Supporting Access to Large Digital Oral History Archives

Samuel Gustman,[1] Dagobert Soergel,[2] Douglas Oard,[3]
William Byrne,[4] Michael Picheny,[5] Bhuvana Ramabhadran,[5] and Douglas Greenberg[1]

ABSTRACT

This paper describes our experience with the creation, indexing, and provision of access to a very large archive of videotaped oral histories – 116,000 hours of digitized interviews in 32 languages from 52,000 survivors, liberators, rescuers, and witnesses of the Nazi Holocaust. It goes on to identify a set of critical research issues that must be addressed if we are to provide full and detailed access to collections of this size: issues in user requirement studies, automatic speech recognition, automatic classification, segmentation, summarization, retrieval, and user interfaces. The paper ends by inviting others to discuss use of these materials in their own research.

Categories and Subject descriptors

H3.3 [**Information Systems**]: Information Storage and Retrieval, H3.7 Digital Libraries, I2.7 [**Computing Methodologies**]: Natural Language Processing – *speech recognition and synthesesis*

General Terms.

Design

Keywords

Cataloging, Oral history, Research agenda

INTRODUCTION

Using the very large digital oral history archive created by the Shoah Foundation as an example, this paper identifies issues of access to such large collections of video data, outlines a general research agenda for which this archive can serve as an excellent test bed, and invites others to discuss use of this test bed in their research.

In 1994, after releasing Schindler's List, Steven Spielberg was approached by many survivors who wanted him to listen to their stories of the Holocaust. Spielberg decided to start the Survivors of the Shoah Visual History Foundation (VHF) so that as many survivors as possible could tell their stories and have them saved, resulting in a collection that could be used to teach about the horrors of intolerance. His original vision had the VHF performing four tasks:

1) collecting and preserving survivor and witness testimony of the Holocaust;
2) cataloging those testimonies to make them available;
3) disseminating the testimonies for educational purposes to fight intolerance;
4) enabling others to collect testimonies of other atrocities and historical events or perhaps do so itself.

Today the VHF has completed part one of this vision, collecting almost 52,000 testimonies (116,000 hours of video) in 32 languages to form a 180 terabyte digital library of MPEG-1 video. Work on the second goal is in progress: Extensive human cataloging (giving clip boundaries, clip summaries, and descriptors) has been completed for over 3,000 testimonies. Streamlined human cataloging (giving time-aligned descriptors) for the bulk of the collection is scheduled to extend over the next five years. VHF has also taken initial steps towards accomplishing the third goal: eight documentaries, two CDROMs, several museum exhibits, and one book have been created from the archive to date for educational purposes. VHF has made substantial progress towards realizing the fourth goal as well by developing collection techniques, a digitization workflow, and support for human cataloging, and testing these techniques on a large scale. A database-oriented search system designed to support intermediated access to the collection is also available. This paper describes the architecture of the present system and identifies some research challenges and approaches for meeting them.

The very large collection with its cataloging data provides an excellent set of training data for the development of automated text processing and classification methods, leading to significant advances in automated and computer-assisted methods for providing access to oral history archives. Transcription of videotapes in several languages is underway to produce training sets for automatic speech recognition. There are users of many different types, from historians to teachers to the makers of documentaries, who are anxious to use this collection for many different purposes, offering rich material for user requirements and usability studies. The ultimate goal is not only to improve access to the Shoah foundation collection but to develop a set of tools that will be useful for other oral history collections and audio materials generally. Each institution can then select the tools that will best match its goals, philosophy of access, and economic constraints.

JCDL'02, July 13-17, 2002, Portland, Oregon, USA.

[1] Survivors of the Shoah Visual History Foundation, P.O. Box 3168, Los Angeles, CA 90078-3168, (sam,doug)@vhf.org

[2] College of Information Studies, University of Maryland, ds52@umail.umd.edu

[3] College of Information Studies and Institute for Advanced Computer Studies, University of Maryland, oard@glue.umd.edu

[4] Center for Language and Speech Processing, Johns Hopkins University, byrne@jhu.edu

[5] Human Language Technologies Group, IBM T.J. Watson Research Laboratory, (bhuvana,pichney)@us.ibm.com

SYSTEM ARCHITECTURE

Figure 1 shows the architecture of the VHF system. There are two main data flows: The actual video data and the metadata. The Production Database supported the logistic tasks of identifying survivors and scheduling and conducting the interviews. The Physical Tape Management Database supports the process of digitization and tracking the tapes as they are used for cataloging and other purposes. The Foundation Central Database serves as the repository of metadata derived from the pre-interview questionnaire (PIQ) and cataloging and possibly data on user projects. The ADIC 400 Terabyte Tape Archive and the Foundation Central Database serve as the backbone for physical and intellectual access to the testimonies.

COLLECTION, DIGITIZATION AND PRESERVATION

VHF developed an intensive campaign to contact Holocaust survivors and others who might wish to provide a permanent testimony of their experiences in a videotaped interview. To conduct the interviews, VHF established a world-wide network of coordinators; they had access to scheduling systems and databases containing interviewee, interviewer, and videographer contact information. Once an interview was scheduled, the interviewer would call or visit the interviewee beforehand and take them through the Pre-Interview Questionnaire (PIQ), which became a pivotal piece of data. The original intention of the questionnaire was to give the interviewer a chance to review the interviewee's experience beforehand so that the interviewer would have the opportunity to research any topics specific to the interviewee's experience before the interview. Now the PIQs are used as a structured, searchable description of the testimonies – important testimony level metadata. These PIQ data are also heavily integrated with the cataloging of the actual video, as discussed below.

