International Journal on Digital Libraries manuscript No.
(will be inserted by the editor)

Accessing the Spoken Word

Jerry Goldman[1], Steve Renals[2], Steven Bird[3], Franciska de Jong[4] Marcello Federico[5], Carl Fleischhauer[6], Mark Kornbluh[7], Lori Lamel[8], Douglas W. Oard[9], Claire Stewart[10], Richard Wright[11]

1 Department of Political Science, Northwestern University, USA; e-mail: j-goldman@northwestern.edu
2 CSTR and School of Informatics, University of Edinburgh, UK; e-mail: s.renals@ed.ac.uk
3 LDC, University of Pennsylvania, USA; and Dept of Computer Science, University of Melbourne, Australia; e-mail: sb@ldc.upenn.edu
4 CTIT, University of Twente, The Netherlands; e-mail: fdejong@ewi.utwente.nl
5 ITC-IRST, Trento, Italy; e-mail: federico@itc.it
6 Library of Congress, USA; e-mail: cfle@loc.gov
7 MATRIX and Department of History, Michigan State University, USA; e-mail: mark@mail.matrix.msu.edu
8 LIMSI-CNRS, Orsay, France; e-mail: lamel@limsi.fr
9 College of Information Studies/UMIACS, University of Maryland, USA; e-mail: oard@umd.edu
10 Library, Northwestern University, USA; e-mail: claire-stewart@northwestern.edu
11 BBC Information and Archives, UK; e-mail: richard.wright@bbc.co.uk

The date of receipt and acceptance will be inserted by the editor

Abstract. Spoken word audio collections cover many domains, including radio and television broadcasts, oral narratives, governmental proceedings, lectures, and telephone conversations. The collection, access and preservation of such data is stimulated by political, economic, cultural and educational needs. This paper outlines the major issues in the field, reviews the current state of technology, examines the rapidly changing policy issues relating to privacy and copyright, and presents issues relating to the collection and preservation of spoken audio content.

Key words: Spoken document retrieval, preservation, privacy, copyright, browsing, speech technology, content annotation, content management

1 Introduction

Christiansen defines a "disruptive technology" as one that rebalances the playing field in ways that fundamentally change the value proposition [9]. That describes well the remarkable potential of recent advances in speech processing to transform the information society. Humans are story tellers, optimized through evolution to produce and understand speech. Durability and accessibility have given primacy to the written word for thousands of years. The exclusivity of text may now yield some authority to the spoken word since digital storage for any medium (text, images, audio, video) is identical. In addition, developments in speech technology have improved the capability to access spoken word collections rapidly and effectively. This paper arises from the DELOS/NSF working group in Spoken Word Audio Collections. It describes the present state of knowledge and outlines a research agenda for digital library research in this area.

Well disclosed spoken word collections[1] can offer substantial value to individuals, organizations, and society, in many different areas including rapid access to archived lectures, the personalized delivery of news broadcasts, and memory augmentation (for instance replaying a conversation several weeks later). A broad range of commercial, nonprofit, community and government organizations are potential users of natural, content-based access to speech archives. Such services could aid in efficiently disseminating information by routing segments of recorded meetings to remote team members, by improving services—for instance by mining help-desk calls—and increasing the efficiency of markets by alerting analysts to breaking news.

Benefits that accrue to society as a whole include the preservation of our cultural heritage through access to oral history and folklore collections of unprecedented scale, searchable access to government records such as parliamentary debates, and important new capabilities that can enhance scholarly inquiry in areas such as sociolinguistics.

To capitalize on the opportunities provided by spoken word collections requires meeting three challenges. First is the dependency on imperfect, but steadily improving speech processing technologies for automated segmentation, transcription, and annotation. Second, if we are to meet the needs of real users, we must draw on expertise in content management from curators of many types of digital collections. And third, we must bring these communities together with the information access communities if we are to build systems and processes that match the needs of real users. These three per-

[1] We focus in this paper on audio, but the ideas apply equally well to multimedia materials in which audio can be used as a basis for access, eg video.

spectives explain the formation of the working group that has authored this paper.

The paper is organized as follows. Section 2 reviews the present state of speech technology. Section 3 considers best practice for the management of spoken word collections, and section 4 addresses policy issues raised by these emerging capabilities. Section 5 proposes a research agenda to capitalize on the unique opportunities that the emerging technologies present and finally section 6 provides a brief conclusion.

2 The Technology Landscape

Speech recognition technology has made substantial advances in accuracy over the past decade. This has formed the basis for a variety of systems that index audio and multimodal archives, particularly for television and radio news broadcasts (eg [30, 33, 23]). Together with a variety of other indexing and content extraction technologies, the rapid development of large vocabulary speech recognition systems has made content-based access to some types of spoken word archives readily available. A significant impetus for these developments has been the technology evaluation programs (and the concomitant resource collection) in speech recognition, text retrieval and information extraction, coordinated in the USA by the Defense Advanced Research Projects Agency (DARPA) and the National Institute for Standards and Technology (NIST).[2] These speech and language technologies have had a focus on a few economically important languages, in particular North American English.

By using speech recognition to convert speech into text, detailed text representations can be generated for spoken content. These are not exact renderings of the spoken content, but they enable specific words and phrases to be indexed. This core capability is well suited for a variety of tasks. Since speech recognition systems can label recognized words with exact time stamps, the time information can be used to direct users to relevant audio fragments (perhaps with links to related content, such as video).

This section will describe how functionality for browsing and search, speech processing, and content annotation each contribute to the disclosure of spoken word content, and outlines the outstanding research issues.

2.1 Browsing and search

The ability to browse and search a spoken audio collection presupposes that the user is able to discover that the collection exists and gain access to a copy. Increasingly, spoken audio collections are being documented using Dublin Core Metadata,[3] which provides a core set of 15 descriptors that can be used to catalog a resource (including title, creator, language and rights). The Open Language Archives Community (OLAC) [28] provides additional descriptors that are appropriate for spoken audio resources, including a set of language identifiers that uniquely identify the language(s) spoken in the recording, and a classification of 'discourse type' (eg drama, formulaic discourse, or singing). OLAC participates in the Open Archives Initiative, which provides comprehensive infrastructure for cross-archive searching [17]. At the time of writing, over 30,000 resources in some 24 language archives can be searched simultaneously via the OLAC Gateway.[4]

2.1.1 Speech retrieval

Browsing accessible spoken word collections typically relies on a time-stamped transcription, along with other annotations. A wealth of tools exist for manual transcription and annotation, and these are in widespread use for languages, recording conditions, and coding tasks that are not well-supported by automatic speech recognition technologies [5].

