FRANCISKA DE JONG
WILLEMIJN HEEREN
ARJAN VAN HESSEN
ROELAND ORDELMAN
ANTON NIJHOLT
University of Twente
Human Media Interaction (HMI)
Enschede, the Netherlands
{fdejong, hessen, ordelman, anijholt}@cs.utwente.nl

Automated Metadata Extraction for Semantic Access to Spoken Word Archives[1]

1. Introduction

Although oral culture has been part of our history for thousands of years, we have only fairly recently been enabled to record and preserve that part of our heritage. Over the past century millions of hours of audiovisual data have been collected.[2] Typically, audiovisual (A/V) archival institutes are the keepers of these collections, a significant part of which contains spoken word materials, such as interviews, speeches and radio broadcasts. More recently more small-scale initiatives have also contributed to the wealth of oral narratives, and even highly individual contributions in the form of spoken blogs add to the amount of content of which the oral culture consists. These materials have great potential for a wide range of applications, e.g., research on social and historical questions, courseware development, reuse for cultural education and new creative productions such as documentaries. From an archival perspective the preservation and maintenance of spoken word collections requires adequate item descriptions. It is generally acknowledged, however, that many A/V collections are poorly annotated, and as a consequence, poorly accessible.

The difficulties regarding disclosure and access that are exemplary for A/V archives in the cultural heritage domain can be illustrated quite clearly by taking a look into the state of the archives of the Dutch regional radio channel 'Radio Rijnmond'. The Radio Rijnmond collection consists of broadcasts recorded and kept since the launching of the radio channel in 1983, amounting to tens of thousands of hours today. Apart from the music programmes it spans several genres, including local news reports, talk shows, interviews, and includes a considerable amount of speech fragments that from an historical perspective form rich and valuable content. Only a small amount of the collection has been disclosed, i.e. manually annotated through standardized archival descriptions of the recordings' content and context (production date, producer etc.). The annotations are commonly referred to as 'metadata'. Each hour of broadcast material has been annotated with at least some information on its context, and description keywords have been assigned to content. In some cases a description of a few sentences was added. The majority of this collection, however, remains in the deposits, partly on analogue data carriers and undisclosed, until resources allow for annotation. To access the disclosed part of the Radio Rijnmond collection, the catalogue can be searched online. Search results are presented as a list of document descriptions, but these are not linked to the physical recordings: there is online access to metadata, not to the speech. To listen to the recordings one has to visit the archive's listening room and subsequently request copies for further exploration and use. This procedure obviously lacks the flexibility to encourage interested individuals to explore this rich collection, and discourages reuse, large-scale research on collection parts, or exploitation in educational settings.

In general, to move towards more exploitable spoken word collections, the quantity and quality of the annotations, widely acknowledged key factors for (re-)usability of audiovisual materials, have to be increased. There are several lines of R&D involved in dealing with both the existing backlog and the inherent potential for use. First there is the issue of ongoing digitization, and the development of standards ensuring long-term preservation of digital audio materials. Then there is the issue of storage and web-based access to digital audiovisual content. R&D relevant for these two issues is on the agenda of archiving professionals and portal developers. On the other side there is (a) research into (semi-)automatic annotation schemes and index generation for fine-grained access to audiovisual data, and (b) research into the design of user interfaces that support users with modes of interaction with spoken audio content that fits the needs of use in a wide range of scenarios. In order to increase the chances that this type of R&D activities will indeed be applicable outside the laboratory and can be taken up in real-life scenarios, attention is also needed for issues of accuracy of automated analysis for less well studied audio types, such as conversational speech and other heterogeneous content (so-called surprise data; [9]), and for affordability of integrating innovative ways of working in existing archival workflows. User interfaces and automation of annotation for spoken audio content will be the topics addressed in this paper.

The role of metadata for content exploration will be presented more elaborately in Section 2. A rough overview of the ways in which natural language processing and search technology can be deployed to support the process of metadata generation is given in Section 3. The design issues involved in developing user interfaces for the interaction with audio content will be outlined in Section 4. Section 5 describes the migration from research in laboratory settings into the delivery of support for real-life use scenarios. Section 6 concludes the paper.

[1] Keynote speech in this Symposium.
[2] A report on European collections gave estimates of over nine million hours of audio and over ten million hours of video [12].