Key sections of the PIQ

Survivor information (name variants, vital dates, languages, education, occupations, military service, political identity, religious identity)

Prewar life (prewar address, affiliations)

War time (ghettos, camps, hiding, resistance, refugees, death marches)

Postwar (displaced persons camps)

Family background (parents, children, relatives., data about them)

Interviews were conducted in 57 countries, typically in a survivor's home. Interviews were structured to cover prewar, wartime, and postwar experiences in a ratio of 20:60:20. The average duration of a testimony is just over two hours. Both participants typically wore lapel microphones (crucial for obtaining recording quality suitable for automatic speech recognition) that were recorded on separate channels, but only the person being interviewed appeared on camera. The video is typically a head-and-shoulders shot, but at the conclusion of the interview survivors were offered an opportunity to display any artifacts they wished to have recorded.

The interviews were recorded on Sony Beta SP, the most common tape stock used internationally. Once the Beta SP tapes and PIQ were returned to the VHF, both were digitized, the PIQ into a TIFF file and the testimony into a 3 MB/sec MPEG-1 stream with 128 kb/sec (44 kHz) stereo audio. Three factors led to the choice of this standard in 1996 when digitization started:

1) It gives an acceptable picture on an ordinary TV set.
2) There were many tools for working with MPEG-1.
3) The widespread adoption of MPEG-1 at the time provided reasonable assurance that future versions of the standard would maintain backward compatibility.

Half of the testimonies have been digitized to date; once digitization is completed, the video will occupy 180 Terabytes. During digitization, a preservation copy (in Digital Betacam format, for offsite storage) and two VHS access copies (one for use in the VHF and one for the interviewee) were created simultaneously. Over 650,000 physical tapes and other physical objects (e.g., PIQs) are managed in the archive.

MANUAL CATALOGING

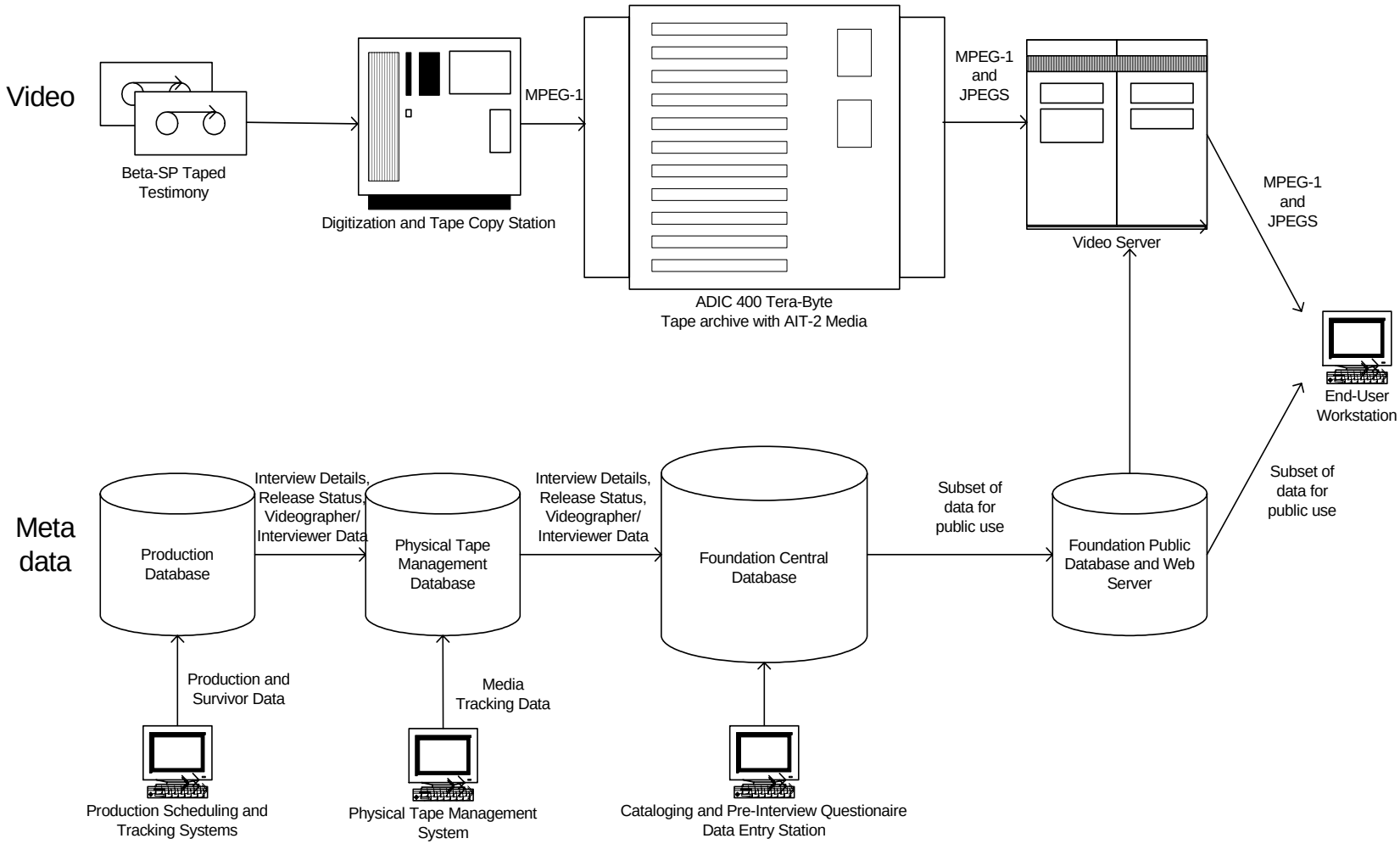
In 1998, with the collection phase coming to a close, VHF began large-scale cataloging by human catalogers. Data extracted from the PIQ provide detailed testimony-level metadata, but rapid content-based access within long linear narratives requires passage-level metadata as well. We defined a complex object (Booch 1994), called a "clip," with the following structure:

1) in and out time codes that define the boundaries of a topically coherent clip;
2) a set of descriptors, each of which identifies an event, time period, and/or location that is described in the clip or a concept for which the clip is relevant;
3) a set of person objects, either newly created or from the PIQ, one for each person mentioned in the clip;
4) a structured written clip summary;
5) objects such as maps, still images, and other clips (e.g., from documentary video) that can provide additional context to the events described in the clip.