Transcriptions generated by an automatic speech recognizer are usually errorful. However, since human speech is somewhat redundant, retrieval effectiveness has proven to be fairly robust in the presence of recognition errors up to a word error rate of 30–40% [11]. An intuitive presentation of retrieval results is hindered in several ways. It is not easy to read speech recognition output due to the errors and because most automatic transcripts do not include punctuation (although such markup is currently being addressed by the speech recognition community). Recognition of unknown words (which are common in freely compounding languages such as German and Dutch) and proper names can also be problematic. Simply put, accurate retrieval often requires some listening.

Interactive search of a spoken word collection involves query formulation, automated ranking, selection, and replay [25]. Query formulation and automated ranking is similar to text retrieval, with the searcher formulating a query as free text or as a Boolean expression, and the system returning a set of documents, hypothesised to be relevant to the query, based on matching the query to the transcription. The selection stage, which allows searchers to rapidly discover the most promising documents from the system-ranked list, differs from standard text retrieval, since it is likely to be based on terse indicative summaries, rather than raw transcripts. Because such summaries may not provide enough information to support a final selection decision, modern systems also typically provide searchers with the ability to replay segments of individual recordings or to view the complete automatic transcript of the segment.

Recorded speech poses both challenges and opportunities for the interactive retrieval process. The key challenges are deceptively simple: automatic transcription is imperfect and listening to recordings can be time consuming. Some important opportunities include potential use of speaker identifica-

[2] http://www.nist.gov/speech/tests/

[3] http://dublincore.org/documents/dces/

[4] http://www.language-archives.org

tion, speaker turn detection, dialog structure, channel characteristics (telephone or recording studio) and associated audio, such as background sounds, to enhance either the sorting or the browsing process. Multimedia integration, particularly with video or background text documents, also offers some important opportunities for synergy. For example, a document list returned from an initial textual query aimed at finding spoken words, might be refined using selection based on key frames extracted from video.

2.1.2 Topic detection and tracking

Speech recognition can also be used as a basis for fully automated search processes, as demonstrated in the Topic Detection and Tracking (TDT) evaluations [32]. The TDT evaluations include five tasks for automatic processing of broadcast news: story segmentation, clustering (an unsupervised learning task in which systems seek to cluster stories together if they report on the same event), topic tracking (a semi-supervised learning task in which systems seek to identify subsequent news stories that report on an event described by one or more example stories), new event detection (in which systems seek to identify the first story to report on each event), and story link detection (in which systems seek to determine whether pairs of stories report on the same event). Some European projects [26] have addressed similar themes.

2.1.3 Cross-language retrieval

When searchers lack the language skills needed to pose their query using terms from the same language as the spoken content that they seek, some form of support for translation must be embedded within the search system. Such a capability might be useful to searchers who can understand the spoken language but find it easier to formulate queries in another language, if the context is multimodal and the principal object of the query is not linguistic (eg an image), or if suitable translations can be provided of the target documents. At present, speech-to-speech translation has been demonstrated only in limited domains, such as travel planning, but development of more advanced capabilities is the focus of a substantial research investment [31].

Cross-language information retrieval is based on query translation, document translation or interlingual techniques [24]. Query translation architectures are well suited to situations in which many query languages must be supported. In interactive applications, query translation also offers the possibility of exploiting interaction designs that might help the searcher better understand the system's capabilities and/or help the system better translate the searcher's intended meaning. "Document translation" is actually somewhat of a misnomer, since it is the internal representation of the spoken content that is translated. Document translation architectures are well suited to cases in which query-time efficiency is an important concern. Document translation also offers a greater range of possibilities for exploiting linguistic knowledge because spoken content typically contains many more words than a query, and because queries are often not grammatically well formed. With interlingual techniques, both the query and the document representations are transformed into some third representation to facilitate comparisons. Interlingual techniques may be preferred in cases where many query languages and many document languages must be accommodated simultaneously, or in cases where the conforming space is automatically constructed based on statistical analysis of texts in each language.

2.2 Speech technologies

The browsing and search techniques described above rely on speech and audio technologies such as audio partitioning, speech enhancement, speech recognition and speaker identification [22]. The first speaker-independent large vocabulary continuous speech recognition systems were developed in the early 1990s. In the mid-1990s the emphasis switched to the recognition of broadcast news and to conversational telephone speech, which have remained the focus of research. More recently there has been an extension to additional languages and to more challenging tasks such as the transcription of conversational data from meetings (with multiple talkers) and speech recorded in noisy (ie, realistic) conditions.

2.2.1 Audio partitioning

Audio partitioning is concerned with segmenting an audio stream into acoustically homogeneous chunks and classifying them according to a broad set of acoustic classes, for instance speech and music. In many systems, the classification of speech segments is refined by considering factors such as the signal bandwidth, the gender of the speaker, the speaker's identity, or the level of noise. The difficulty of this task increases with the level of detail required. For instance, while detecting speaker turns in conversational speech is often relatively easy, it can be very difficult when two (or more) talkers are speaking at the same time.

A variety of statistical algorithms for acoustic segmentation have been developed in recent years. The most influential have operated by dividing a segment into two parts if it is more probable that the observed acoustics come from two segments with distinct audio characteristics [8]. The task of labeling segments is typically treated as a statistical classification problem using Gaussian mixture models or neural networks. Audio partitioning has been applied most successfully to broadcast news transcription. The application to other audio collections poses problems of portability and robustness of the methods, particularly if there are multiple acoustic sources (both speech and non-speech) or if the audio signal is degraded.

2.2.2 Speech enhancement

Signal processing techniques can be applied to speech to enhance both intelligibility by human listeners and the accuracy of subsequent automatic processing, particularly speech recognition. Human perception is far more robust than present automated approaches to speech recognition [19], so speech enhancement is the focus of a substantial research effort (eg [21]), mainly concerned with the accommodation of environmental factors (such as vehicle noise or reverberation due to room acoustics) and the effect of transmission channels (such as cellular telephones).

