events [Jing et al. 2007]. Optimization strategies typically focus on improving the recognition of named entities, using either domain-specific techniques (e.g., vocabulary priming) or more general approaches. Among other things, this may involve language model tuning.

MALACH, a US NSF project (2001–2007), was the first to use ASR technology for searching automatically transcribed interviews [Byrne et al. 2004]. Among the key findings from that project are: (1) an automated transcription tuned to the collection can yield adequate accuracy to support searches based on natural language queries and (2) controlled vocabulary indexing by subject matter experts still yields better retrieval effectiveness, in part because less commonly spoken names are often mistranscribed and because dates are rarely spoken (but are readily inferred by trained indexers). These results were obtained with transcription systems that were tuned to the specific collection, however, and tuning to a new collection can be expensive. The best estimates from the MALACH project are that this approach would presently be cost effective only for collections larger than about 1,000 hours.

Transcription of speech found in natural environments, where the speech is typically spontaneous and often conversational, has been the focus of research for quite some time. Automatic annotation of meeting recordings has received a lot of attention recently, stimulated in part by yearly NIST Rich Transcription evaluations. In addition to measuring transcription accuracy, these evaluations have also evaluated specific preprocessing steps, such as audio segmentation, that have implications for both ASR system design and for direct use in systems that are designed to support information access. The IST-FP6 AMI project (Augmented Multi-party Interaction, 2004–2006) funded much of the ASR research in the meeting domain. Because meeting recordings have much in common with interviews and, in particular, because they vary in similar ways with regard to acoustic environment, recording conditions and content, work on meetings is highly relevant to the design of systems to improve access to oral history.

In determining the accuracy requirements for ASR, it is clearly important to have specific tasks in mind when asking what constitutes results that are good enough. Supervised machine learning techniques for topic segmentation, for example, place a greater premium on consistency than on raw

accuracy, while bag-of-words retrieval techniques are robust in the presence of occasional errors. Extractive summarization, by contrast, requires that consecutive words be correctly recognized (so higher error rates may yield shorter and less informative snippets), and more sophisticated analysis (e.g., the entity tagging used in question answering systems) may be even more sensitive to recognition errors.

In the remainder of this section, some techniques to overcome, or at least to limit, the effect of poor ASR accuracy are discussed.

3.1.2.2 *Transcription enhancement: Alignment.* In many cases, closely associated collateral text will be available along with a spoken word collection. Typical examples are subtitling, teleprompter texts in news shows, minutes of meetings, written versions of lectures or speeches, and interview notes. Time synchronization of these resources with the audio track can be worthwhile since it can serve as an affordable source of searchable annotations. Digital audio can be dynamically aligned with the words in the text using speech recognition tools that provide accurate time labels on the basis of an acoustic model and the words in the text. Due to the low complexity of the task, alignment is relatively robust to mismatches in acoustic conditions so the careful tuning that is needed for full speech recognition can often be avoided [Brown et al. 1995]. When single-pass matching proves to be inadequate (e.g., because of gaps or phrase compression), a two-stage process can be used in which a rough alignment is performed first to identify reliable anchor points, and then by a word-level alignment within the resulting segments follows [de Jong et al. 2006; Moreno et al. 1998].

3.1.2.3 *Vocabulary optimization.* To adequately support spoken document retrieval, it is crucial that the speech recognition system can adapt to differences in word choice across the collection in order to reduce the number of out-of-vocabulary (OOV) words (words unknown to the system that therefore cannot ever be recognized). In particular, in SDR, it is important to reduce the number of OOV query (QOV) words: words that occur both in a user's query and in the spoken content but that could not be recognized correctly because they were OOV. OOV words impede access to spoken content in three ways: (1) given a query with a QOV word, the QOV word leads to a word miss in searching; (2) the (necessarily incorrect) replacement hypothesized by the ASR system can potentially induce a false hit for other queries, and (3) when speech transcripts are used as a basis for further processing stages, OOVs may introduce unrecoverable errors.

Several solutions to this problem have been proposed, including use of larger recognition vocabularies, dynamic adaptation of vocabularies based on collateral materials or metadata, and asking interviewees to provide a list of salient terms before or after their interview [Rosenfeld 1995; Auzanne et al. 2000; Allauzen and Gauvain 2005]. Another strategy for dealing with QOV words is to avoid speech-recognition vocabulary restrictions by creating audio representations based on subword units (e.g., phonemes) rather than complete words [Ng 2000; Smeaton et al. 1998]. Subword approaches introduce other types of errors, however, since vocabulary knowledge adds a useful source of constraints for the speech recognition task. For this reason, there is now considerable interest in exploring ways of combining word-based and subword techniques.

3.1.2.4 *Multipass approaches.* Speech recognition systems applied to difficult tasks often use multiple passes in order to flexibly incorporate specialized technology or knowledge sources that can be used to tune system parameters optimally to the needs of a processing stage. Typical examples are adaptation of acoustic models to specific speakers (e.g., using vocal tract length normalization [Welling et al. 1999], heteroscedastic linear discriminant analysis [Burget 2005], maximum likelihood linear regression [Legetter and Woodland 1995]), or high-order language models (e.g., word 4-grams). Word lattices that encode the most likely word sequences are often used as an intermediate data structure between system stages. Other multipass approaches first do an acoustic-phonemic search to generate a

dense phoneme network followed by one or more passes in which specialized decoding steps are applied (e.g., using morphological and/or domain knowledge).

3.1.2.5 *Speaker-dependent speech recognition.* Speech recognition technology can also be used indirectly in the annotation process by using a dictation process (restatement of the content) together with a speaker-dependent speech recognition system. After an enrollment session in which the dictator typically reads aloud predefined texts, system parameters that are tuned to that specific speaker are computed. Next, dictation is performed by listening to the spoken content and restating it. This process is sometimes used by court reporters, for example, because a usable transcript can be generated more quickly than from a stenographic transcription. In a related approach, speaker-dependent dictation systems could also be used when new interview collections are created. In such an approach, both the interviewer and the interviewee would perform enrollment to tune the system to their voices before the actual interview takes place. One risk with such an approach, however, is that spontaneous conversational speech differs substantially from read speech, so improvements resulting from speaker-dependent system tuning may be less likely from conversational speech than would be the case for the more highly structured dictation task.

3.1.3 *Audio Partitioning and Classification.* For speech recognition purposes, an audio stream is usually partitioned into segments that are categorized according to a set of acoustic classes that may be fairly broad (e.g., speech, non-speech and music) or more fine-grained (e.g., channel characteristics, speaker gender, or even speaker identity). These segment classifications can subsequently be used to select the most appropriate speech recognition configuration. Audio partitioning and classification is also a prerequisite when non-speech content should also be indexed.

3.1.3.1 *Speaker diarization.* Automatic detection and tracking of individual speakers within an audio recording and/or between different recordings can yield useful metadata that would be difficult to produce manually. The labeling or categorization of audio sources within a spoken document is referred to broadly as *diarization*, but the term is often used as a synonym for speaker diarization at the NIST Rich Transcription evaluations, which incorporated this 'Who Spoke When?' task starting in 2002. Although accurately naming individual speakers in a large set of candidates (where training data may be scarce) in very large collections remains an active topic of research, techniques for the somewhat simpler task of determining whether speakers at different times are the same person are now probably sufficiently accurate and robust to support some types of speaker-based information access.