We created three additional structures to support querying these clips:

1) A thesaurus of descriptors following NISO Standard Z39.19, which accommodates the whole/part, inheritance, and associative relationships needed to support cataloging and search (see Figure 2).
2) A complex object for person that contains all names and aliases; information about the pre-war, wartime, and postwar experiences of the person; and any other information that was provided about the person in either the PIQ or the testimony.
3) A structure called *project* for storing a set of clips and optional descriptors, persons, and descriptive text, paralleling the structure of an individual clip. A project may be created by the system or by the user.

Figure 1. Survivors of the Shoah Visual History Foundation system architecture



The cataloging process populated the clips with the first four data items identified above. The in and out time codes were set by the catalogers based on their interpretation of natural breakpoints that divide the narrative into coherent and complete stories, with an average clip duration of 3.5 minutes. The cataloger also created a three-sentence summary of the clip and any necessary person objects, and linked appropriate thesaurus descriptors and persons to the clip. This process took about 15 hours for each hour of video. Over 3,000 testimonies were processed in this way, resulting in an exceptionally rich set of metadata for more than 100,000 clips that can be used as a basis for developing automated techniques. (See Figure 3 for a view of the cataloging interface.) But even with good tools, manual full-description clip-level cataloging for 116,000 hours of video would cost over $150 million!

We took two approaches to resolve the cost issue. First, we devised a new "real-time" cataloging methodology. We found that establishing clip boundaries and writing structured clip summaries consumed the lion's share of the cataloger's time, while passage-level indexing with descriptors and persons could be accomplished quickly. The new real-time cataloging system automatically creates one-minute clips to which the cataloger assigns descriptors and persons while listening to the tape (without pausing), linking descriptors and person objects with points in a testimony at a one-minute granularity. Also, the value of the clip boundaries and summaries created by the cataloggers is in question. Text summaries do not convey the emotional content of the testimonies to the user. For students, the main end-users for the VHF archive, the text summaries cannot substitute for the full video because it is the emotional content of the testimonies that vividly shows the horrors of intolerance and thus gets messages of tolerance across to students. For the same reason it is questionable whether summaries provide a sufficient basis for selection of testimonies for use in the classroom.

Our second approach to more efficient cataloging is to look into fully or partially automate cataloging. A consortium of the VHF, the IBM Thomas J. Watson Research Center, Johns Hopkins University, and the University of Maryland will work on this problem (see below).

ACCESS ARCHITECTURE

Search

Search takes place at two interacting levels: (1) the whole-testimony-level supported by PIQ data and (2) the within-testimony level supported by cataloging data, which enabling both browsing within testimonies and retrieval access to specific places within testimonies.

The metadata extracted from the PIQ supports direct access to individual testimonies. All questions in the PIQ are answered with either persons, thesaurus descriptors, dates, or Boolean combinations of those elements. This testimony-level metadata is stored in the same Sybase database that contains clip-level and project-level metadata. Because all three structures reference the same thesaurus and person database, the user can search the testimony metadata to find a set of interesting testimonies, then search the clip metadata within those testimonies to find the specific passages that are of interest. Similarly, the user can navigate between clips and entire testimonies or projects, based on common metadata attributes or on metadata relationships encoded in the thesaurus.

Content Delivery

Video content delivery is presently provided over dedicated OC3 lines, using an EMC server with a one terabyte local disk cache and a 400 terabyte tape robot. If a requested video does not exist on the local disk cache, the server looks for it on any networked disk caches. If that too fails, the video is downloaded to local cache from the tape robot. Because each tape is a serial device, tape access time depends on the position of the file on the tape that contains it. Sub-second access is typical for access from disk, while 5-10 minute latencies are typical for access on tape. Standard software such as Windows MediaPlayer can be used to display the retrieved MPEG-1 clips.

A RESEARCH AGENDA

Introduction

The previous sections have described how, given sufficient resources, human cataloging of videotaped oral history testimonies can produce rich and useful metadata. While such effort is economically feasible for specialized or partial collections, for very large collections, such as the Shoah Foundation collection, the resource requirements are staggering, particularly if dividing videos into content-based clips and providing summaries is appropriate to a collection and its uses. Some archives may not be able to do any kind of detailed cataloging, such as assigning passage-level descriptors. The solution to this problem must be sought in advances in automatic speech recognition (ASR); automatic or computer-assisted classification/ categorization; automatic or computer-assisted segmentation and summarization, as appropriate; and in database and retrieval technology.

The VHF archive and similar large oral history archives present a number of challenges for these methods:

- Speech that is often difficult even for a human listener: spontaneous, emotional, disfluent speech; whispered speech; heavily accented speech; speech from elders; speech with background noise and frequent interruptions; speech that switches between languages; words, such as names, obscure locations, unknown events, etc., that are outside the recognizer lexicon.
- Clips, which may be a useful unit of retrieval, may be hard to delineate – topic boundaries may be ambiguous (as opposed to clips in news recordings, for example). Often interviewees jump back and forth between stories and topics, and a passage may be fully understood only after listening to a later passage in the testimony.