Audio restoration is mainly concerned with the improvement of intelligibility and the listening experience, applied to recorded material. In addition to environmental factors, analog recordings might be degraded when they are first created (if the microphone had an imperfect frequency response), during duplication, during storage (due to media decay), as a result of prior use, and during replay. Most current approaches to audio restoration are based on a statistical model of additive noise bursts (eg "thumps" from Dictabelt loops) [14].

2.2.3 Speech recognition

Speech recognition is concerned with converting the speech waveform (an acoustic signal) into a sequence of words. Automatic speech recognition (ASR) is a challenging problem, with a set of complicating factors. The audio signal may contain background music or crowd noise in addition to speech. Significant acoustic differences between speakers arise due to anatomical differences, and an individual speaker's acoustics may be dependent on factors such as their state of health at the time the recording was made. Finally, a speaker's choice of words and speaking style may exhibit variations that relate to the social context.

Current approaches to speech recognition are statistical in nature [35]. A statistical speech recognition system comprises a *language model* that governs the generation of word sequences (by estimating the probability of producing any given word sequence), and an *acoustic model* which describes the generation of the audio signal from a word string. These generative models are inverted to perform speech recognition: given an observed acoustic signal, find the string of words most likely to have generated it. A set of well understood algorithms and models are used to perform this process efficiently.

The power of statistical speech recognition lies in the fact that the acoustic and language models can be trained from large amounts of speech and text data. This training process requires annotated speech corpora for all languages and audio data types of interest, with the resulting recognition accuracy depending strongly on the availability of a sufficient quantity of representative accurately transcribed speech. Speaker independence is obtained by estimating the parameters of the acoustic models on large speech corpora containing data from a large speaker population.

State-of-the-art systems are typically trained on several tens to hundreds of hours of manually transcribed speech and several hundred million words of related texts. While the same basic technology has been successfully applied to different languages and types of speech, there have been many advances in speech recognition accuracy over the last decade. These advances can be partially attributed to advances in robust feature extraction, acoustic modeling with effective parameter sharing, unsupervised adaptation to speaker and environmental condition, efficient decoding algorithms, the availability of huge audio and text corpora for model estimation, and increased computational power [12].

Despite these improvements in accuracy, speed and robust operation remain as challenges. Present techniques allow a tradeoff between speed and accuracy in a limited range, but even the fastest systems generate words several orders of magnitude more slowly than other components of an information access system can index those words. Speech recognition is thus presently a dominant factor in the overall cost of providing automated access to spoken word collections. The difficulty of providing robust operation in the presence of differing acoustic conditions and speaking styles is an equally important limitation in many applications. Present techniques rely on the availability of a coherent set of representative examples. Application of these techniques to a new task or domain therefore often requires a retraining process, which can become quite expensive if a substantial amount of manual transcription is required. Reducing the porting costs and increasing model genericity are very active research areas in the ASR community. Another outstanding challenge is the recognition of previously unseen words (ie, those not occurring in the audio or textual training data) since these are unknown to the ASR system.

2.2.4 Speaker identification and tracking

Accurately identifying a speaker is an unsolved research problem, despite several decades of research [7]. The problem is quite close to that of speech recognition in that the speech signal encodes both linguistic information (ie the word sequence which is of interest for speech recognition) and paralinguistic information including speaker identity, mood, emotion, and attention level. The characteristics of a given individual's voice change over time (short and long periods) and depend on the speaker's emotional and physical state. The identification problem is also highly influenced by the environmental, recording, and channel conditions. For example, it is very difficult to determine if a voice is the same in the presence of background music or noise.

Several types of speaker recognition problems can be distinguished: speaker identification, speaker detection and tracking, and speaker verification (also called speaker authentication). In speaker identification the absolute identity of the speaker is determined. In contrast, for speaker verification the task is to determine if a speaker is who they claim to be. Speaker tracking refers to finding audio segments from the same speaker, even if the identity of the speaker is unknown. Automatically identifying speakers and tracking them throughout individual recordings and in recording collections can allow digital library users to access spoken word documents based on who is talking. Some of the recent speaker tracking research can potentially allow speakers to be located in large audio corpora using a sample of speech, even if the absolute identity of the speaker is unknown. Most of today's working speaker recognition systems use statistical approaches similar to speech recognition. Current research issues include the use of multiple types of acoustic, supra-linguistic and phonetic attributes, and the incorporation of machine learning approaches.

2.3 Content annotation

A considerable amount of value can be added to a spoken audio collection by the incorporation of automatically extracted annotations ranging from punctuation and speaker labelling to complete summaries.

Most current approaches to the extraction of linguistic content operate on transcripts. In addition to problems arising from speech recognition errors, such approaches also lose some of the distinctive elements of speech communication: a spoken message contains more than simply what was said and who said it. The prosody—timing, intonation and stress—of the speech signal offers a great deal of information about the emotional state of the speaker, "punctuation" in the speech and disambiguation of the intended message (questions have a rising intonation, for instance). When there are multiple speakers, a further source of information is the interaction between the speakers (the pattern of speaker turns). While accurate identification and annotation of such paralinguistic characteristics remains an open research problem, improved speech processing technologies has led to growing interest in such areas.

2.3.1 Annotation and transcription

The extraction of information from spoken audio ranges from annotations relating to meaningful segmentations based on topic, speaker, acoustic conditions or punctuation, to named entities (people, organizations, and locations), attributes, facts, and events. Segment annotations are important for further processing: for instance, machine translation algorithms require the identification of sentence boundaries, and summarization algorithms perform better if topic boundaries are available. The difficulty of information extraction is related to the natural language processing required to recognize complex concepts, the intrinsic ambiguity of named entities (eg "Barcelona" could denote a city or a football team, depending on the context), and the steady evolution of language, whereby new words, particularly names, routinely appear in the media, while others disappear or occur with a much lower frequency.

Recent research on information extraction from spoken audio has been carried out in Europe (via several EC projects) and the US (under various DARPA and NIST programs). Much current work has been applied to broadcast news, with state-of-the-art performance achieved by both statistical and rule-based systems [3,15]. Open research issues include the extraction of more complex entities, the identification of relations among entities, the development of domain-independent systems, and application to other speech domains (eg conversational speech).

2.3.2 Summarization

By speech summarization we usually mean techniques that reduce the size of automatically generated transcripts in a way that resembles summarization technology for text documents. The goal is to present the most important content in a spoken document in a condensed form, sensitive to the needs of the user and the task. Furthermore, speech summaries may be more readable than automatically generated transcripts, since they do not include disfluencies, repairs, repetitions, etc.