3.1.3.2 *Postprocessing transcription.* The raw output of current speech recognition systems consists of a stream of words without punctuation or formatting. This sort of structural information is, however, important both for human readability and for NLP techniques that assume fluent, well-formatted input (e.g., translation and summarization). Substantial research effort has therefore been devoted to the enrichment of speech recognition output for which the NIST Rich Transcription Metadata Extraction (MDE) task serves as the principal evaluation venue. Specific tasks in that evaluation include identification of sentence boundaries and identification of specific types of speech disfluencies (repairs, restarts, and fillers).

3.2 Automated Metadata Generation

ASR transcripts can be indexed directly to support search, but they can also be the basis for several kinds of Natural Language Processing that are suited for the capturing of the semantic layers in speech. This analysis can be thought of as a form of automated metadata generation (although, as we have just seen, the term metadata is also sometimes used to refer to some aspects of the speech technology). Metadata generation (used in the sense we mean here) yields a mark-up of the transcribed text

indicating which predefined semantic units have been detected (a classification process often referred to as *tagging* or *extraction*). These more atomic pieces of information can then be searched for (e.g., using a relational database) or they can be used to generate structured summaries. This has been done for a variety of element types, including names (e.g., for people, organizations, and locations), times, and (at a higher aggregation level) events. NLP techniques can also be used to link transcribed text to semantic units or concepts from a domain-specific ontology or a thesaurus to support search and navigation. For some application areas, metadata models have been proposed that focus on the capturing of such automatically extracted content descriptors. An example is the event table defined by NIST in the context of the benchmark for Automatic Content Extraction (ACE) from text.

ASR and many modern NLP techniques have one important thing in common; both rely extensively on statistical techniques. This has two important consequences when constructing an ASR-to-NLP cascade: (1) there can be a benefit to using rich data structures (e.g., word lattices) that encode uncertainty to connect the components, and (2) the NLP results (and hence the resulting metadata) will themselves naturally be subject to some uncertainty. The first of these problems can be accommodated by adopting a unified representation framework for all components in the cascade. Several such processing environments have been proposed over the years, at present, IBM's open-source Unstructured Information Management Architecture (UIMA) seems to be the most popular. Representation of uncertainty is relatively rare in present metadata standards (for the simple reason that most such standardization efforts probably implicitly assumed human-produced metadata) although MPEG-7 does include provisions for storing lattices so, at least for the near term, it would be advantageous to integrate the ultimate information access application into the same unified environment as the processing cascade.

3.2.1 *Topic Segmentation.* Some of the same features used as a basis for topic segmentation in news broadcasts also seem to be useful for interview collections, although not always in the same ways. For example, silence duration, which is positively correlated with story boundaries in news, turns out to be negatively correlated with topic changes in recorded interviews. Speaking rate, by contrast, seems to be positively correlated with topic changes in recorded interviews [Byrne et al. 2004]. Experience from the MALACH project also indicates that the optimal granularity for browsing search results changes when conditioned on a query (generally becoming shorter), and that it may also vary with the ultimate use that is intended. Hierarchical topic segmentation models may therefore be more appropriate than a a strict partition as a basis for indexing. Many approaches to the evaluation of topic segmentation accuracy assume a fixed partition (e.g., for story segmentation in broadcast news in the TDT evaluations, and for passage retrieval in the TREC HARD track), but the XML element retrieval task in recent INEX evaluations may offer some insight into the design of an appropriate evaluation framework for this task.

We also have some evidence that certain types of interviewer questions may be useful as a basis for topical segmentation of conversational speech. Other question types seem to be useful as a basis for browsing interviews in a question-answer format similar to a frequently asked questions list. These results are, however, based on a single relatively small study [Zhang and Soergel 2006] so more work is needed.

3.2.2 *Topic Classification and Summarization.* Automatically assigned topic labels can enhance access of spoken content in two ways. First, as with any content, the labels can provide indexing vocabulary that extends beyond what was actually said. For example, people talking about a specific event may not mention the date or location, but if automatic topic classification is used to recognize the event, then a known time and/or location for that event can be added to the index. Second, and more specific to speech, at higher word error rates, topic labels may provide more readable (although perhaps less rich) summaries than extractive summarization.

So-called extractive summaries can help users to recognize relevant interviews and/or interview segments. In contrast to text summarization, speech summarization on the basis of ASR output has to cope with a number of additional issues: sentence boundaries in speech are not clear, sentences may be incomplete, speech disfluent, and recognition transcripts far from error-free, especially in conversational speech, [McKeown et al. 2005]. Extractive summarization for ASR output has been widely applied to broadcast news (e.g., [Hori and Furui 2003]). More complex 'generative' summarization techniques, which rely on complex NLP to fill slots in a story template from which a fluent summary is then generated, are harder to use with spoken content because ASR errors have a cascading effect on the subsequent NLP.

Machine learning techniques that were originally developed for topic classification with text sources are fairly mature, and the most widely used techniques (e.g., k-nearest-neighbors, and support vector machines) are fairly robust to at least modest ASR error rates. Spoken content offers additional opportunities, however. For example, the relative position of a segment within an interview can be used to bias a classifier in ways that improve overall accuracy (e.g., in many oral history collections, people would be more likely to describe events early in their life in the early part of an interview). The sequential nature of storytelling can also be exploited by drawing evidence form earlier segments to improve classification accuracy. Used together, such techniques can yield better classification accuracy than more general topic classification techniques that are insensitive to presentation order [Olsson and Oard 2007].

Topic classification is not without substantial cost, however. If the classification is to be meaningful, development of a suitable concept hierarchy or ontology will be required. And if the classifier is to be trained automatically (which is in almost every case the most accurate approach), representative hand-labeled training data will also be needed. When suitable training data can be constructed from the available metadata (perhaps with written rather than spoken text), construction of a usable classification system can be relatively straightforward, although at present little is known about how best to tune such a system to spoken content when written content was used for training. If both the ontology and the training data must be developed from scratch, however, topic classification could ultimately be more expensive than manual transcription at least for relatively small collections.

3.2.3 *Speech Indexing.* Indexing methods for spoken audio based on automatically extracted transcripts and annotations typically assume that longer items are split up into segments that the retrieval system can treat as documents. The annotations and/or transcripts are treated as any other bag-of-words to be indexed. Simple document expansion and query expansion techniques can be used to compensate for undesirably large numbers of likely query terms that cannot be recognized in advance due to OOV effects [Woodland et al. 2000; Jourlin et al. 1999]. While this typically yields improvements when averaged over many topics, the effect on individual topics can be adverse. Moreover, user studies have repeatedly revealed that in interactive applications real users would often willingly trade some potential of retrieval effectiveness if doing so would lead to more understandable and predictable system behavior [Zhang et al. 2007]. These points suggest that expansion-based techniques may well be useful in some cases but that, for interactive applications, some usability evaluation would be wise before committing to their use.

Since the unstructured spoken content can be annotated at many levels (e.g., structural metadata, transcription, topic labels, and speaker features), using an XML-based approach to storage and retrieval can also be attractive [Blok et al. 2006]. Different techniques will, however, sometimes yield overlapping annotations, and the strictly hierarchical structure of XML does not handle such cases well. For example, a topic change may occur in the middle of a single speaker utterance, and a speaker change may occur in the middle of a topic statement. Simple cases like these can be accommodated by imposing a

hierarchy, but real situations can be considerably more complex (e.g., topic segmentation systems may hypothesize a topic shift in the middle of the utterance of a multiword expression that is hypothesized by named-entity recognition as a proper name). So-called standoff annotation techniques (such as those implemented in UIMA) can easily accommodate such cases, however.

3.3 Support for Interaction and Access

It is, of course, also important to help users actually access content that has been identified as potentially relevant by some search process.