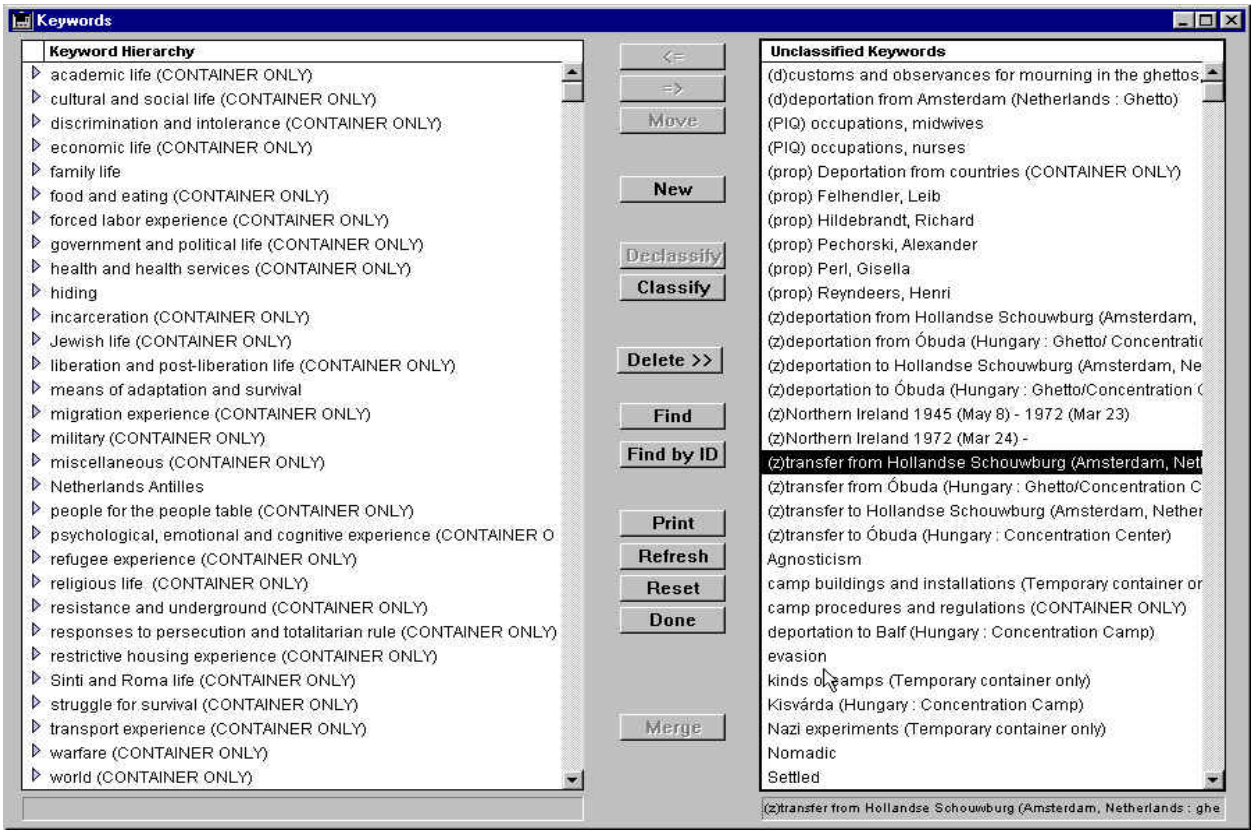


Figure 2. The VHF Thesaurus

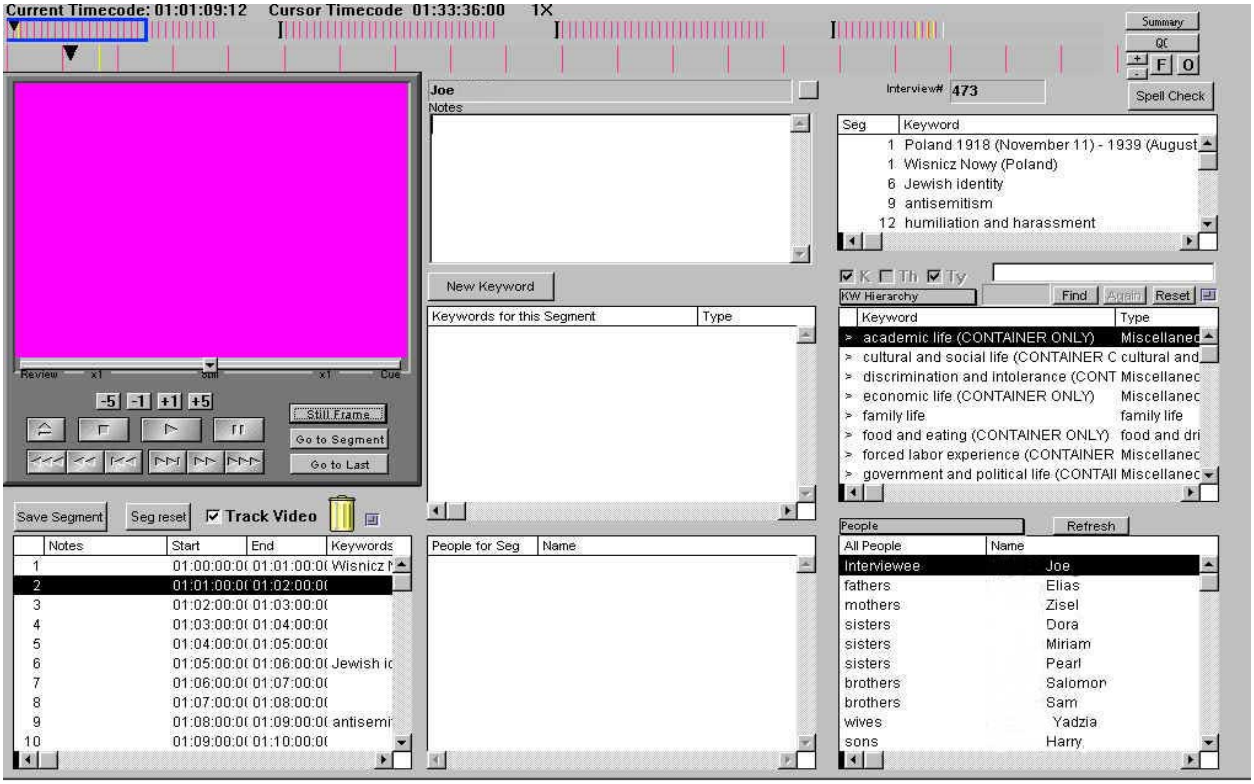


Figure 3. VHF Cataloging interface

- Users are often interested in abstract concepts, such as *Jewish-Gentile relations*, *reasons for post-war emigration to non-European countries*, the *psychological processing of Holocaust memories*, or *material suitable for fourth-graders*. (The VHF Thesaurus, built by subject experts, includes many such abstract concepts that support searches of this type.)

The remainder of this section gives an overview of the research issues that arise from these challenges. Figure 4 presents a potential system architecture that provides a context for the individual research questions. It is based on a sequence of processes that create different kinds of evidence that can then be used singly or in combination in retrieval algorithms and for presenting results to the user. Many of the research issues are given in the form of ideas on possible strategies that need to be explored and tested. Many of these ideas build on work done elsewhere, but the techniques must be adapted to a more challenging environment. In this short paper we can give only illustrative references.

The introduction outlined the opportunities offered by the VHF collection to study these research issues by providing training data that support building system components and can serve as a test bed for evaluation.

User Requirements

System design should be informed by knowledge of user requirements [Bates 1996]. We should know who the potential users are and for what purposes they wish to use the materials. What kinds of search criteria or access points (person, place, time, emotional state of the interviewee, subject, etc.) and what specific subject descriptors do they need? One source for such data is the analysis of requests; VHF has records on close to 600 advanced access requests that can be analyzed. Specific data on the use of these materials by teachers are also available and will be analyzed. How do teachers use such materials for tolerance education? How do historians and students of oral history use such source materials? Some literature exists on how historians work with oral history transcripts [Ulargiu 2000], and an analysis of that literature with a focus on design implications will be useful. But little is known about how educators, makers of documentaries, students of non-material culture, historians, and others use video or audio materials that do not have transcripts. How do users search such materials? On what basis do they make relevance judgments? What metadata do they need? What, if any, differences are there between making relevance judgments on speech versus written text? How does easy access to specific places in the audio or video affect their work? Would clip summaries or a running list of time-aligned themes/narrative descriptions of what is discussed in a recording be helpful, or would summaries result in a disincentive to users to look at the recording itself and be exposed to the power of the original