Speech summarization is a rather young area, and is currently based on approaches developed for text, applied to speech transcripts, typically involving the extraction of key sentences and their compression [16]. It is still an open issue how well these textual based methods work on ASR-generated speech transcripts. Other issues include the use of recognition confidence scores, alternative word choices and the incorporation of non-textual features such as prosody and interaction patterns.

It is possible to summarize speech using the audio alone, and prototype speech skimming systems have been developed [4]. An important issue in this case is the development of accelerated audio playback, which is an interesting signal-processing task if intelligibility and speech characteristics (such as intonation) are to be maintained as much as possible. This area is rather closely related to speech synthesis.

3 Content Management

The management of spoken word collections concerns: acquiring, formatting and processing the sound file; attaching metadata; packaging of data and metadata; and issues relating to the sustainability of the content. We discuss content management in reference to an organization that wishes to obtain and maintain content for the long term, and also wishes to make that content available to its community of users during the same period. We write from a particular perspective representing public archives, such as research libraries or national collections, and to some degree our ideas reflect organizations with a broad public responsibility. The technical concepts apply as well to corporate, private, or for-profit archives.

3.1 Content acquisition

Digital speech can be acquired in one of four ways: creation, deposit, capture, or digitization. Digital recordings of the spoken word are routinely created as a part of many activities. Examples include: news broadcasters preparing stories, air traffic controllers communicating with aircraft in flight, and individuals recording messages on a telephone answering machine. Copyright deposit laws and archival retention schedules are another source of spoken word materials; in such cases, some degree of coordination between the content creators and the depository institution is needed. Digital speech can also be captured when it is transmitted, for instance, digital radio. Finally, existing analog recordings can be digitized.

Existing spoken word collections cover an enormous range, from the earliest recordings of public speeches and broadcasts on wax cylinders and 78rpm records, to oral histories on cassette tape, through to contemporary digital recordings of broadcast news. The social and historical implications of these collections are striking. A survey carried out in 2002 by the PRESTO project estimated that national broadcast archives hold on the order of 100 million hours of spoken language recordings, 80% of it in analog form.[5] The important fact about all analog material on tape is that it will perish within a few decades and that it is expensive to digitize. The PRESTO survey estimated that the preservation reformatting of these analog recordings would cost roughly US $100 per hour. Archival reformatting is generally carried out in real time and produces files at the resolution of compact disks and sometimes higher. Furthermore, as digital systems replace analog systems, and as recording and storage costs decline, there will be an accelerated growth in the creation of spoken word documents, and a corresponding demand for effective strategies for archiving and retrieval.

One aspect of preservation and archiving concerns the initial acquisition of content. For instance, extensive bodies of tape-recorded testimony resulting from oral history projects languish in small local libraries and historical societies. Similarly, many scholars who study language and dialect possess personal collections of sound recordings that have resulted from their research. In some cases, these may constitute the only record of an extinct language. There is clearly a public good to be served by placing these pre-existing, analog-format materials (or copies) in larger and more robust institutional archives.

More recent sources of digital content can be found on the World Wide Web and other online contexts. This content is often ephemeral and short-lived. Archivists sometimes refer to this online content as intangible to distinguish it from digital content distributed in fixed media like compact disks. In recent years, the Library of Congress (US) and other national libraries have begun to collect and archive Web content, although to date this has generally not included sound recordings like radio webcasts. Those with an interest in spoken language, of course, will encourage the expansion of current collecting in order to secure this important cultural record for future generations. Production organizations, such as broadcasters, share with public institutions the social responsibility of safeguarding this content.

In Europe, legal deposit legislation obliges publishers to place copies of printed matter in national libraries. Recent cases in France, Sweden, and Denmark have extended the definition of 'publication' to websites. This legislation is somewhat in advance of comprehensive Web archiving and preservation technology, but the action has launched a process of archiving Web content in Europe, including initial attempts to take audiovisual content from websites. National broadcast organizations like the BBC in Great Britain, and other major producers of media websites, are also actively involved in archiving content, including audio and video. Finally, the Internet Archive[6], an independent non-profit organization in the United States, is attempting to archive as much of the World Wide Web as is practical, and in a project shared with the Library of Congress has already made an impressive collection of broadcast coverage (audio and audiovisual) of the terrorist attacks in the United States on September 11, 2001.

Two technologies of interest regarding the acquisition of intangible born-digital content are those that identify or filter content, for instance in the context of web harvesting; and those that capture the transmitted bitstream, which may be in some proprietary format. There are few tools available to accomplish these goals.

3.2 Content format

The core content element for those with an interest in spoken language is the sound recording itself. In the digital realm, this is represented by a bitstream, typically contained in a computer file. Spoken language processing is able to take advantage of a range of file and bitstream types, even when the quality as judged by an audiophile is only good. Archivists with an eye on the long term, however, must be concerned with which formats will endure and remain playable as time passes. This raises questions such as "Is the file migratable?", "Is the file in a format for which we can expect playback-system emulations in the future?", and "Does the archive have a system for normalizing digital content into a form that the archive proposes to maintain for the long term, and can this element be normalized into an appropriate form?".

Regarding born-digital files acquired by an archive, the format question is challenging. For example, if a RealAudio stream is captured from a webcast, can the capturing archive count on the continued existence of playback software and/or emulations for the long term, or should this bitstream be reformatted into a different structure, such as a PCM rendering, in hope of increasing the likelihood of longterm playability? Will this kind of digital reformatting produce audio artifacts that mar the listen-ability of the recording? Is there a normalization strategy that may be helpful? Questions like these animate many digital library community discussions at this time; the spoken language community can contribute to this broader investigation by means of applications research or demonstration projects devoted to its particular type of content.