2. Metadata and Content Exploration

In the information science field the annotation of content is commonly referred to as metadata. In principle each information object comes with a record describing what is known about the object and its content, such as when (date) and where (location) the data were produced, and the author or agent that produced the data. For certain specialized cultural heritage collections, for example in cases where the data sets represent sequences of scholarly observations (repeated measurements, transitions detected in series of images, changes in language use over time, etc.), more refined annotations may also have been attached, such as the reliability and validity of certain metadata fields. The format of metadata may vary and include keywords from standard lists or thesauri, free text labels, numerical details and links or references to other information objects.

Influenced by the transformation of content collections into digital libraries, the role of metadata and annotation processing is becoming more and more diverse and facetted. At least three trends can be observed:

1. Access tools link distributed collections at the semantic level captured in the metadata: the tombstone details that are beyond dispute and that describe items as monolithic objects (author, date of excavation, location of interview, etc.), but also annotations that are the result of scholarly analysis and interpretation of the content;
2. Enrichment of the primary content descriptions can more easily be extended with annotations from secondary users; this may involve elements reflecting dynamic annotation processes, such as protocols for error correction, as well as changes due to the dynamics in analysis frameworks, or to the dynamics in the availability of contextual details, including geo-spatial and timing information; this enrichment process can be linked to concepts such as Web 2.0 and crowd sourcing, depending on the profile of the annotators involved;
3. Human-generated metadata can be enriched with automatically extracted content features, e.g. resulting from all kinds of content mining, including text stylistics, genre classification, date detection, etc. Partly due to the probabilistic models on which machine-generated annotations are likely to be based, this type of metadata enrichment may introduce new forms of uncertainty and fuzziness in addition to the annotation layers indicating reliability and validity.

These trends are likely to be key factors in the introduction of data-driven analysis models in the cultural heritage domain. The role of metadata will become even more important with the emergence of the so-called Fourth Paradigm [8][3] and its likely uptake within the e-humanities. The integration of distributed content collections (including the available metadata) and state-of-the-art methods for text mining may lead to novel models for knowledge discovery on the basis of digital libraries content as data. If spoken word collections are made increasingly available online, and if the recordings become part of a networked collection of multimedia datasets that are richly annotated and linked to a myriad of user communities, the oral cultural heritage may benefit in unprecedented ways from these methodological and technological innovations. This perspective has been one of the drivers for the more in-depth investigation that our research group (Human Media Interaction: HMI) has undertaken into the options and challenges for the annotation of oral history collections and the coupling of this specific type of digital library objects to other libraries within the Netherlands, across Europe, and even further away.

3. Spoken Document Retrieval for Cultural Heritage Collections

There is wide agreement that speech-based, automatically generated annotation of audiovisual archives may be an alternative for and/or complementary to semantic access based on manual annotation, (e.g., [1, 3]). As the automatic annotation process generates time-labels, time-stamped indexes can be built that allow searching within documents at various levels (words, speaker-turns, topics). The results of automated transcription can be searched directly, and transcripts can also serve as a basis for many techniques from the field of Natural Language Processing (NLP) that can help to characterize the semantic content of speech in more refined ways.

A typical layout of a spoken document retrieval system is shown in Figure 1. The system's user interface allows the user to formulate search requests, and also shows the user the results (i.e. speech results + metadata). To match the user's needs to the index, the query is processed and subsequently checked against the index using information retrieval technology. Automatic speech recognition (ASR) together with some pre- and post-processing (including NLP) can be used to generate a textual transcript and to turn the text into a time-coded index for an A/V collection.

[3] Cf. also (http://research.microsoft.com/en-us/collaboration/fourthparadigm/)

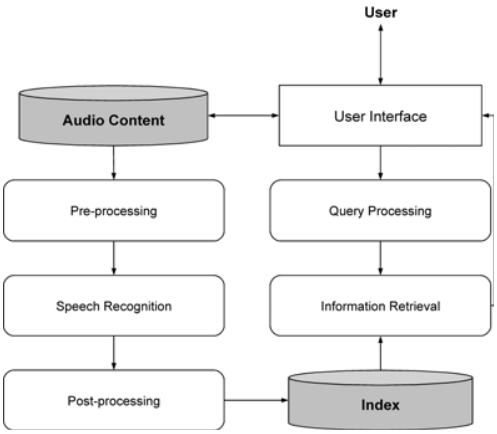

Figure 1: A generic spoken document retrieval system.