3.3.1 *Visualization for Relevance Assessment.* Approaches to content representation that support effective browsing are essential to help users evaluate search results. Contrary to browsing text results, leafing through audio documents is not an option. Listening to all fragments in a result list is simply too time-consuming. In the simplest case, for retrieval based on ASR transcripts, we can simply treat those transcripts as text and form snippets around the query terms in the usual way. At low word-error rates this can work well, but, as the error rate increases, the snippets must become shorter if excessive errors are to be avoided, and shorter segments provide fewer contextual cues.

Indeed, because it can take a considerable amount of time to play even a few audio segments, many operational systems that provide access to audio use manually created metadata to provide a second extended description page that users can skim before selecting audio for replay. User studies have shown that, as with other genres, topicality is the dominant criterion that users consider when selecting spoken word content for examination (e.g., Kim et al. [2003]). This suggests the utility of techniques such as clustering, classification, summarization, extraction (e.g., for dates and proper names) and timelines, all of which can convey cues about topicality [Morang et al. 2005]. Replay duration is also an important factor that can be included in the initial hit list and/or the extended description.

3.3.2 *Efficient Audio Replay.* A study with broadcast news content found that visually depicting information about audio content helped users to find facts and judge relevance [Whittaker et al. 1999]. Moreover, when compared to a case in which users could only listen to audio using a standard player, the addition of visual information reduced the amount of audio that was played in the browsing process and also reduced the overall search time. Another way of supporting faster browsing is to use pitch-controlled time-compressed speech to allow users to speed up audio playback. Techniques are now available which are, in some applications, able to support 3:1 speedup without a substantial loss of intelligibility [Hürst et al. 2004]. Fatigue can become a concern when using accelerated replay over extended periods, however.

Interface affordances for the display of speech transcripts include a bouncing ball metaphor (in which the highlighting moves to the present word as the speech occurs), a vertically scrolling marquee (usually without word highlighting, as in closed captioning of video), and a horizontally scrolling single-line ticker (named for its similarity to ticker tape).

3.4 Cross-Media Access

Microphones without cameras are becoming less common, and recording video along with audio can fundamentally change the nature of what is created. Video, focusing on the speaker(s), may add facial expressions, gestures, movement and gaze information, to name a few. In professionally produced programs such as documentaries an even wider range of uses for video are possible (e.g., illustrating what is being talked about). This potential was extensively explored in the Informedia project (e.g., Christel and Yan [2007]). Among the most interesting conclusions from that project are that a unified index can be built from multiple sources of evidence (e.g., OCR of on-screen captions can be used to train face recognition and speaker identification systems, which can then often identify the same person

even when captions are not present) and that key frame images can provide a very useful basis for recognizing relevant segments in documentary video. Key frames will naturally be less useful with just head-and-shoulders video of a speaker, although there are exceptions (such as when an interviewee displays an artifact for the camera).

3.5 Evaluation

As with many disciplines that are fundamentally empirical, advances in speech retrieval depend critically on systematic evaluation. In information retrieval, a distinction is typically drawn between batch evaluation using test collections (which emphasize affordability and repeatability) and user studies (which emphasize fidelity and contextual factors).

3.5.1 *Batch Evaluation Using Test Collections.* Reusable test collections permit rapid iterative exploration of alternative system designs, an approach generally referred to as an evaluation-driven research paradigm. In essence, test collections model some (but not all) salient aspects of a task. The NIST Rich Transcription, Text Retrieval Conference, Spoken Term Detection, Topic Detection and Tracking, and Automatic Content Extraction evaluations all fit into this framework, as do similar evaluation venues in other places (e.g., the Cross-Language Evaluation Forum (CLEF) in Europe).

3.5.2 *User Studies.* It is important to recognize that there are two types of users for curated collections, the curators themselves and the ultimate users of the content. A survey among archivists administering the Dutch National Broadcast Archive, for example, found that imperfect automated transcription raised serious questions about the usefulness of automatically generated metadata. Adequately addressing such concerns will surely be an important factor in the technology adoption process.

For work with end users, two basic types of user studies are possible. Often initial studies focus on learning how people actually perform their tasks. Of course, the existence of some kind of process for performing that task is a prerequisite to performing such a study, and studies of this sort are naturally more useful for elucidating how techniques that already exist will be employed than they are for envisioning the potential use of technologies that do not yet exist. Nonetheless, we must start somewhere. Once the task is reasonably well understood, then new technologies can be created and (after initial batch evaluations) tried out with actual users using controlled study designs and quantitative measures. Measuring the utility of some new technique in such a setting can help to steer further developments but returning to qualitative studies to see how the new techniques will actually be used in less structure settings is important as well. One way to think of this process is as an iterative multidimensional in-depth long-term case study (MILCS) [Shneiderman and Plaisant 2007].

4. A RESEARCH AGENDA

In this section, we draw on the state-of-the-art previously described to identify research challenges that have not yet been met.

4.1 Acquisition and Digital Object Creation

A survey of British oral history in 2000 identified 44 institutionally managed collections in that country alone and that does not include the countless interviews languishing on cassette tapes in the bottom drawers of individual scholars [Ulargiu 2000]. Very little of this diversity has been explored in the NGSW, MALACH, or CHoral projects, and we cannot even hope to characterize the consequences of that diversity until we begin assembling research collections. Moreover, no materials, not even those being collected today, have been created in ways that intentionally optimize the opportunities for subsequent automated processing. We therefore start by looking at how we can work with creators of recorded interviews to best serve their needs by helping them to best serve ours. Most of the issues addressed

are not particularly challenging technically; most urgent are initial laboratory experiments, followed by field studies and then canonization/promulgation of the resulting best practices.

4.1.1 *Recording Conditions.* The textbooks from which practitioners of oral history learn include detailed treatments of issues such as microphone placement, but the practices that they advocate were originally designed to maximize intelligibility to the human ear. Unfortunately, automated processing places greater demands on signal quality and signal conditioning. At least three issues can be identified that could benefit from special accommodations: (1) near-field (e.g., head-mounted) microphones yield substantially greater transcription accuracy (e.g, because of reduced reverberation effects), (2) stereo channels associated with each speaker (interviewer and interviewee) and minimized cross-channel acoustic pickup can reduce the difficulty of speaker segmentation, and (3) a bookmark function built into recording devices would allow the interviewer to unobtrusively designate points of interest that could be exploited during alignment.

4.1.2 *Speaker Adaptation.* One widely used technique for improving speech recognition accuracy in systems designed for personal dictation is to enroll the speaker (i.e., to adapt the acoustic model) by having them read from some prepared text. Read speech, however, bears little resemblance to the spontaneous conversational speech that is commonly found in interviews. So while read speech may be worth exploring as a basis for interview-specific (and, indeed, channel-specific) speech recognition, alternative approaches to speaker enrollment also need to be explored. One possibility is to have the interviewer transcribe a brief, representative portion of the interview. Known-speaker evaluations of speech recognition accuracy typically yield better transcription accuracy than transcription of previously unseen speakers so this might be a productive line of research using actual interviews.

4.1.3 *Vocabulary Adaptation.* One of the most serious limitations of present speech recognition systems is the undesirably large number of likely query terms that cannot be recognized in advance for the simple reason that they are not known to the system. These terms are in the long tail of the distribution of language use, but they are good query terms for exactly that reason. While automated techniques such as phonetic wordspotting can partially mitigate this effect, some of the problems might also be quite easy to overcome at query time. All that would be needed is for the interviewer to jot down any unusual terms that they hear as the interview progresses and then check with the interviewee afterwards to verify the spelling and meaning of those terms. This could then be entered by the interviewer as interview-specific metadata to prime the vocabulary of an ASR system.