message? (This issue should be explored as a trade-off between time and quality; the results may well depend on the nature of the material, the quality of the clip boundaries and summaries, the nature of the question, and the characteristics of the users.) Would clip boundaries be helpful or impose a particular view on the user and distort the historical record? Would users want to establish their own clip boundaries? Would sharing such clip boundaries among users be useful? Answers to these questions require specific empirical studies, and the many users interested in the VHF archives will provide many opportunities to conduct such studies. In the context of projects that develop systems, an iterative strategy of several user study - system development cycles is possible.

Support for Cataloging, Search, and Exploration

This section presents research issues in automatic speech recognition, further processing of the ASR output for metadata creation, use of these metadata in retrieval, and user interface issues.

Automatic Speech Recognition

Automatic speech recognition (ASR) [Young 1996, Jelinek 1998] is the basis of all other text processing steps. ASR is divided into two steps, recognition of phonemes in context from the acoustic signal and derivation of terms (words and perhaps phrases) from the phonemes (usually multiple term hypotheses with probabilities attached). Both processes are driven by statistical models derived from training data: The *acoustic model* makes associations between acoustic signals and phonemes in context; it may be highly speaker-dependent. The *language model* gives probabilities of word n-grams (pairs, triples, …) and is used to generate and select word hypotheses. A class-based language model contains n-grams of which some elements refer to word classes, such as person names or place names; this is helpful in recognizing proper names, and the VHF Thesaurus can be used to obtain word-to-class mappings. The language model may be dependent on a language community (such as Polish-born survivors speaking English, the influence of Yiddish).

Processing the difficult speech in the survivor testimonies requires significant improvements to ASR, among them

- methods for acoustic segmentation – dividing the acoustic signal into segments by the categories of speech (emotional speech, speech in different languages, etc., see challenges above), since each of these segments requires a different acoustic model and may require a different language model;
- methods for rapidly adjusting the acoustic model to the speaker [Ramabhadran 2000, Zweig 2001];
- methods for optimizing the language model for retrieval by using appropriate task-dependent loss functions geared towards retrieval, for example, giving higher weight to words that are important for searching and/or for automatic classification [Goel 1999].

A special case of word recognition is named entity recognition [Kubala 1998, Franz 2000], especially recognition of personal names and place names, both important search criteria (access points) in oral history.