A side-effect of using automatic transcription technology is that it introduces errors, which causes errors in the index. Still, spoken document retrieval has been proven successful in the broadcast news domain [4], and has been developed for a larger set of domains, including voice-mail messages [21], webcasts [14], and meeting recordings [22]. Spoken document retrieval was declared a solved problem with the results of [4]. However, this mission accomplished proclamation can be considered to be overgeneralizing the results for a singular domain and to have hampered the thorough exploration of what is needed for other domains. In the rest of this section we will focus on R&D efforts needed to develop spoken document retrieval technology for cultural heritage collections, including historical recordings of speeches and broadcasts (e.g., [5]), and speech recorded in the context of oral history research (e.g., [15]).

3.1 Alignment as ASR Light

Alignment is the process of using an ASR system to recognize and time-code utterances in a fragment, where (a subset of) the words occurring in the fragment, but not their timing, are known beforehand. Alignment of text and audio uses a subset of the features used in a standard ASR engine. It works best when available transcripts closely follow the speech, such as can be the case with accurate minutes from a meeting, with news broadcasts for which autocues were used, or with fully written out speeches. When the available collateral text allows for the successful application of the alignment strategy, alignment has a number of benefits: it saves the development and tuning of collection-specific automatic speech recognition, it is accurate (e.g., with respect to collection-specific words) and it is fast.

An example of an archive for which we have applied full-text alignment is the historical collection of radio speeches that Queen Wilhelmina of the Netherlands addressed to the Dutch people during World War II, the so-called Radio Oranje collection. The alignment from an off-the-shelf speech recognition engine was used [17]; performance was adequate for this task: over 90% of all word boundaries were found within 100 ms of the reference, i.e., within the correct syllable. Details for the alignment procedure and performance figures can be found in [7]; details on the search interface for this collection can be found in Section 4.

Another audio object for which alignment could be applied is the audio recording of the reading of *The Diary of Anne Frank.* Of course here a perfect transcript was available, and the main goal was not primarily in time-coding the audio, but in enhancing the user experience for the recording by automatically coupling each audio fragment to the pieces of text and to (semi-)automatically identify links to related pieces of information, such as an interactive time line, and relevant biographical details. Cf. Section 4.3 for details on the interface that was designed for this purpose.

3.2 Research Topics in Speech Recognition and Retrieval

Though several research projects have been concerned with developing well-performing speech recognition for a number of languages, indexes based on Automatic Speech Recognition (ASR) will always be noisy to a certain extent due to limitations in the recognition performance. This has led researchers to question the suitability of the current standard for automatic speech recognition output, the so-called 1-best ("one-best") transcripts for indexing purposes (e.g., [19]). Searching indexes that contain errors induced by speech recognition are expected to challenge users more than regular text search, and search in audio documents may also be considered relatively difficult.

Generally speaking, automatic speech recognition is used to find the sequence of words that most accurately represents the speech content. The best performing systems do this by using models at three different levels: (i) the acoustic level, where each frame of audio (of around 25 ms) is matched against phoneme, i.e. speech sound, models, (ii) the word level, which limits the allowable sequences of phonemes to meaningful ones, i.e. words, and (iii) above the word level, which introduces a preference for word sequences similar to those seen in a representative text sample of the language.

Speech indexing on the basis of ASR is of sufficient quality for broadcast news content, i.e. mostly read speech. Many collections from cultural heritage, however, are made up of speech that can best be categorized as spontaneous. This type of data is different from read speech in form as well as substance, and poses challenges because it is usually a mismatch to the type of speech a speech recognition system was developed for.

Challenges at the acoustic level. A typical speech signal consists of a combination of clean speech (containing natural variation in e.g., pronunciation), a speech channel (including transducers and acoustics), and additive noise (applause, tape hiss). The assumption underlying speech sound models is that natural variation in speech leads to a somewhat predictable variation in the models. This type of variation can be handled by using Gaussian Mixture Models and Hidden Markov Models. The speech channel, e.g., a telephone line, is usually static for a given fragment of audio. This can be dealt with by training models on material that has similar channel characteristics. Additive noise is a problem that is much more difficult to solve. It is one of the most important reasons why some collections from the cultural heritage domain show much reduced speech recognition performance as compared to broadcast news.