4.1.4 *Workflow Management.* Acquisition and ingestion of additional text data sets that can be used for interview-specific optimization of speech recognition models and query expansion is technically feasible (see Section 3.1.2.3), but it complicates the workflow. The same holds for linking interviews to related images. Format conversion and standardization issues need to be settled to support integration of related content. Workflow issues also influence the optimal division of labor between curators of interview collections and the technical staff with expertise in speech processing. The key to addressing these needs is to do locally what is best done locally (e.g., digitization and creation of metadata records), and do centrally what is best done centrally (at present, the types of automated processing that can benefit from collection-scale statistics). Combining decentralized deposits with centralized processing requires both standards development and a process for deploying and supporting software.

4.2 Representing Spoken Content

Much of the research on ASR to date has focused on improving transcription accuracy, but, in practical applications, we must focus in a balanced way on the three As: automation, accuracy and affordability. Most interview collections would seem to be what is commonly called surprise data (i.e., recordings

made under unpredictable conditions that use unanticipated vocabulary for which only small sets of training data are available), to speech researchers.

4.2.1 *Human in the Loop Processing.* Almost all of ASR research has focused on building fully automated systems, but ultimately those system will actually be used by real people. It would therefore behoove us to explore opportunities to integrate ASR-generated indexing and description with both semi-automatic user-assisted annotation approaches and completely manual user-generated content, such as text descriptions posted to blogs or the types of uncontrolled tags that are often used to label content in so-called shared bookmark systems. Ultimately, our goal should be to foster development of a new generation of tools that leverage the participation of domain experts. At the most basic level, interviewers often produce some description as a byproduct of their work. This might range from handwritten notes taken during the interview to detailed analytical work based on the recorded content. If we can capture this byproduct more effectively, we can produce high quality metadata and links to collateral content that could be exploited in systematic ways. At the other end of the spectrum, recent developments for harvesting metadata generated by user communities (what is often now referred to as Web 2.0) should also be explored. Spoken content can be particularly compelling, thus increasing its potential to attract interest from user communities that are large enough to permit substantial metadata harvesting. End-user annotation raises serious concerns about metadata quality, of course, but these concerns are not unique to spoken content and considerable progress is being made in expanding our understanding of these factors (e.g., vocabulary convergence effect and the development of social structures for managing postediting). Innovative workflow designs will be needed to support the integration of automated, semi-automated, and manual annotation, and systems will then be needed that can support continued enrichment of metadata acquired in this way throughout the content life cycle.

4.2.2 *Improving ASR Accuracy.* Access to recorded interviews will undoubtedly benefit from continued investments in automatic transcription of spontaneous conversational speech in meeting collections and in other applications (e.g., telephone calls to help desks). The unique characteristics of recorded interviews will require some specialized investments as well. For example, the acoustic signal from older recordings may be degraded by mechanical characteristics of the recording and replay devices. Present audio restoration techniques model these factors in ways that are designed to reduce adverse effects on human hearing. It seems reasonable to expect that investments in techniques tuned to the requirements of downstream automated processing could yield some further improvements in ASR applications. Interviews are typically much longer than telephone calls (and they typically have fewer speakers than meetings do), thus potentially offering more scope for application of the types of unsupervised speaker adaptation techniques that have been successfully applied elsewhere. When head-and-shoulders video is available (as in the case of videotaped interviews), automated analysis of lip motion could also be used in conjunction with the acoustic channel to improve transcription accuracy.

4.2.3 *Improving ASR Affordability.* Automatic transcription involves both fixed costs that must be incurred regardless of the size of the collection, and variable costs that accrue in a way that depends on the size of the collection. Present ASR techniques require manual transcription of substantial quantities of representative speech, and we do not yet know how best to leverage that investment across diverse collections. Investing in techniques that could be adapted to new collections with a minimum of collection-specific training is therefore important. Even more importantly, automated transcription of spontaneous conversational speech has not yet progressed beyond the research laboratory, which means that training present systems still requires specialized expertise. Ruggedized systems that can be reliably trained for new domains by end users would dramatically reduce fixed costs. Variable (i.e., per interview) costs reflect computation costs, which are already relatively low and still decreasing.

Up to at least 1,000 hours, fixed costs presently dominate. Variable costs can become important for enormous collections such as the British Library's National Sound Archive, but, even in that case, the cost of running ASR is likely to be no higher than costs for retrospective digitization. The bottom line, therefore, is that investments in reducing fixed training costs promise the greatest potential benefit.

4.2.4 *Augmenting ASR with Wordspotting.* While present ASR systems are able to transcribe most of the words that are spoken, searchers exhibit different patterns of term usage when posing queries. Their natural preference for highly selective terms substantially amplifies the QOV rate for query terms. One solution is to use phonetic wordspotting for the remaining query terms. Most research to date on wordspotting has focused on optimizing accuracy and speed, but relatively little has yet been done to optimize the way ASR and wordspotting results are combined. The fundamental problem is that present statistical techniques rely heavily on estimating likelihood values rather than probabilities. While some ad hoc techniques have been developed, a systematic investigation combining ASR with wordspotting for searching spontaneous conversational speech still remains to be done. Scalability to very large collections also needs investigation. For example, a system with one false alarm per ten hours of speech could be quite useful with a 100-hour collection, but application of the same system to a 10,000-hour collection could be problematic.

4.3 Segmentation

Isolation of appropriate units is particularly important in spoken word collections, both because skimming unnecessarily long speech segments can be time-consuming and because structural metadata can be used to enhance access. Segmenting spontaneous conversational speech is, however, quite different from the work that has been done to date on segmentation of news broadcasts into discrete stories.

Topic-based segmentation of spontaneous conversational speech is relatively complex because topics exist at many degrees of granularity. Experience from the MALACH project suggests that hierarchical topic segmentation models might therefore be more appropriate than a single partition. Evaluation measures for hierarchical topic segmentation are not yet sufficiently well understood so we still need fundamental research on evaluation issues before we can make progress on system design.

Speaker segmentation research has to date focused on either recorded two-wire (i.e., single channel) telephone conversations (where evidence from multiple channels is not available) or recorded meetings (where more than two speakers are typically present). Interviews offer both unique characteristics (e.g., systematic regularities in turn-taking) that could be exploited to optimize diarization accuracy and unique combinations of characteristics that have not yet been well explored in other settings.

4.4 Automated Metadata Generation

Transcribed and segmented speech is an important source for automated metadata generation. Interviews impose a few extra challenges, however, at least in part due to the relatively high rate of transcription errors.

4.4.1 *Challenges for Transcript Annotation.* Annotating ASR transcripts with syntactically or semantically meaningful categories is a challenging problem because simple sliding-window bag-of-word recognition models are typically not sufficiently accurate, and the more sophisticated sequential models (e.g., Hidden Markov Models) are adversely affected by transcription errors that interrupt the correct rendering of word sequences. Two broad classes of approaches seem to be worth exploring in this context: (1) sequence models that operate on representations that specifically encode uncertainty (e.g., pinched word lattices), and (2) multiscale models that leverage a broader context (interview-scale, collection-scale, or query-specific) to constrain the category assignments that would have been hypothesized using narrower sequential contexts.