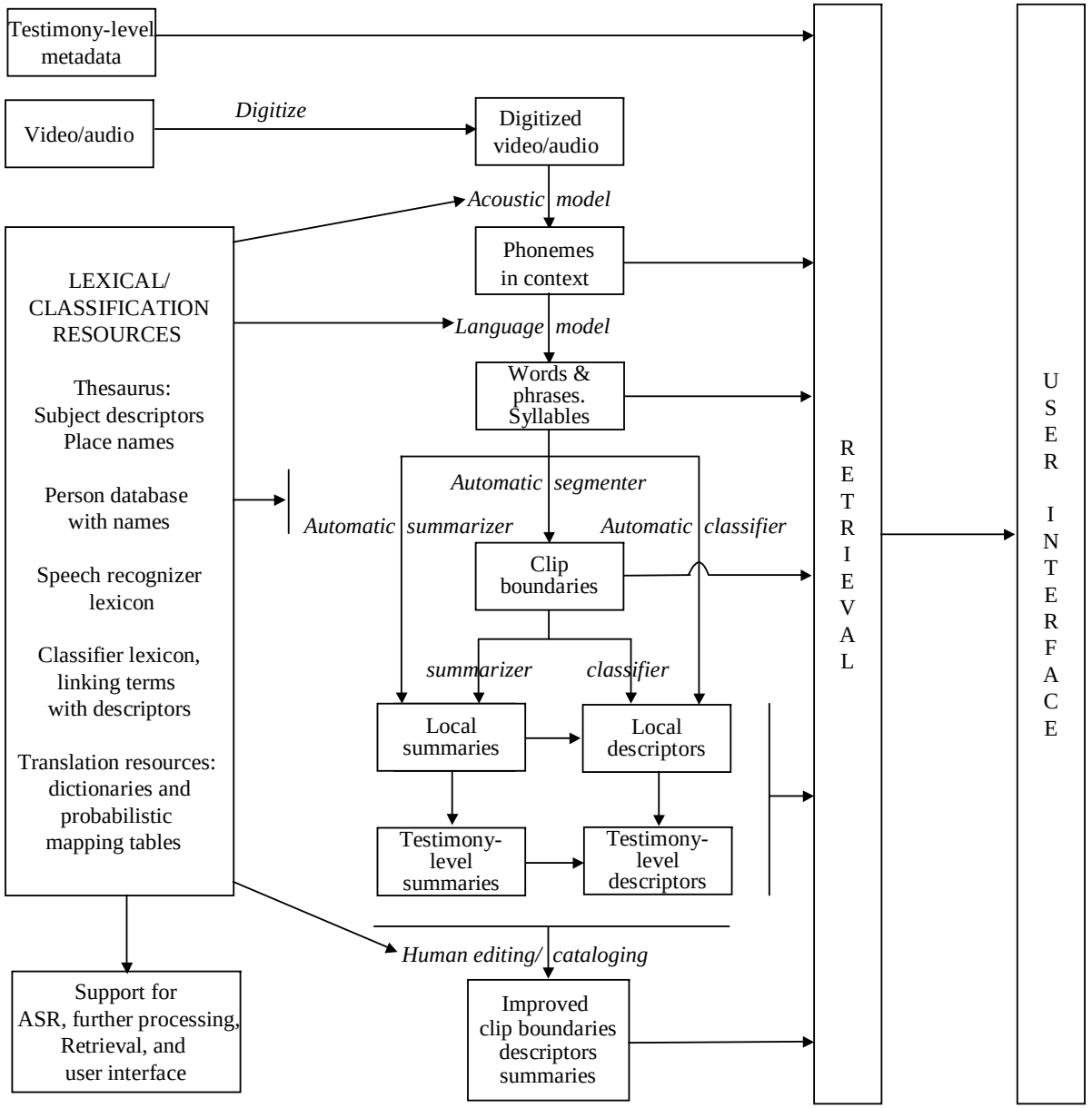


Figure 4. Architecture of an oral history information system using automatic speech recognition

This is a difficult problem since these names are often not in the speech recognizer's lexicon and not present in transcribed training data. Moreover, names come in many variations (Hebrew name, Yiddish name, diminutives, first name only). One strategy is to obtain names from a large list pertinent to the domain and derive their pronunciations automatically. The VHF collection offers an opportunity to study this problem through its large person database (ultimately around 2.5 million names) that is populated from the PIQs and additional names assigned by catalogers and its large list of places (over 20,000 locations).

Improvements in ASR might also be achieved by analysis of facial expressions and gestures; the VHF collection offers the opportunity to study this issue with videos taken in a frontal view and digitized with high resolution.

The ultimate goal of ASR is a readable transcription. A more modest goal, more realistic for difficult speech in the short run, is to produce sufficient word and phrase information for further text processing as described below.

A goal related to speech recognition is emotion detection, which would provide an alternate mode of access.

Further Text Processing

Further text processing can derive additional retrieval cues and metadata useful for presentation. There are many challenges here. We need to improve existing techniques for the automatic assignment of multiple descriptors from a thesaurus or taxonomy. Also, studies in the automatic determination of clip boundaries and automatic generation of summaries are needed to study their usefulness for different collections (see caveats above). Furthermore, assignment of time-aligned descriptors and the creation of summaries may depend on clip boundaries. To investigate these questions, collections with clip boundaries are needed as training data.

Automatic determination of clip boundaries or theme changes (automatic segmentation) is a difficult problem [Dharanipragada 2001, Johnson 1999]. One strategy is to combine data from acoustic segmentation and speaker turns with semantic methods. The idea of a scope being associated with a descriptor or term based on its category, discussed below under retrieval, might also be applicable to the determination of theme changes.