Challenges at and above the word level. At the word level, an automatic speech recognition system only produces words that are in a predefined lexicon. A typical lexicon size would be 100,000 words, whereas the Oxford English Dictionary for instance contains over 600,000 different words. Creating a lexicon that covers most of the speech content in a collection is quite easy, making one that covers all of it is practically impossible. In fact, many of the words that may not be in the lexicon are potentially the most interesting words in the collection: named entities (e.g., Lexington Street, Mr. Johnson) or rare terms (e.g., xylophone). After all, it is reasonable to assume that whenever a speaker decides to use such a word there is a compelling reason: ignoring the occurrence of low-frequency words is therefore likely to be detrimental to the representation of the speech content. Defining a lexicon that is optimized for topics in the cultural heritage domain is not trivial, however, since it requires digitally available texts on comparable topics.

Due to acoustic similarity, multiple word-level transcriptions with similar acoustic likelihoods can be generated for most speech segments. Determining which the most likely one is requires contextual knowledge. For speech from the broadcast news domain newspapers and other written sources with similar content can be used to generate that knowledge, but it is unlikely that this holds for spontaneous speech. Take, for instance, an interview collection on personal experiences from detention in a World War II concentration camp. Despite the general topic being known, it is not easy to predict the wording that will be used. Euphemisms, archaic expressions and foreign words can easily pop up in this type of speech, and as the speaker is gathering his thoughts while speaking, there will be disfluencies and ungrammatical sentences. Named entities may pose even more of a challenge, since their very existence can be introduced in these collections, meaning that there is no previous record of them.

Finding a representative collection of text first requires knowledge of the properties of the speech, and then requires large corpora of digitally available text that match those. Since spontaneous speech does not follow the same constraints as most of written language and usually there are not much truly spontaneous speech transcriptions available, it is difficult to model it correctly using standard statistical techniques. Some of these challenges are alleviated through the fact that spoken document retrieval does not require the same deterministic approach to speech recognition as the approach proven effective for e.g. the dictation task. It is not critical to determine the most likely sentence, as all transcription alternatives may be included in the index. To ensure optimal retrieval performance a confidence score could be added per alternative. The calculation of such a confidence score, however, may be hampered by a poor acoustic match due to additive noise. Moreover, when a word is not in the lexicon, it will not appear in the index at all.

The quality of speech recognition output is traditionally measured using the Word Error Rate, i.e. the percentage of erroneous words in the transcript. Broadcast news speech may be transcribed with as little as 10% word error rate [13], whereas spontaneous speech typically results in an error rate of over 40% (e.g., [5]). In case of additive noise or speech channel mismatches we found that this may even rise to over 60% [16]. When applying speech recognition output in spoken document retrieval systems, the consensus seems to be that an error rate of less than 30-40% renders a system usable [4]. For most collections in the cultural heritage domain it is currently unfeasible to automatically generate high quality speech transcriptions (i.e. with error rates under 20%). This means that the standard approach –modeling spoken document retrieval as information retrieval applied to an automatically generated transcription– may not be the preferred route. For the kinds of spoken materials of interest here, the quality of automatic transcripts is likely to remain a bottleneck. Hence efforts have been directed towards improving quality of systems applied to spontaneous speech and suboptimal recordings. At the same time, the retrieval engine should exploit the automatically generated textual representations of speech in an optimal manner.

3.3 Research Topics in Automated Semantic Annotation

The automation of conceptual or semantic annotation has the potential to reduce the pile of undisclosed content in archives and to enhance the support for facetted search. As explained above, the use of (ASR) for the exploitation of the linguistic content can be helpful in bridging the semantic gap between low-level media features and conceptual information needs through: (a) the generation of a textual representation that can be the basis for semantic analysis, such as extraction of named entities, automatic classification, topic clustering or even summarization, and (b) the classification of audio content in terms of extralinguistic conceptual features such as emotion and affect. For a wide range of topics performance levels of usability aspects could be improved.

Retrieval of out-of-domain words. The potential for semantic analysis is hampered by the fact that a large number of user queries to multimedia collections involve named entities, such as personal names and locations. Especially named entities run the risk of being out-of-vocabulary. If a term does not occur in the ASR vocabulary it can't figure in the ASR transcripts and will be irretrievable. One way of reducing this problem is by carefully annotating at least the names of places and persons that are associated with a certain multimedia document during the description process. Archivists can help to improve access to multimedia collections by identifying

related sources with relevant terminology, so that these can be fed into the automatic processing workflow. In order to benefit from the use of collateral data for cost-effective annotation and access, research into optimizing the workflow of multimedia archiving is necessary.