[5] The survey identified on the order of ten million hours of sound recordings of all types in Europe, http://presto.joanneum.ac.at/projects.asp#d2

[6] http://www.archive.org

3.3 *Metadata*

Metadata includes *bibliographic* or *descriptive* information, generally defined as providing the names of works and their creators, information about the physical manifestation of the work (such as format, publisher, date, rights), and subject language, for instance terms that describe what a document is about.[7]

Two additional types of metadata have been identified by those who create or manage digital objects (digital manifestations of a work). *Structural metadata* consists of information about the structure and organization of a multipart digital object. For instance, a spoken word digital object (or collection of objects) may contain multiple files or bitstreams as well as transcripts (and in some cases, images), and structural metadata will express their relationships, such as the sequencing of a series of audio segments, the correspondence of sound with textual transcripts, and so on. The second additional type of digital-object metadata (which may overlap with bibliographic information) is *administrative metadata*, which includes detailed technical information about bitstream encoding, specialized rights metadata, and provenance information about the object's history. Some commentators consider transcriptions of spoken language recordings to be metadata because they support searching within a corpus of information. Transcripts may be conceptually a part of a single digital object together with the sound recording. In other cases, they may be defined as "works of their own," a phenomenon strikingly represented by the US Congressional Record.

3.4 *Packaging*

The structure of the Open Archival Information System (OAIS) reference model expresses phases of the digital content life cycle.[8] A key feature of the OAIS model is the content or information package, conceived of as an object that bundles data and metadata for the sake of content management. Content packages include files or bitstreams that represent the content (eg a WAVE file), metadata, and encapsulation schemes (eg Unix tar files). Formats that bundle these content elements together include MPEG–21[9], MXF[10](Media Exchange Format), METS[11] (Metadata Encoding and Transmission Standard), and others. The unresolved aspects of packaging standards do not require laboratory research but rather the establishment of conventions to aid in the preparation and structuring of content. These conventions and associated practices are essential to the practice of archiving and thus to the long-term availability of spoken language content.

At a higher level, archivists and librarians follow a number of standards and guidelines pertaining to digital content, including the practical guidance provided by the International Association of Sound and Audiovisual Archives (IASA) [1].

Regarding standardized representations of transcripts, it is worth noting the existence of the MPEG–7 Spoken Content Description Scheme, which addresses speech recognition output (including support for alternative word choices).[12] At present, this standard has not been widely adopted in the spoken language processing community. Other alternatives for representing the synchronization of sound and transcribed text include proposed W3C standards such as SMIL[13] and EMMA.[14] Meanwhile, humanities scholars have invested much effort in developing the Text Encoding Initiative (TEI) markup language and associated conventions in order to exchange textual renderings of printed and written documents.[15]

Packaging raises a number of unresolved questions such as: Considering that many of the spoken language recordings with enduring interest to society, for instance oral histories, are important documents for the humanities, there may be merit in investigating an expansion of the TEI schemes to spoken language transcripts, including a recommended approach for indicating elapsed time? Are there actions or conventions that will make transcripts usable by researchers from multiple disciplines? For example, will renderings that feature the "annotation graphs" employed by workers in the spoken language processing and speech recognition communities be comprehensible or helpful to oral historians, folklorists or cultural anthropologists? Or what might specialists in the latter fields do to make their content more useful to the spoken language processing and speech recognition communities?

Finally, we ask if there are there special classes of metadata pertaining to content of interest to the spoken language community not addressed by any other standards? Is there a process that will describe these classes and take the actions that may be needed to establish a standard? To what degree will practices in this area address the concerns about the portability and hence the permanence of language data expressed by Bird and Simons [6]?

3.5 *Sustainability*

Digital content in technical terms is sustainable. Sustainability in financial terms, however, is another matter. It is a focus of concern, although not specifically within the group's expertise. What business case can be made to support the existence of a spoken language archive? We note the following aspects to this topic:

- The acquisition and archiving of content of evident social value should be supported by society, ie, government

[7] The term "metadata" is often used in the speech recognition community to refer to content annotations (eg the NIST rich transcription evaluation, http://www.nist.gov/speech/tests/rt/); we do not use the term in this way.

[8] http://ssdoo.gsfc.nasa.gov/nost/isoas/

[9] http://www.chiariglione.org/mpeg/standards/mpeg-21/mpeg-21.htm

[10] http://www.g-fors.com

[11] http://www.loc.gov/standards/mets/

[12] http://www.chiariglione.org/mpeg/standards/mpeg-7/mpeg-7.htm

[13] http://www.w3.org/AudioVideo/

[14] http://www.w3.org/TR/emma/

[15] http://www.tei-c.org/TEI

libraries and archives that are funded by taxes. This includes content of interest to scholarship, like oral histories, selected lectures by academics, judicial proceedings, and the records of government, for example recordings of the deliberations of political bodies.
- Content owners will sponsor content of commercial interest. For example, many broadcast recordings have continuing value in commerce, as material for rebroadcast or for sale to others engaged in program production. This content is likely to be preserved, although society should encourage this preservation and stand by to receive material when commercial interests retire it.
- Content of interest to the law enforcement and national security communities will be archived by its members and in some cases may pass into public archives in the future. Some of this content has evident social value. The business case for longterm preservation should be considered.
- The generation of content intended for specific spoken language research purposes is often funded for the purposes of that research. Some of this content has evident social value, and the business case for long-term preservation should be considered.

4 Policy

Collectors of spoken word audio materials must address a number of complex privacy and copyright issues relating to the collection, retention and distribution of works. These policy issues cannot be ignored, but the legal frameworks that define them offer incomplete and sometimes conflicting guidance. Privacy and copyright are two of the most rapidly changing aspects of United States and EU law. We provide a brief analysis of some key issues.

4.1 Privacy

Privacy is not a precisely defined concept. The issues of data and communications privacy have been very widely debated, across both the US and the EU. Less commonly discussed aspects of privacy may be equally relevant to a spoken-word archive. For example, what are the legal implications of recording a public meeting?

Some issues surrounding audio and video capture in public are not dissimilar to those debated when face-recognition technology began to be used to scan for potential criminals in crowds at airports and other public places [10]. Here, the expectation of privacy is one of anonymity, but this expectation is not always codified in law. Several US state courts have resisted attempts to curtail video and audio recording in public, finding that no reasonable expectation of privacy can exist in a public place [27]. Use of recording technologies for public surveillance in the United Kingdom has been common for some years, though the government in 2000 signaled its intention to regulate such surveillance in accordance with its 1998 Data Protection Act, passed to harmonize U.K. laws with the 1995 European Union Data Protection Directive.[16] Other EU nations, including Greece and Sweden, also interpret the EU Directive (revised in 1998 and 2000) to specifically pertain to public video surveillance and closely regulate its use.