4.4.2 *Event Detection.* Although some extraction-focused techniques (e.g., named-entity detection) are now reasonably well explored, even in ASR applications, work on more challenging problems such as event detection is still in its infancy. Recorded interviews provide human perspectives on actual events so references to events are often of particular importance. Mention of events might address specific historical events that have a common basis across observers (e.g., the terrorist attacks of September 11, 2001) or generic personal event types for which the time and location of specific instances of those events will vary (e.g., a wedding or a funeral). Present approaches to event detection in text require substantial amounts of tuning to the characteristics of the data, and they have been reported to be quite sensitive to transcription errors [Jing et al. 2007]. Common characteristics of recorded interviews (e.g., temporal ordering with some digressions, turn taking, or prosodic cues) offer potentially useful features that have not yet been fully explored and that might help to yield improved accuracy in event-detection tasks.

4.4.3 *Richer Diarization.* In addition to the words spoken and the structural metadata created, speaker features beyond who-spoke-when could be used for indexing interviews. This would call for investments in more sophisticated speaker diarization, integrating features such as gender recognition, language identification, dialect/accent identification, sociolect identification and recognition of the emotional or cognitive state of the speaker. Quite a lot remains to be done on many of these challenges individually, and integrating them to create richer descriptions of noncontent cues will likely pose additional challenges.

4.5 Search

Effective navigation and retrieval critically depends on ranking techniques that can estimate the relative value of an interview or interview segment. Improving present techniques in ways that draw on the full range of available features therefore could also large dividends. Drawing an analogy to full text search, we might call this full audio search.

4.5.1 *Complex Queries.* Search tools that are based on the words actually spoken allow searchers to access content in ways that were not envisioned when the systems were first designed. Searching descriptive metadata (i.e., conceptual and structural metadata), by contrast, can support searches that are both more inclusive (e.g., to find implied concepts that were not explicitly spoken about) and more precise (e.g., to focus only on interviews with individuals with a specific profile). Allowing structured queries in which both approaches can be combined is therefore an obvious thing to try.

Recorded interviews often exhibit complex interactions between topic, location and time. Affordances for each of these have been explored individually, but their combination poses new challenges. These have been explored for knowledge discovery from text, for example, in the context of the ECAI (Electronic Cultural Atlas Initiative) but not in depth for spoken audio. For instance, lifelines that simultaneously intersect in location and time could be used to identify multiple perspectives on the same event using a query-by-example framework. This functionality could build on a combination of advanced tools for event detection (see Section 4.4.2), domain models optimized for capturing temporal and location information (see Makkonen and Ahonen-Myka [2003]), and event-aware metadata models (see Hunter and James [2000]). Close coupling of query and selection affordances (described in the following) could also help to support fluid strategies for query refinement that span topic, location, and time in ways that permit a more nuanced exploration of a collection than would otherwise be possible.

The presence of an image channel in some types of recorded interviews (e.g., for televised documentary programs) offers an additional challenge for effective searching, because, in addition to conceptual metadata and transcripts, they may be annotated with low-level features. There has been a good deal of research on indexing text descriptions (e.g., photo captions) as a way of helping people to find specific

images, but recorded interviews with associated images pose the dual problem of using images as a basis for finding how those images are being described. There has been some work on similar problems in the context of news video that could be a starting point.

Actually achieving this potential will require more sophisticated integration of spoken words and metadata as a basis for search than has yet been tried. Advanced query interfaces may also be needed to handle structured queries such as "give me male, native French speakers talking about ..." (see also Section 3.1). Complex queries raise a number of important research questions, including how best to store and access multiple annotation layers, how best to combine evidence from multiple annotation layers for ranked retrieval, and how we can best support the development of appropriate mental models and effective search processes by the users of our systems. While similar issues are present in many types of collections, the linear nature of speech and the extended length of some types of interviews mean that improving access to recorded interviews will require greater emphasis on segment-level metadata. Recorded interviews therefore provide an excellent perspective from which to explore this fundamental issue.

4.5.2 *Topical Expansion.* Searching informal language in conversational media (e.g., speech or email) has to date received little attention. Conversational speech exhibits far greater variation in lexical choice than written text, and thus somewhat less scope for learning from one speaker how another might express a similar idea. Some form of vocabulary expansion would therefore be a useful capability. Such techniques might be based on unsupervised machine learning (e.g., clustering terms in a low-dimensional projection such as that produced by Latent Semantic Analysis), supervised machine learning (e.g., using topic classification systems trained on labeled text as a crosswalk for vocabulary expansion), or hand-engineered, rule-based techniques (e.g., associating specific events with the known date on which they occurred).

4.6 Presentation of Aggregate Result Sets

Value estimation techniques can be used to rank interviews (or portions of interviews) in decreasing order of likely utility, but if such a ranked list is to be useful, the searcher will need some way of recognizing where their limited time is best invested.

4.6.1 *Selection Support.* Support for selection must compensate for some of the idiosyncrasies present in automatically generated transcripts of spontaneous conversational speech. Supervised machine learning has been applied to a number of paraphrase problems (e.g., Barzilay and Lee [2003]) so these techniques offer one possible starting point for tuning disfluency repair to supporting selection (rather than the more common goal of supporting comprehension). Because of variation in transcription accuracy, confidence-based reranking of candidate snippets also merits investigation. The absence of clear sentence boundaries in spontaneous conversational speech could result in infelicitous snippet selection, suggesting that integration of automated techniques for detecting sentence-like units in spontaneous conversational speech (e.g., Roark et al. [2006]) could also be helpful. User studies will, of course, be needed to find the optimal mix of capabilities.

4.6.2 *Social Network Depictions.* Recorded interviews often include descriptions of relationships between entities (e.g., family relationships or more complex relationships between people, objects, and events). When specific types of relationships are important factors in satisfying a searcher's information needs, network depictions can be useful as a basis for selection. For example, a searcher might specify a set of people in whom he is interested, and an upper triangular matrix visualization (with those people as columns and as rows) might show which pairs of people are mentioned together in at least one interview. Clicking on a cell would then bring the searcher to a list of interviews with that characteristic.

Network visualization is a well-studied problem, but visualization of relationships detected in spoken content adds the additional complexity of how uncertainty can best be characterized (since entity-detection errors are likely to occur, and relation detection, which presumes detection of at least two entities, will undoubtedly be even more error prone.) Because entities and relations may be referred to in different ways in different interviews, coreference resolution will also be important. That too is a well-studied problem in text, but little work has yet been done on coreference resolution in spontaneous conversational speech.

4.7 Presentation of Individual Search Results

Once a user has selected promising passages from the set suggested by a search system, we need to provide them with ways of expeditiously making final assessments of utility.

4.7.1 *Visualization.* Graphical representations associated with personally meaningful events (which have been called *memory landmarks*) have been found to trigger a useful degree of recall in a personal information-management application [Ringel et al. 2003], and subsequent work on that topic has led to a modeling framework of detection of personally meaningful memory landmarks [Horvitz et al. 2004]. The challenges in applying similar techniques to large collections of recorded interviews are the identification of analogues to memory landmarks that are meaningful to searchers who did not participate in the events being recounted and development of an appropriate modeling framework for controlling automatic selection and placement of those landmarks.

4.7.2 *Transcript-Coupled Replay.* We might speculate that coupling result presentation with query formulation in some way might be desirable. Several interface affordances have been proposed for the display of audio transcripts. Most of these place the initiative on the system rather than the searcher, and to date most have been tested with fairly accurate transcripts. As yet, however, we have little basis other than our own intuition for selecting among these alternatives. Hence, a series of comparative user studies under controlled conditions is needed, followed by field studies to assess the degree to which the laboratory results transfer to practical applications.