Support for multimodal search. For spoken word content, semantic search is of course feasible insofar as the manual annotation layers contain words that are suited as search terms, such as proper names. For the nonlinguistic image content, concept detection tools based on automated labeling (cf. [20]) are a key contributor to conceptual search. Additional conceptual annotation layers can also be generated based on the automatic labeling of segments with speaker identities and affect labeling. Coupling of these types of labeling to the labeling based on the detection of concepts in the content layers for other modalities (e.g., image, collateral text) has only recently become a topic of research. As shown by initial experiments conducted in the context of IST project MESH[4] the number of the parameters involved (length of segment, density of speech track, voice-over versus dialogue, user preference for results in image or speech content) requires a very careful design of the experiments in order to draw any conclusion on how to set up the fusion of retrieved multimodal search results.

Link generation across collections. Scaling up state-of-the-art analysis tools for single collections into techniques that can be applied across collections to identify semantic links, trends and motifs seems a matter of time. Information in scripts, audio tracks, wikis or blogs will be used for the cross-modal detection of people, places, events, etc., and for link generation between all kinds of entities occurring in audiovisual content. In addition to automatic content labeling and link detection, there also seems to be room to exploit the navigation behavior of users as a source that could be exploited for the labeling of content. The concept of a myriad of information trails crossing the web generated by users that either implicitly or explicitly leave their traces is the starting point of more recent initiatives for European collaboration among the keepers of open data collections assuming that users can be engaged in the annotation process: in combination with machine learning techniques and the support of selection and feedback tools, they will enable the gradual improvement of tagging performance. In particular the design of a sound evaluation framework, including ground truth annotations will be a crucial prerequisite for success.

4. Research Topics in User Interface Development

When searching for fragments of speech, what do end users need? What kinds of questions do they pose? How can we help them to find the fragments of interest efficiently? Which entities should be extracted with highest performance in order to serve the needs of each of the various user types best? Relatively few studies into user requirements for searching audiovisual archives in the cultural heritage domain have been carried out. At the start of the R&D work to access technology for spoken word archives at HMI no working examples of online access systems to spoken word documents using automatically generated content representations were available. To get started with the development of a user interface for spoken word audio content, we therefore carried out an initial requirements analysis stage and launched an online system for word-level search in a spoken word collection: the ‘Radio Oranje’ speeches.

4.1 Radio Oranje Online

As mentioned above, the Radio Oranje search engine gives access to the speeches that Queen Wilhelmina (1880-1962) addressed to the Dutch people in occupied areas during World War II. The recordings and manual transcripts were preserved by the Netherlands Institute for War Documentation[5] (NIOD) and the Netherlands Institute for Sound and Vision[6] (S&V). Earlier, the collection could only be searched by reading the transcripts at the NIOD and then visiting S&V to obtain copies of the audio. The demonstrator system[7] provides an example of how state-of-the-art technology for indexing and multimodal presentation can boost the accessibility and enliven the perception of such collections (see Figure 2).

[4] http://www.mesh-project.eu/
[5] http://www.niod.nl
[6] http://portal.beeldengeluid.nl/
[7] http://hmi.ewi.utwente.nl/choral/demo