Open monitoring and recording of telephone transactions and monitoring of employees' electronic communications for business purposes is also widespread [10]. The right of employees to opt out of such data gathering has been weak or non-existent. The EU is leading the push to expand data privacy regulations to include employee-monitoring activities, which may have the effect of discouraging such monitoring beyond the EU [13]. Most European Union nations have appointed a central data protection agency, charged with oversight of all personal data collection and processing, and grant individual citizens a mechanism for review, change or removal of their own information.

Given the need for oversight and the ease of access to such information once stored in digital form, some difficult choices face the custodian. What balance should be struck between protection of the individual and benefits of large spoken word collections for worthy public purposes, such as scholarly inquiry, political discourse, law enforcement, artistic expression? The regulations governing research on human subjects, which clearly advocate informed consent and limited gathering and use of personal data, may offer guidance in this case.

Collecting agencies should determine whether individuals have granted permission for a recording to be made, implicitly or explicitly. A signed consent form or recorded consent are the best safeguards, but may not always be available. Presenters and announcers, interviewers and interviewees, audience members and call-in guests, parties in a conversation: all such participants must be considered when determining whether privacy rights are an issue. A public figure, such as a politician or a known lecturer, is unlikely to substantiate an invasion of privacy claim were his speech to be recorded. The more public the citizen, the less likely he or she is to be able to make a claim.

4.2 Copyright

When providing access to spoken word materials the principal issues are whether the materials are protected by copyright, whether auxiliary rights must be taken into consideration when archiving digitally, and, all rights notwithstanding, whether an argument can be made to proceed with providing access.

Copyright legislation has changed dramatically over the past decade, both in the United States and in Europe. The rise and demise of Napster and other online fileswapping services have focused the attention of the technology, content, legal and consumer advocacy communities on the issue of digital audio distribution. Despite this attention and debate, clear rules have failed to emerge, and are unlikely to surface in the

[16] CCTV code of practice at http://www.dataprotection.gov.uk

near term, particularly for non-commercial use by libraries and archives.

As signatories to the Berne convention (as revised in 1971, and subsequently amended) [34], the United States and the European Union member nations have reciprocity in copyright protection so that materials created or published in one nation will, for the most part, enjoy the same protections in other nations. Copyright statutes generally reserve for the copyright holder the exclusive right to reproduce, display, distribute copies of, and perform or broadcast the work. The European Union issued a copyright directive in 2001 that matches many of the provisions in the United States Digital Millennium Copyright Act (DMCA) of 1998. Both extend encryption protections with harsh anti-circumvention language. The results of this implementation do not yet offer clarity or guidance.

In general, sound recordings have historically been accorded fewer protections than other types of works, though some recent initiatives have the effect of increasing their protection.[17] In the United States, sound recordings were not protected by federal copyright law until 1972, and recordings made before that date are still not federally protected (although they may be under state copyright laws). Works fixed after 1977 receive at least 70 years of protection; in the European Union there is a 50 year duration of copyright for sound recordings.

There may be layers of authorship embedded in a single sound recording, and each act of authorship may be subject to separate protection. For a musical work, the composition and arrangement might both be protected even if the physical recording itself is not. A more relevant example of layered rights may be seen in observing several separate acts of creation that might be said to be encompassed within a sound recording of a news broadcast: a typescript, background music, and interviews with news subjects. It is unclear how stringently these protections will be pursued and enforced.

The Berne convention (article 2bis) suggests that signatory states may wish to exempt certain works from copyright protection, such as political speeches, legal proceedings and public lectures [34].

Although several countries mandate or encourage legal deposit, it is not true that physical ownership confers ownership of the underlying intellectual content, unless a deed of gift or some other condition of acquisition explicitly transfers copyright along with the physical artifact. Therefore, even though national and depository libraries and archives have wonderful, unique and precious audio collections at their disposal, they must look carefully at exemptions in the copyright law before providing access, for most audio content is likely to be subject, in some degree, to copyright protections.

The "fair use" clauses are a natural starting point. Most countries have made some provision for reproducing copyrighted works for certain purposes. Those specifically mentioned include teaching, criticism, news reporting, and parody. In all cases, the language of the copyright law is non-specific as to the particulars. The United States copyright law's fair use clause cites "amount and substantiality of the portion used in relation to the copyrighted work as a whole" as a factor, but states that it must be balanced along with three other factors and offers no specifics about what a "substantial portion" might be.[18] In practice, fair use clauses are problematic. Their vagueness has led to self-censorship in many domains, including education, entertainment, and publishing. The content community has been successful in characterizing fair use as an archaic loophole [20].

Specifics of copyright law vary from country to country, even among Berne convention signatories. It may be that a productive collaborative activity will be to establish some reasonable "acceptable risk" policies and practices, which need not be overly concerned with a narrow reading of any one copyright statute. As an inspiring example, the Australian National Archives recently decided to digitize archival materials and make them freely available, regardless of copyright status, to help overcome the "tyranny of distance" [18].

4.3 Moral rights

In addition to the set of rights recognized in copyright laws, other rights may come into question with audio archiving projects. Among these are so-called "moral rights," those that allow the creator of a work some lasting ability to control the context in which it is used and how (or whether) authorship is attributed. Generally, copyright governs economic rights, but moral rights are less tangible and involve the integrity of a work [29].

Moral rights are established in the copyright laws of several countries, but are not universally supported and protected. The United States, for example, grants rights of attribution and integrity only to authors of works of visual art, and extends them to the end of the author's natural life. However, German and French copyright law extend these moral rights to authors of all works and allow them to be transferred to heirs. Moral rights allow the author to associate or disassociate herself from works, including derivative works, and, in the case of French law, to prevent release of or removal from public availability already published works. It is possible that moral rights will play a role in evaluating spoken word collections, particularly in the cases of unscripted or extemporaneous speech in oral histories, interviews, meetings, and the like, where it is perhaps more likely that a subject will wish to retract or withdraw.

Archives will set local policy based on the laws governing their country and the legal preferences of the parent institution, if any. A standard list of questions to ask when considering whether or not a risk can be managed [2] include the age of the material, whether it was produced for commercial purposes, whether any rights are transferred by consent forms, whether access can be brokered to reduce concerns

[17] http://homepages.law.asu.edu/∼dkarjala/OpposingCopyrightExtension/legmats/HarmonizationChartDSK.html

[18] Title 17, US Code, Chapter 1, Section 107

about worldwide distribution (eg “thumbnail” equivalents), and whether digital liability insurance be obtained.