4.7.3 *Audio Skimming.* Although there has been a good deal of work on speech compression, little is known about the utility of the compressed speech as a basis for selection in the context of a full audio search system. User studies are needed, both to characterize the added utility from accelerated audio replay and to determine the degree to which decontextualized measures of intelligibility predict utility when used in an environment rich in complementary selection cues.

4.8 Repurposing

Sometimes content use occurs entirely within the context of the search system as might be the case when the searcher wants to find an interview and then listen to it. In other cases, external systems are involved as in the case of using a word processing system to create an annotated bibliography of interviews that address some specific topic. The creation of derivative works such as documentary programming and prepackaged audiovisual materials for classroom use is one example, and the form of content sharing popularly known as remix is another. As podcasting and personal video hosting services have shown, the ability to easily access, manipulate, and share content can stimulate the production and sharing of derivative works. The link-based architecture of the Web further facilitates this by supporting aggregation and organization by reference both within and across collections. It seems important to facilitate reference not just to interviews but to regions within interviews and to do so in a manner that is compatible with the existing architecture of the Web. Extending this framework to written text that is synchronized with spoken words seems straightforward; less clear is what should

be done when no accurate transcript is available. Some inspiration might be drawn from the use of audio icons in screen readers designed for use by the blind however.

4.9 Community Activities

While individual research teams can productively attack many of the challenges identified, several important issues will require broader participation.

4.9.1 *Test Collections.* The application of indexing techniques to spontaneous conversational speech poses new challenges to the evaluation-driven research paradigm. Evaluating search systems using predefined passages is at best an imperfect model of the real task faced by a speech retrieval system, but work on alternative evaluation designs has started only recently [Pecina et al. 2007]. Collection diversity is also an important issue. At present, the only available test collections for ranked retrieval from recorded interviews are drawn from a single source of interviews. Without a diverse set of test collections, it will not be possible to know whether we are tuning our techniques to the task or simply to eclectic characteristics of that one source.

4.9.2 *Iterative Design.* Research studies in which real users participate are far less common than automated batch evaluations, and studies with participation by truly representative users are scarcer still. The potential for synergy between observational studies, design and implementation efforts, and quantitative studies seems particularly important when the design space is as broad as we face in this case. Of course, different users will undoubtedly have very different needs so several studies will be needed if we are to explore the full design space.

4.9.3 *Web-Scale Services.* The World Wide Web now makes it possible to deploy systems that support real users at a scale that was previously unimaginable. Investments in providing public access to a few high-profile collections of recorded interviews using automated techniques would certainly be worthwhile since only in this way will we begin to appreciate the true diversity of users and uses. Deploying such systems will undoubtedly raise thorny issues related to rights management (e.g., a content creator's moral rights to protect the integrity of their work), and personal privacy (e.g., adequate anonymization for data used from system log files). In some cases (e.g., oral histories for which an edited transcript is the authoritative version), redaction of some unreleasable audio may also be needed. Nonetheless, large-scale use offers many opportunities (e.g., including social tagging and recommendation) so it would be well worth pursuing.

5. CONCLUDING REMARKS

Setting a broad agenda is one thing, but making it work requires that we establish some priorities. This broad question has been the focus of several recent efforts, including recent workshops at several major conferences [Aroyo et al. 2007; de Jong et al. 2007a, 2007b; van den Bosch et al. 2007]. As tempting as it might be to look for a silver bullet, it is clear that no single technology holds the key to improving access in the ways that a more balanced investment would make possible. We therefore believe that priorities should be placed on two types of activities: (1) investments early in the processing pipeline that would have broad impact on many types of downstream processing, and (2) activities that bring multiple communities together to address a shared problem.

Among technologies, it seems unlikely that any single investment could be more urgent than enhancing ASR. State-of-the-art word-error rates on spontaneous conversational speech are still far higher than for genres in which people speak more predictably (e.g., broadcast news and dictation), the fixed cost of tuning systems to a new domain severely limits the range of cost-effective applications, and the processing resources that must presently be devoted to actually performing ASR dominate the variable

technology costs for providing access to large collections. These three challenges are coupled, building robust systems requires that we work on many genres and with many exemplars in each genre, and doing this affordably will require that we create efficient and easily adapted systems.

It is a telling comment that the largest public speech-retrieval test collection in the world today (which contains around 1,000 hours of oral history interviews) is not much larger than the well-known Cranfield test collection for text retrieval, which was shown in the early 1990's to be far too small and far too homogeneous to support development of Web-scale applications. If we are ever to move access to recorded speech to Web-scale, we must begin by building ASR systems that can handle the types of content diversity that we can reasonably expect to encounter at that scale and that can process speech at least as quickly as obtain it.

Comparing research on Cross-Language Information Retrieval with research on speech retrieval is illuminating in this regard. Although two teams had been working on CLIR before 1994, it is reasonable to mark that year as the start of a broad push to attack the problem. Thirteen years later, there are well over 100 active research groups, working in dozens of languages. We might mark 1996, the first year of the TREC Spoken Document Retrieval track, as a similar starting point for speech retrieval. Now, over a decade later, there are perhaps a few dozen research groups working actively on spoken document retrieval in just a handful of languages. Why the difference? Simply because the cost of entry is far higher for speech. Until we tackle this problem, speech retrieval is likely to remain a niche market.

In the meantime, we can best steer the overall capacity towards the most productive investments by bringing together researchers from different communities around realistic problems. This requires that we both build bridges between technology developers and potential user communities and that we build bridges between different communities of technology developers. Among developers, differentiated communities such as ASR, NLP, and IR tend to emerge naturally, at least in part because the focused effort that such communities foster results in increased productivity. Trying to build bridges across communities is, therefore, somewhat of an unnatural act, and one that risks actually reducing productivity in the short run as we spend the time to first learn to communicate, and then begin to learn from each other. If we are to succeed at this effort, strong incentives will likely be needed. Indeed, large disciplinary projects of this type are becoming increasingly common in many parts of the world. These projects tend to take us outside our comfort zone, and because they often involve new col,laborations they perhaps involve a somewhat greater element of risk. But if we are right that this is a necessary step, then we are fortunate to already be gaining some of the experience that we will need.

As a concluding remark, we should stress that research on the topics described in this article would clearly benefit from global coordination and from international collaboration. At the most basic level, no one nation could hope to address every aspect of these important challenges on its own. But the reasons to work together are even more pressing than that. Our cultural heritage belongs to all of humanity, which means that no single nation could possible bring an understanding of all of the content, all the users' needs, and all the ways of using that content to meet those needs. In the end, we're all in this together.