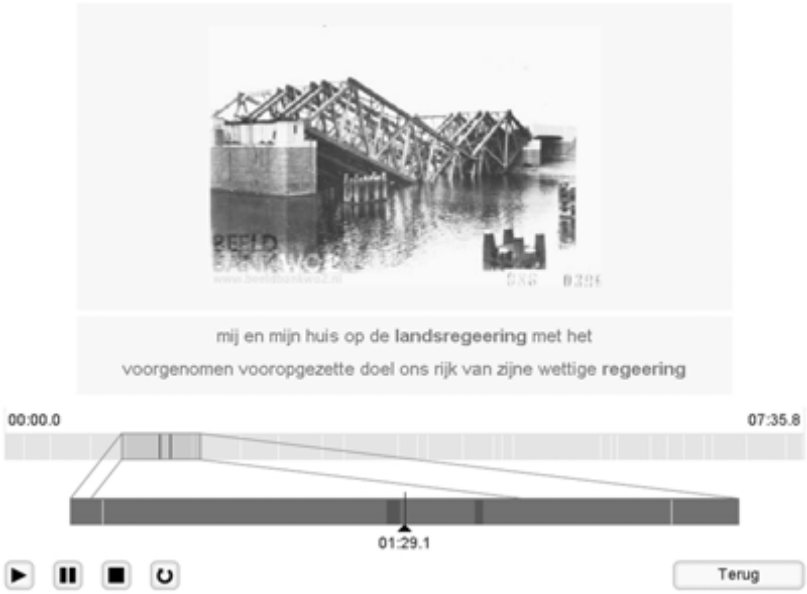

Figure 2: Audio playback page for the Radio Oranje system

The text versions of the speeches were synchronized with the audio using an alignment tool (cf. Section 3.1). On the basis of the alignment a time-stamped index was created that – apart from supporting fine-grained access to the speeches – allows the integration with additional functionalities, such as interactive visualization of the audio content and subtitling. Finally, time-synchronized links to images from a topically related photo database are automatically provided. For this purpose topic labels were assigned to the audio documents using a coarse semantic classification tool.

Interactive user interfaces help the user in building a mental model of the speech content, (e.g., [23]), and facilitate navigation in the audio, (e.g., [18]). The timeline visualization designed for this collection contains different types of information, including segment boundaries and query term locations. This functionality allows the user full control over audio playback and it always presents fragments within the context of the audio document they were taken from. A first evaluation with 10 academic students with a background in humanities showed that they immediately used the timeline visualization to locate relevant intervals, and also that subtitling aided intelligibility. An evaluation with 23 students of Library and Information Sciences showed that users valued the location markers indicating query terms in the timeline as well as the presence of context during playback. Through feedback from these students, but also from discussion with archiving professionals we learned that especially in the cultural heritage domain the presentation of context is important; documents and artifacts should be interpreted in their proper setting.

4.2 The Buchenwald Portal

A successor of the Radio Oranje project is the interface development for the 'Buchenwald' website[8], a Dutch multimedia information portal on World War II concentration camp Buchenwald [16]. The collection, maintained by NIOD, holds both textual information sources and a video collection of testimonies from 38 Dutch camp survivors. For each interview with duration between a half and two and a half hours, an elaborate description, a speaker profile and a short summary are available. The retrieval engine matches queries against the index based on automatically generated transcripts and against the various types of texts. Search results are listed and contain context information (interview duration, location, and date) and content information (speaker profile, short summary) (see Figure 3), and a link to the video file and the elaborate description (see Figure 4). For video navigation a time line visualization was based on the functionality developed for the Radio Oranje portal, slightly adapted in order to be able to differentiate between markings relevant at the document level and those relevant within the document, at the fragment level.

A user evaluation of this system through analysis of its user logs (1096 sessions), showed a 1:2 ratio for users typing a query versus requesting a list of all available documents. This pattern has also been found for the Radio Oranje logs. We estimate that many of the visitors to these web sites so far have mainly been 'looking around', i.e. browsing instead of searching. User sessions were generally short, and search queries often consisted of one or two fairly general, but topic-related terms. Moreover, when a user typed a query it contained a named entity in almost 60% of the cases. We also found that the functionality to access the interviews and the related texts were being used fully and – in comparison with traditional audiovisual archives – frequently. Finally, we have found that the demonstrator systems serve a clear purpose in the discussion with content providers, archivists and historians as to the use, possibilities and restrictions of such technology for disclosure of audiovisual digital library collections.

[8] http://vuurvink.ewi.utwente.nl:8080/Buchenwald

Figure 3: Result listing for the Buchenwald collection.

4.3 Interfaces for the Navigation of Multimodal and Interlinked Content

In many use scenarios people are not just interested in finding (a list of) pieces of information in a certain modality on a certain topic, but in exploring available information from several perspectives, in particular in the case of historical narratives. Here, visualization of chronological details, collateral texts, references to related publications, etc., can all have added value for the way in which people interact with content or for the way in which the information can be used for scholarly analysis. Requirements for navigation in multifaceted, multimodal or otherwise heterogeneous content are not yet well explored, and it is an open question whether generic guidelines can be given for the generation of a well balanced choice from multiple perspectives. HMI approaches these questions by focusing on content for which real users who can engage in evaluations are likely to be found. An example is the time-aligned multimodal presentation via multiple screens of *The Diary of Anne Frank*.