5 A Research Agenda

We have structured the research agenda for spoken word archiving into three principal areas: technology; privacy and copyright; and archiving and access. The main priority is to advance each area individually, and to foster integration among them. It is clear that each area informs the others.

5.1 Technology

Audio/signal processing: Many spoken-word collections of interest, particularly historical collections, have deteriorating audio, due to media degradation or imperfect analog recording technology. Other audio signal processing challenges arise from multiple overlapping speakers, as found in meetings, low signal quality due to far-field microphones, as found in courtrooms, and effects of other sound sources and room acoustics.

Speech and speaker recognition: Any spoken audio collection raises two immediate questions: (a) What was said? (b) Who said it? Speech and speaker recognition technologies now work to minimally acceptable levels in controlled domains such as broadcast news. However, there is a large gap between machine and human performance [19] and it is well acknowledged that improvements are needed in the modeling techniques at all levels: acoustic, lexical and pronunciation, and linguistic (syntactic and semantic). Achieving substantial improvements will require new tools to address less controlled collections of spoken audio. Without such tools, the costs in labor to access spoken-word collections will be prohibitive. The creation of these tools also enables the hearing-impaired public to access and use these materials.

Multilinguality: The universal accessibility of spoken language technologies depends on porting to languages beyond those on which current technologies focus. There is a lack of coverage for many languages and the collection and management of linguistic resources is required. Further, in a multilingual context, automatic language identification is essential. In particular many collections (eg, meetings at a European level, some oral narratives) feature speakers switching between different languages. It is possible to construct adequate baseline systems based on current knowledge, but issues such as within-utterance language change pose interesting and challenging research problems. Finally particular research challenges are raised by those languages that are unlikely to become economically important, may be endangered and may have no written form.

Content annotation: The use of a spoken-word collection can be enhanced by the automatic generation of content annotations. The automatic identification of names and numbers, and punctuation has been demonstrated. However, it would be advantageous to annotate many other elements particularly paralinguistic features such as attitude, style, emotion and accent, discourse features, and features such as decision points in meetings, and interaction patterns in a conversation. At the outset, such annotation must be done manually, and tools have already been developed for many of these annotation tasks. New research must be undertaken to develop the data models and coding schemes for these new annotation types. For scalability, new tools should support both collaborative annotation, in which networked colleagues share the task and quickly resolve questions about the correct annotation; and mixed initiative annotation, in which the system observes the work of the human annotator and gets better at suggesting the correct decision. Once these annotated corpora have reached sufficient size they can then be used to train fully automatic annotation software.

Information access technology: We know quite a lot about supporting access to broadcast news, but far less about how best to support access to extended sessions of spontaneous speech. There is also a need for focused assessment of the needs of specific user groups that to date have been understudied. Some examples include teachers and students, scholars in the humanities and social sciences, and individuals employing personalization and memory augmentation systems.

Presentation: The final technological research area that we have identified is presentation. Currently this involves little more than playing an audio clip and displaying its transcription. There is an enormous need for research in this area, for instance the construction of audio scenes, presentation of higher-level structure, summarization, and presentation of non-lexical information in speech.

5.2 Content Management

Acquisition: Speech is an ephemeral medium, and at present much of what we might wish to have access to in the future is not being captured. Content capture is often an incidental process; the Internet Archive evolved from an early effort to index the Web, Deja News resulted from the distributed design of USENET news servers. No comparable source presently exists for capturing the millions of hours that are presently being webcast each year, however. Aging analog media pose an equally grave concern; without digitization, substantial quantities of irreplaceable content may well be lost forever before its value can even be recognized.

Preservation: Open research issues include standards for preservation and development of sustainable digital repositories. Issues that need to be addressed include: funding; automating digitization and metadata capture; migrating and refreshing/augmenting collections. Computerized automated capture and preservation of collections clearly underlies the development of this entire area.

Content structure: This area spans metadata, item structure, annotation, discovery and delivery issues, such as network bandwidth. Metadata vocabularies have been developed, but this area still needs further research, particularly when the archived items have a complex structure. Additionally, metadata needs to be aggregated and services offered on the aggregated collection. Models and tools for annotation are a rapidly evolving research area, particularly in the area of distributed and collaborative annotation.

Media storage: Even with the rapidly declining costs of spinning disks, most preservation-quality audio collections will continue to require supplemental digital storage media for the raw audio files at least into the foreseeable future. Research is needed on various media (CD, DVD) and best practices for storing, checking, and refreshing.

5.3 Policy

A number of policy issues arise when discussing spoken-word collections, and it is impossible to treat the technologies in isolation from these issues.

Privacy: Privacy is a major problem, particularly for some spoken-word collections when individuals do not have an expectation that their statements will be archived, although they have spoken in a public forum such as a company board meeting or a political rally. It may not be possible to offer a comprehensive solution to the privacy problem, particularly for materials where contact with the original collector or subject has long since been lost, but research in this area can accomplish some practical goals. Future collectors must be armed with reasonable policies to obtain clearances and document applicable rights.

Copyright: The impact of copyright varies by collection, and by national jurisdiction. Because the legal terrain here is difficult to understand and is undergoing rapid change, a practical approach for cultural institutions to take may be to implement "acceptable risk" policies. These policies set forth overarching principles of respect for subjects and for the creators' intellectual property rights, but balance them against a need to provide access to important cultural heritage materials. Issues to research include: copyright exemptions (eg, for educational purposes), classes of works that do not qualify for copyright protection, digitization for preservation and mediated access, and questions collection custodians should pose to determine copyright status and likely consequences of wide availability of digital surrogates.

6 Conclusion

The digital revolution has the potential to do for spoken language what the printing press did for written language. For the first time, the spoken word can be preserved for the long term and made accessible to those far beyond hearing range and in ways that open up new possibilities for human culture. Researchers have the tools and capabilities to transform access to the spoken word, preserving an essential aspect of cultural heritage, and stimulating a diverse set of communities: speech and language technology; digital libraries and information sciences; and a wide range of user communities.