Automatic assignment of descriptors may be most useful at the clip level, even if the boundary information is used only to time-align the descriptors within a testimony. Another possibility is to simulate the present manual cataloging process by having the automatic classifier scan a testimony and assign a descriptor whenever enough evidence has been accumulated; the results can then be compared with descriptor assignment based on automatically generated clips. Testimony-level descriptors can be derived either by applying the automatic classifier to the entire transcribed text or (probably preferable) by deriving a testimony-level set of descriptors from the clip-level descriptors by a process of consolidation and abstraction.

Summaries can be formed simply as sets of descriptors. Where readable transcriptions cannot be produced, fluent summaries may not be possible. One possibility to be investigated is this: derive typical sentence templates from the training data and see whether these sentence templates can be filled from the words identified by ASR. Instead of basing summaries on clips, it might be possible to have an automatic summarizer read through a transcript and output a theme summary when enough evidence for that theme has accumulated, creating *time-aligned themes*. These, in turn, could be used as a basis for descriptor assignment. The thoroughly cataloged sample of the VHF collection can provide a baseline for experimenting with these ideas. A final question is how to create a testimony-level summary (or a project-level summary) from a set of clip-level summaries or themes.

For later application of ASR and text processing output, especially if that output is to be further edited by people, the ASR system should assign degrees of confidence to its results, such as labeling stretches of tape by the degree of difficulty or labeling ASR-generated terms or classifier-generated descriptors by the degree of confidence one should have in them. That way a human editor can focus on pieces the machine could not do well (for example, a cataloger might read stretches labeled as difficult in order to assign additional descriptors). Retrieval algorithms could also take this information into account

Manual cataloging and its interaction with text processing

The results of this automatic text processing can be used directly or can serve as a tool for assisting human cataloging, speeding up that process. Of particular interest are two questions: (1) How much time do the catalogers save? (2) How does the quality of the results compare with the quality of entirely human cataloging (where quality must be ultimately measured by retrieval performance)? Conversely, the time-aligned subject descriptors and proper names assigned in real-time cataloging can be used to improve the output of ASR and further text processing. The VHF collection provides the opportunity to study this approach as thousands of real-time-cataloged testimonies become available within the next year.

Retrieval Algorithms

ASR and subsequent text processing and/or human cataloging produce many types of evidence that can be used in retrieval: time-aligned phonemes and terms, time-aligned descriptors from the thesaurus, clip boundaries, clip summaries (either as a set of descriptors or as text) or time-aligned statements of themes. These are in addition to the testimony-level metadata available from the pre-interview questionnaires (PIQs). This opens up many possibilities for retrieval based on any of these types of evidence used singly or in combination. For example, one might use assigned descriptors to search for an abstract concept and combine it with a named-entity search for a proper name, with backoff to phoneme-based search if the named entity is not found. Due to the many languages in the collection

(with the added complication of several languages found within the same document), all these techniques must be extended to cross-language searching [Franz 2000, Oard 2001]. This includes methods for automatically or semiautomatically creating quality translations of thesauri. A further issue is the relationship between the quality of automated speech recognition, retrieval algorithms, and retrieval performance.

Retrieval may target testimonies as a whole or specific places within testimonies; a specific place could be a clip with boundaries defined based on content (a story within a testimony with a beginning and an end), a moving window of a fixed or user-definable length, or a scope with fuzzy boundaries. When there are content-based clips, one can apply simple Boolean or ranked retrieval to retrieve them. But when there are only time-aligned text terms or descriptors, AND queries (queries that combine several search criteria) require different methods. One can use time proximity search, requiring terms or descriptors to occur within a given distance (for example, five minutes).

A more sophisticated method that promises to be more effective is based on the hypothesis that descriptors of different categories have a different scope. For example, place names might have a wide scope, possibly extending to the mention of the next place name; activity descriptors may have a scope of only a few minutes. Thus, each descriptor is in force within a window that is based on its time stamp and its scope, and each place in a tape has associated with it a set of descriptors in force at that place. An AND search would retrieve all places where all required descriptors are in force. This principle can be generalized to ranked retrieval and to proximity searching. By mapping terms to thesaurus descriptors that in turn lead to categories, free-text terms can be assigned a scope.

Within-testimony data can be combined with testimony-level data; for example, a user might be interested in *reasons for immigration* (passage-level descriptor) for survivors living in *Brazil* (from the PIQ). One might even combine with characteristics found in external databases; for example, by linking from place names to a gazetteer that has information about the places one might search for pre-war experiences in medium-sized cities.

User Interfaces and Usability Testing

A key issue in user interfaces is assisting end users with formulating a good query. Several tools for query elicitation can be explored. One could display a query frame with certain categories of criteria (place, time, concrete subjects, abstract themes) to assist the user in thinking through all aspects of her query. In descriptor-based searching, users need assistance with finding the right descriptors. This can be supported by mapping free-text entry vocabulary to suggest thesaurus terms [Buckland 1999], using the existing cataloging data for training, and by providing a browsable thesaurus hierarchy to be used by itself or after descriptor candidates are found through mapping.

A second issue is assisting users in interacting with oral history data and with exploring the wider context of oral history testimonies. This includes a number of subordinate issues (see User Requirements), for example:

- What representations (descriptor lists, summaries, full transcription, full audio, full video) are available to the user? How are they used? How useful are they? [Merlino 1999] Do surrogates detract from exposure to the actual recording by giving the user the impression she does not need to examine the recording itself? If no or only limited surrogates are available, fast access to the full audio or video becomes essential, making compressed video where the entire collection can be stored in disk cache an important area of study.
- Methods for interacting with testimonies (for example, searching for all occurrences of a word).
- Linking to documents related to the testimony: documents written by the interviewee, documents based on the testimony, documents or testimonies by people mentioned in the testimony.
- Assistance to users in defining their own clips, grouping these clips into projects, assigning project-level metadata, retrieving projects. and possibly sharing these projects with others. This raises database, interface, and usability problems.
- Presenting a timeline of the events discussed in a testimony as an aid in navigation and in comparing several testimonies. Linking to relevant events in an external events database to provide context.
- Links from place names to maps and images, again to provide context.