APPENDIX

Links to urls of Projects and Initiatives

Digitization and research projects:
AMI IST-FP6 project , `http://www.amiproject.org/`
CATCH programme, `http://www.nwo.nl.catch`
CHoral project, `http://hmi.ewi.utwente.nl/project/CHoral`

DARPA EARS, http://www.darpa.mil/ipto/programs/ears/ ECAI (Electronic Cultural Atlas Initiative), http://www.ecai.org/
ECHO IST-FP5 project , http://pc-erato2.iei.pi.cnr.it/echo
MultiMatch IST-FP6 project, http://www.multimatch.org/
PRESTO IST-FP6 project, http://presto.joanneum.ac.at/index.asp
PrestoSpace IST-FP6 project, http://www.prestospace.org/

Benchmarks:
CLEF (Cross-Language Evaluation Forum), http://www.clef-campaign.org/
CLEF-CLSR (Cross-Language Speech Retrieval), http://clef-clsr.umiacs.umd.edu/
INEX (INitiative for the Evaluation of XML Retrieval) http://inex.is.informatik.uni-duisburg.de/
NIST, http://www.nist.gov/ TDT, http://www.nist.gov/speech/tests/tdt/
TREC, http://trec.nist.gov/
TRECVID, http://www-nlpir.nist.gov/projects/trecvid/

Standards in encoding and archiving:
CHORUS IST-FP6 Coordination Action, http://www.ist-chorus.org/
IASA (International Association of Sound and Audiovisual Archives), http://www.iasa-web.org/
ISAAR(CPF), http://www.ica.org/biblio/ISAAR2EN.pdf
ISAD(G), http://www.ica.org/biblio/isad_g_2e.pdf
EAD (Encoded Archival Description), http://www.loc.gov/ead/
GTAA, for an example cf. http://ems01.mpi.nl:8080/GTAABrowser/
METS (Metadata Encoding and Transmission Standard), http://www.loc.gov/standards/mets/
NIST ACE, http://www.nist.gov/speech/tests/ace
TEI (Text Encoding Initiative), http://www.tei-c.org/
W3C (World Wide Web Consortium), http://www.w3.org/

Radio/TV archives:
- Australian ABC archives, http://www.abc.net.au/archives/av/database.htm
- FIAT/IFTA (International Federation of Television Archives), http://www.fiatifta.org/
- Netherlands Instituut voor Beeld en Geluid, http://www.beeldengeluid.nl/

Examples of digital institutional collections in the oral history domain:
- University of Michigan-Dearborn Voice/Vision Holocaust Survivor Archive, http://holocaust.umd.umich.edu/
- Oral Histories of the American South, http://docsouth.unc.edu/sohp/index.html
- Kilbirnie-Lyall Bay Community Centre, http://kilbirnie.natlib.govt.nz/
- Oral History of British Photography, http://www.bl.uk/collections/sound-archive/historyphoto.html
- Voices of the Colorado Plateau, http://archive.li.suu.edu/voices/voicesx3.html
- Community Memory of Veterans History, http://www.loc.gov/vets/
- Migration Audio Archiv, http://www.migration-audio-archiv.de/audioweb/
- Voice/Vision Holocaust Survivor Archive, http://holocaust.umd.umich.edu/
- Stories of the Dreaming website, http://www.dreamtime.net.au/dreaming/index.htm
- British Library National Sound Archive, http://www.bl.uk/nsa
- National Library of Australia, http://www.nla.gov.au/ohdir/index.html
- Library of Congress, http://www.loc.gov/index.html

- Shoah Foundation for Visual History http://www.usc.edu/schools/college/vhi/
- Story Corps, http://www.storycorps.net
- Monument Jewish Community in the Netherlands, http://www.joodsmonument.nl/
- Radio Oranje, http://hmi.ewi.utwente.nl/choral/demo

REFERENCES

ALLAN, J., Ed. 2002. *Topic Detection and Tracking: Event-Based Information*. Kluwer Academic Publishers, Boston, MA.

ALLAUZEN, A. AND GAUVAIN, J.-L. 2005. Open vocabulary ASR for audiovisual document indexation. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 1013–1016.

AROYO, L., KUFLIK, T., STOCK, O., AND ZANCANAR, M., Eds. 2007. *Proceedings of the User Modeling Conference Workshop on Personalization Enhanced Access to Cultural Heritage*.

AUZANNE, C., GAROFOLO, J., FISCUS, J., AND FISHER, W. 2000. Automatic language model adaptation for spoken document retrieval. In *Proceedings of Recherche d'Information Assistée par Ordinatour (RIAO)*. 132–141.

BARZILAY, R. AND LEE, L. 2003. Learning to paraphrase: An unsupervised approach using multiple-sequence alignment. In *Proceedings of HLT-NAACL*. 16–23.

BLOK, H. E., MIHAJLOVIC, V., RAMIREZ, G., WESTERVELD, T., HIEMSTRA, D., AND DE VRIES, A. 2006. The TIJAH XML information retrieval system. In *Proceedings of the 29th International ACM SIGIR Conference*. 725–725.

BROWN, M. G., FOOTE, J., JONES, G., SPÄRCK JONES, K., AND YOUNG, S. J. 1995. Automatic content-based retrieval of broadcast news. In *Proceedings of the 3rd ACM International Conference on Multimedia*. ACM Press. 35–43.

BURGET, L. 2005. Combination of speech features using smoothed heteroscedastic linear discriminant analysis. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP'04)*. Jeju island, KR.

BYRNE, W., DOERMANN, D., FRANZ, M., GUSTMAN, S., HAJIC, J., OARD, D., PICHENY, M., PSUTKA, J., RAMABHADRAN, B., SOERGEL, D., WARD, T., AND ZHU, W.-J. 2004. Automatic recognition of spontaneous speech for access to multilingual oral history archives. *IEEE Trans. Speech Audio Process. 12,* 4, 420–435.

CHRISTEL, M. G. AND YAN, R. 2007. Merging storyboard strategies and automatic retrieval for improving interactive video search. In *Proceedings of the 6th ACM International Conference on Image and Video Retrieval*. 486–493.

DE JONG, F., OARD, D., ORDELMAN, R., AND RAAIJMAKERS, S., Eds. 2007a. *Proceedings of ACM SIGIR Workshop on Searching Spontaneous Conversational Speech*.

DE JONG, F., ORDELMAN, R., AND VAN HESSEN, A. 2006. The role of automated speech and audio analysis in semantic multimedia annotation. In *Proceedings of the International Conference on Visual Information Engineering*. Bangalore, India, 226–240.

DE JONG, F. M. G., OARD, D., ORDELMAN, R., AND RAAIJMAKERS, S. 2007b. Searching spontaneous conversational speech. *ACM SIGIR Forum 41,* 2, 104–108.

GAROFOLO, J., AUZANNE, C., AND VOORHEES, E. 2000. The TREC SDR track: A success story. In *Proceedings of the 8th Text Retrieval Conference*. Washington, DC, 107–129.

GODFREY, J., HOLLIMAN, E., AND MCDANIEL, J. 1992. Switchboard: telephone speech corpus for research and development. In *IEEE the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. San Francisco, CA, Vol. 1. 517–520.

GOLDMAN, J., RENALS, S., BIRD, S., DE JONG, F., STEWART, C., FREDERICO, M., FLEISCHHAUER, C., LAMEL, L., KORNBLUH, M., SEBASTIANI, F., OARD, D. W., AND WRIGHT, R. 2003. Report of the EU/NSF working group on Spoken Word Audio Archives. http:/www.ercim.org/publication/ws-proceedings/Delos-NSF/SpokenWord.pdf.

GOLDMAN, J., RENALS, S., BIRD, S., DE JONG, F. M. G., FEDERICO, M., FLEISCHHAUER, C., KORNBLUH, M., LAMEL, L., OARD, D. W., STEWART, C., AND WRIGHT, R. 2005. Accessing the spoken word. *Int. J. Digital Libraries 5,* 4, 287–298.

GUSTMAN, S., SOERGEL, D., OARD, D., BYRNE, W., PICHENY, M., RAMABHADRAN, B., AND GREENBERG, D. 2002. Supporting access to large digital oral history archives. In *Proceedings of the Joint Conference on Digital Libraries*. 18–27.

HANSEN, J., HUANG, R., ZHOU, B., DEADLE, M., DELLER, J., GURIJALA, A. R., KURIMO, M., AND ANGKITITRAKUL, P. 2005. Speechfind: Advances in spoken document retrieval for a national gallery of the spoken word. *IEEE Trans. Speech Audio Process. 13,* 5, 712–730.