Although we represent diverse disciplines, we see convergence in the domain of spoken-word collections to address new and challenging issues. In advancing an ambitious research agenda, we envision ancillary benefits across many communities of interest: speech and language technology; software engineering; information science and digital libraries; education; and a set of diverse user communities. Progress requires integration across these areas at the international level. In our judgment, the impact will be substantial. To do any less will risk significant loss to an essential element of our collective heritage.

Acknowledgements. This work arose from the working group in Spoken Word Audio Collections, supported the EU Network of Excellence DELOS, and the National Science Foundation of the USA. The working group was co-chaired by Jerry Goldman and Steve Renals, and met twice in Evanston IL, USA (June 2002) and in Delft, NL (September 2002). In addition to the authors of this paper, Fabrizio Sebastiani (ISTI-CNR) was also a member of the working group.

References

1. The safeguarding of the audio heritage: Ethics, principles and preservation strategy, February 1997. IASA-TC 03 Version 1.
2. Risk management suggestions. In *Multimedia and Web Strategist*, volume 5. January 1999.
3. D. Appelt and D. Martin. Named entity recognition in speech: Approach and results using the TextPro system. In *Proc DARPA Broadcast News Workshop*, pages 51–54, 1999.
4. B. Arons. SpeechSkimmer: A systen for interactively skimming recorded speech. *ACM Trans. on Computer-Human Interaction*, 4:3–38, 1997.
5. S. Bird and J. Harrington, guest editors. Special issue on speech annotation and corpus tools. *Speech Communication*, 33(1-2), 2001.
6. S. Bird and G. Simons. Seven dimensions of portability for language documentation and description. *Language*, 79:557–582, 2003.
7. J. P. Campbell, Jr. Speaker recognition: a tutorial. *Proceedings of the IEEE*, 85:1437–1462, 1997.
8. S. Chen and P. S. Gopalakrishnan. Clustering via the Bayesian Information Criterion with applications in speech recognition. In *Proceedings of IEEE ICASSP-98*, pages 645–648, 1998.
9. C. M. Christensen. *The Innovator's Dilemma*. Harvard Business School Press, Boston, 1997.
10. Electronic Privacy Information Center (EPIC) and Privacy International. *Privacy and Human Rights 2002*, 2002.
11. J. S. Garofolo, C. G. P. Auzanne, and E. M. Voorhees. The TREC spoken document retrieval track: A success story. In *Proc. RIAO 2000*, 2000.
12. J.-L. Gauvain and L. Lamel. Large-vocabulary continuous speech recognition: advances and applications. *Proceedings of the IEEE*, 88:1181–1200, 2000.

13. R. Glover and A. Worlton. Trans-national employers must harmonize conflicting privacy rules. In *The Metropolitan Corporate Counsel*, page 20. Mid-atlantic edition, November 2002.
14. S. J. Godsill and P. J. W. Rayner. A Bayesian approach to the restoration of degraded audio signals. *IEEE Transations on Speech and Audio Processing*, 3:267–278, 1995.
15. Y. Gotoh and S. Renals. Information extraction from broadcast news. *Philosophical Transactions of the Royal Society of London, Series A*, 358:1295–1310, 2000.
16. C. Hori, S. Furui, R. Malkin, H. Yu, and A. Waibel. A statistical approach for automatic speech summarization. *EURASIP Journal on Applied Signal Processing*, 2:128–139, 2003.
17. C. Lagoze and H. Van de Sompel. The Open Archives Initiative: Building a low-barrier interoperability framework. In *Proceedings of the First ACM/IEEE-CS Joint Conference on Digital Libraries*, pages 54–62, 2001.
18. T. Ling. Why the archive introduced digitisation on demand. *RLG Diginews*, 6(4), August 2002.
19. R. P. Lippmann. Speech recognition by machines and humans. *Speech Communication*, 22(1):1–15, 1997.
20. J. Litman. *Digital Copyright*, page 84. Prometheus Books, Amherst, NY, 2001.
21. B. Logan and T. Robinson. Adaptive model-based speech enhancement. *Speech Communication*, 34:351–368, 2001.
22. J. Makhoul, F. Kubala, T. Leek, D. Liu, L. Nguyen, R. Schwartz, and A. Srivastava. Speech and language technologies for audio indexing and retrieval. *Proceedings of the IEEE*, 88:1338–1353, 2000.
23. M. Maybury, guest editor. Special issue on news on demand. *Communications of the ACM*, 43(2), February 2000.
24. D. W. Oard. Serving users in many languages: Cross-language information retrieval. *D-Lib Magazine*, 1997.
25. D. W. Oard. User interface design for speech-based retrieval. *Bulletin of the American Society for Information Science*, 26(5):20–22, 2000.
26. G. Rigoll. The ALERT system: Advanced broadcast speech recognition technology for selective dissemination of multimedia information. In *IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 301–306, 2001.
27. L. E. Rothenberg. Rethinking privacy: peeping toms, video voyeurs and failure of the criminal law to recognize a reasonable expectiation of privacy in the public space. *American University Law Review*, 49:1127, 2000.
28. G. Simons and S. Bird. Building an Open Language Archives Community on the OAI foundation. *Library Hi Tech*, 21:210–218, 2003.
29. M. T. Sundara Rajan. Moral rights and copyright harmonization: prospects for an “international moral right”. In *17th BILETA Annual Conference*, April 2002.
30. H. D. Wactlar, T. Kanade, M. A. Smith, and S. M. Stevens. Intelligent access to digital video: Informedia project. *IEEE Computer*, 29(5):46–53, May 1996.
31. W. Wahlster, editor. *Verbmobil: Foudations of speech-to-speech translation*. Springer Verlag, Berlin, Germany, 2000.
32. C. Wayne. Multilingual topic detection and tracking: Successful research enabled by corpora and evaluation. In *Language Resources and Evaluation Conference (LREC)*, pages 1487–1494, 2000.
33. S. Whittaker, J. Hirschberg, J. Choi, D. Hindle, F. Pereira, and A. Singhal. SCAN: designing and evaluating user interfaces to support retrieval from speech archives. In *Proceedings of ACM SIGIR-99 Conference on Research and Development in Information Retrieval*, pages 26–33, 1999.
34. World Intellectual Property Organization (WIPO). *Berne Convention for the Protection of Literary and Artistic Works*. http://www.wipo.int/treaties/ip/berne/.
35. S. Young. A review of large-vocabulary continuous-speech recognition. *IEEE Signal Processing Magazine*, 13(5):45–57, 1996.