Supporting lexical tools

ASR, human cataloging, retrieval (especially cross-language retrieval), and the user interface all require lexical tools. Examples are listed in Figure 4. Developing such tools creates resources for the whole community. How information can be exchanged and shared among these tools is a question to be explored.

Providing Global Access

Making a large oral history archive, such as the VHF collection of testimonies, available over the Web or Internet 2 presents serious policy issues. The interviewees must be protected; access to their personal information must be regulated through an authentication and authorization process. For example, a survivor may ask that his or her testimony not be shown in some countries for safety reasons or that some personal data in the testimony be protected. Technical problems aside, global authorization and authentication for access to personal data from a digital library pose problems that must be resolved before general access can be established. In the meantime, carefully selected stand-alone subsets of the archive are being made available to museums and communities that request them. Data used for research must be stringently protected from unauthorized access and kept off network.

THE MALACH PROJECT

Our organizations are collaborating on the MALACH (Multilingual Access to Large spoken ArCHives) project, which will work on a number (but by no means all) of these issues. A major focus of the project is on making significant advances in automatic speech recognition applied to difficult speech and on tightly integrating speech recognition with further text processing and retrieval.

CONCLUSION AND INVITATION

We have presented the architecture underlying the creation and operation of a very large digital oral history archive and the problems of providing specific access to such a mass of data. We then outlined a research agenda that grew out of these problems, covering issues in user requirement studies, automatic speech recognition, automatic segmentation, classification and summarization, retrieval, and user interfaces. For these research issues, we identified challenges, strategies to meet the challenges, and opportunities offered by the VHF collection which, enriched by the work of the project, provides a large amount of training data for automatic speech recognition and further text processing tasks in many languages. The research issues we identified are very hard. The solutions arrived at in the context of the difficult speech in the VHF collection will have wide applicability. There is much work to do. We invite other groups to talk to us about arrangements for using our data resources to work on these problems and push the envelope further.

ACKNOWLEDGEMENTS

The authors are grateful to Karen Jungblut, Director of Cataloging, Marina Berkovicz, Manager of Access, Kim Simon, Director of Strategic Partnerships, VHF, and Bruce Dearstyne, University of Maryland, and to many students at the University of Maryland, for sharing their insights in discussions of these problems.
This work is supported in part by NSF grant IIS-0122466

REFERENCES

http://www.vhf.org

http://www.clsp.jhu.edu/research/malach/

Bates, Marcia J., 1996. "The Getty End-User Online Searching Project in the Humanities: Report No. 6: Overview and Conclusions." *College and Research Libraries* 57 (Nov.): 514-523.

Booch, Grady. 1994. *Object-Oriented Analysis and Design with Applications.* 2. ed. Addison-Wesley.

Buckland, M., et al., 1999. Mapping Entry Vocabulary to Unfamiliar Metadata Vocabularies. *D-Lib Magazine* Vol.5 No.1 January.

Dharanipragada, S., et al., 2001. Segmentation and detection at IBM: Hybrid statistical models and two-tiered clustering. *Topic detection and Tracking: Event-Based Information Organization.* Kluwer.

Franz, M., McCarley, J. S., Roukos, S., 1999. Audio-Indexing for Broadcast News, *Proceedings of the Seventh Text Retrieval Conference*, pp. 115-119.

Franz, M., McCarley, J. S., Ward, T., 2000. Ad Hoc, Cross-Language and Spoken Document Retrieval at IBM, *Proceedings of the Eight Text Retrieval Conference*, pp. 391-398.

Goel, V. and Byrne, W., 1999. Task dependent loss functions in speech recognition: A-star search over recognition lattices. *Proc. European Conf. On Speech and Communication and Technology. V. 3, p. 1243-1246.*

Jelinek, F., 1998. *Statistical Methods for Speech Recognition*. MIT Press: Cambridge.

Johnson, S. E., Jourlin, P., Sparck Jones, K., Woodland, P. C., 1999. Spoken Document Retrieval. *The Eighth Text Retrieval Conference TREC-8*, Cambridge University, Nov. See Also http://trec.nist.gov

Kubala, F. and R. Schwartz and R. Stone and R. Weischedel,, 1998. Named entity extraction from speech, in *Proceedings of DARPA Broadcast News Transcription and Understanding Workshop*, (Lansdowne, VA), February.

Merlino, A., and Maybury, M., 1999. An Empirical Study of the Optimal Presentation of Multimedia Summaries of Broadcast News. Mani, I., and Maybury, M. (eds.), *Automated Text Summarization*. MIT Press. pp. 391-401.

Oard, Doug, 2001. The CLEF 2001 Interactive Track . *CLEF-2001 Workshop* in Darmstadt, Germany. http://www.glue.umd.edu/~oard/research.html

Ramabhadran, B., Gao, Y., and Picheny, M., 2000. Dynamic selection of feature spaces for robust speech recognition. *ICSPL*.

Ulargiu, Barbara. 2000 *Accessibility of Oral History Collections: An investigation of current practices and future developments.* Masters Thesis, University of Sheffield, September, 2000

Young, S., 1996. “A review of large-vocabulary continuous-speech recognition”, I*EEE Signal Processing Magazine*, pp. 45-57, Sep.

Zweig, G., et al., 2001. The IBM 2001 Conversational Speech Recognition System, *The 2001 NIST Hub-5 Evaluation Workshop*, May.