HEEREN, W. F. L., VAN DER WERFF, L. B., ORDELMAN, R. J. F., VAN HESSEN, A. J., AND DE JONG, F. M. G. 2007. Radio Oranje: Searching the queen's speech(es). In *Proceedings of the 30th ACM SIGIR Conference*. The Netherlands. ACM 903.

HORI, C. AND FURUI, S. 2003. A new approach to automatic speech summarization. *IEEE Trans. Multimedia 5*, 368–378.

HORVITZ, E., DUMAIS, S., AND KOCH, P. 2004. Learning predictive models of memory landmarks. In *Proceedings of the Cognitive Science Society*.

Huijbregts, M. A. H., Ordelman, R. J. F., and de Jong, F. M. G. 2007. Annotation of heterogeneous multimedia content using automatic speech recognition. In *Proceedings of the 2nd International Conference on Semantic and Digital Media Technologies (SAMT)*.

Hunter, J. and James, D. 2000. The application of an event-aware metadata model to an online oral history project. In *Proceedings 4th European Conference on Research and Advanced Technology for Digital Libraries (ECDL)*. 291–304.

Hürst, W., Lauer, T., and Götz, G. 2004. An elastic audio slider for interactive speech skimming. In *Proceedings of the Norde Conference on Computer-Human Interaction (NordCHI'04)*. 277–280.

Jing, H., Kambhatla, N., and Roukos, S. 2007. Extracting social networks and biographical facts from conversational speech transcripts. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*. Association for Computational Linguistics, Czech Republic, 1040–1047.

Jourlin, P., Johnson, S., Spärck Jones, K., and Woodland, P. 1999. General query expansion techniques for spoken document retrieval. In *Proceedings of the ESCA Workshop on Extracting Information from Spoken Audio*. Cambridge, UK, 8–13.

Kim, J., Oard, D., and Soergel, D. 2003. Searching large collections of recorded speech: A preliminary study. In *Proceedings of the Annual Conference of the American Society for Information Science and Technology*. Long Beach, CA, 330–339.

Klemmer, S., Graham, J., Wolff, G., and Landay, J. 2003. Books with voices: Paper transcripts as a tangible interface to oral histories. In *Proceedings of Computer-Human Interaction (CHI)*. Ft. Lauderdale, FL.

Legetter, C. and Woodland, P. 1995. Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. *Comp. Speech Lang. 9*, 171–185.

Makkonen, J. and Ahonen-Myka, H. 2003. Utilizing temporal information in topic detection and tracking. In *Proceedings of the European Conference on Digital Libraries*. 393–404.

McKeown, K., Hirschberg, J., Galley, M., and Maskey, S. 2005. From text to speech summarization. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 997–1000.

Morang, J., de Jong, F., Ordelman, R., and van Hessen, A. 2005. Infolink: analysis of dutch broadcast news and cross-media browsing. In *Proceedings of the IEEE International Conference on Multimedia*. Amsterdam, The Netherlands, 1582–1585.

Moreno, P. J., Joerg, C., Thong, J.-M. V., and Glickman, O. 1998. A recursive algorithm for the forced alignment of very long audio segments. In *Proceedings of the 5th International Conference on Spoken Language Processing (ICSLP)*. Sydney, Australia.

Ng, K. 2000. Subword-based approaches for spoken document retrieval. Ph.D. thesis, Massachusetts Institute of Technology.

Oard, D., Demner-Fushman, D., Hajic, J., Ramabhadran, B., Gustman, S., Byrne, W., Soergel, D., Dorr, B., Resnik, P., and Picheny, M. 2002. Cross-language access to recorded speech in the malach project. In *Proceedings of the Text, Speech, and Dialog Workshop*. Brno, Czech Republic. 197–212.

Olsson, J. S. and Oard, D. W. 2007. Improving text classification for oral history archives with temporal domain knowledge. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 623–630.

Oomen, J. and Smulders, H. 2006. First analysis of metadata in the cultural heritage domain. MultiMatch report. `http://www.multimatch.org`.

Ordelman, R., de Jong, F., and Heeren, W. 2006. Exploration of audiovisual heritage using audio indexing technology. In *Proceedings of the 1st ECAI Workshop on Intelligent Technologies for Cultural Heritage Exploitation*. 36–39.

Pecina, P., Hoffmannova, P., Jones, G. J., Zhang, Y., and Oard, D. W. 2007. Overview of the CLEF-2007 cross language speech retrieval track. In *Working Notes for the CLEF 2007 Workshop*.

Ringel, M., Cutrell, E., Dumais, S. T., and Horvitz, E. 2003. Milestones in time: The value of landmarks in retrieving information from personal stores. In *Proceedings of Interact*.

Roark, B., Liu, Y., Harper, M., Stewart, R., Lease, M., Snover, M., Shafran, I., Dorr, B., Hale, J., Krasnyanskaya, A., and Yung, L. 2006. Reranking for sentence boundary detection in conversational speech. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Rosenfeld, R. 1995. Optimizing lexical and N-gram coverage via judicious use of linguistic data. In *Eurospeech 95*. 1763–1766.

Shneiderman, B. and Plaisant, C. 2007. Strategies for evaluating information retrieval tools" multi-dimensional in-depth long-term case studies. In *Proceedings of the 2006 AVI Workshop on Beyond Time and Errors: Novel Evaluation Methods for Information Visualization*. 1–7.

Smeaton, A. F., Morony, M., Quinn, G., and Scaife, R. 1998. Taiscéalaí: Information Retrieval from an Archive of Spoken Radio News. In *Proceedings of the 2nd European Conference on Research and Advanced Technology for Digital Libraries (ECDL2)*. Crete, 429–442.

Smeaton, A. F., Over, P., and Kraaij, W. 2006. Evaluation campaigns and trecvid. In *Proceedings 8th ACM SIGMM International Workshop on Multimedia Information Retrieval (MIR)*.

Steijlen, F. 2002. *Memories of The East*. KITLV Press, Leiden, The Netherlands.

ULARGIU, B. 2000. Accessibility of oral history collections: An investigation into current practices and future developments. Masters thesis, University of Sheffield.

VAN DEN BOSCH, A., GROVER, C., AND SPORLEDER, C., Eds. 2007. In *Proceedings of the ACL Workshop on Language Technology for Cultural Heritage Data*.

WELLING, L., KANTHAK, S., AND NEY, H. 1999. Improved methods for vocal tract length normalization. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Phoenix AZ, 761–764.

WHITTAKER, S., HIRSCHBERG, J., CHOI, J., HINDLE, D., PEREIRA, F., AND SINGHAL, A. 1999. SCAN: Designing and evaluating user interfaces to support retrieval from speech archives. In *Research and Development in Information Retrieval*. 26–33.

WOODLAND, P., JOHNSON, S., JOURLIN, P., AND SPÄRCK JONES, K. 2000. Effects of out of vocabulary words in spoken document retrieval. In *ACM SIGIR Conference*. Athens Greece, 372–374.

WRIGHT, R. AND WILLIAMS, A. 2001. Presto - preservation techniques for European broadcast archives. IST-1999-20013.

ZHANG, P., PLETTENBERG, L., KLAVANS, J. L., OARD, D. W., AND SOERGEL, D. 2007. Task-based interaction with an integrated multilingual, multimedia information system: A formative evaluation. In *Proceedings of the Joint Conference on Digital Libraries*. 117–126.

ZHANG, P. AND SOERGEL, D. 2006. Knowledge-based approaches to the segmentation of oral history interviews. MALACH technical report, College of Information Studies, University of Maryland, College Park, MD.

Received December 2006; revised May 2007, August 2007; accepted November 2007