Figure 4: Video playback for the Buchenwald collection

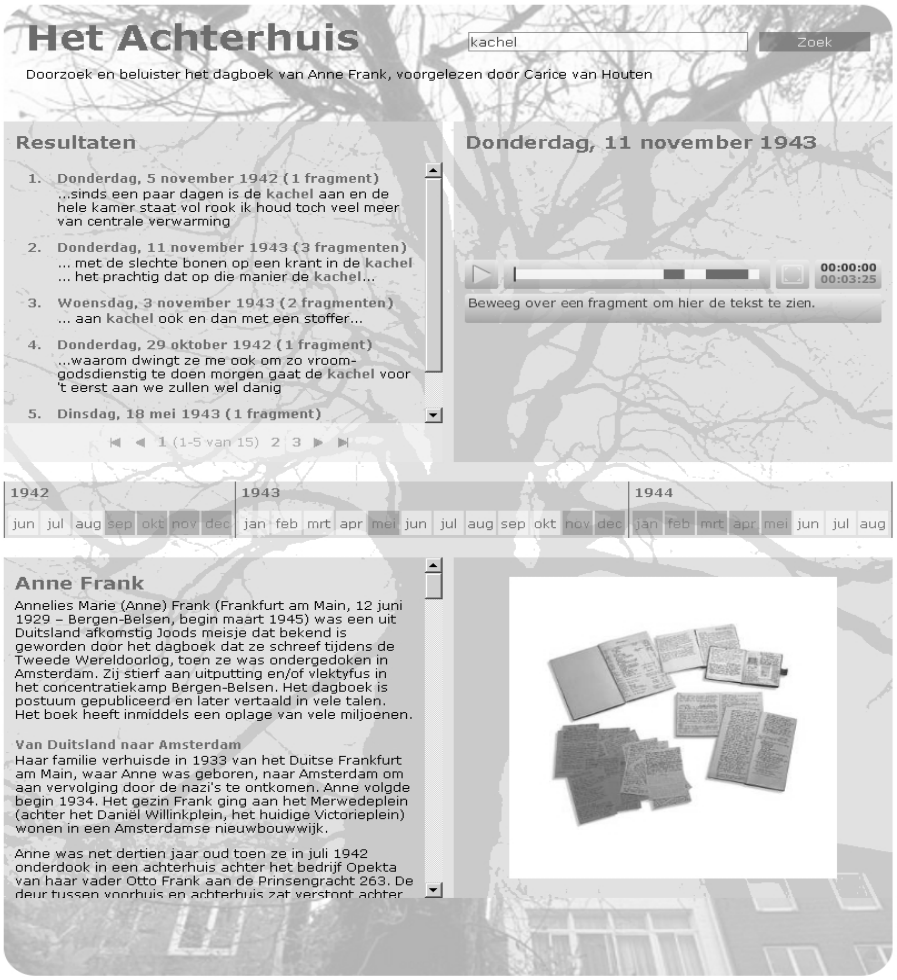

Figure 5: Time-aligned split screens for multimodal presentation

4.4 Next Steps in Improving User Interfaces

The next step in the development of suitable user interfaces is scaling up from relatively small, homogeneous collections to large and heterogeneous collections such as Radio Rijnmond's archives. Query logs showed that often users do not formulate a query, but ask for a list of documents to begin their exploration of a collection, (e.g., [16]). Whereas content listing is a feasible approach for a collection of up to several dozen documents, it is not useful in the setting of digital library access applications that are currently emerging and that will become the standard in the cultural heritage domain in the near future. Portals such as the ones developed in the context of European collaboration between heritage institutions such as MultiMatch[9] and Europeana[10] aim at the integration of distributed multilingual and multimedia heritage collections. Rather than single collections of documents, document clustering (by e.g., year of production, creator, topic, language) could provide users with a way of exploring an archive's contents. This paradigm is already being used by archiving institutes, such as S&V, albeit only for searching document descriptions.

Whereas alignment makes use of (near-)perfect transcripts, automatic speech recognition transcripts contain errors. Consequently, result lists will contain false alarms and misses, i.e. they may contain irrelevant audio fragments and they may be incomplete. This can result in users being unable to find fragments that are present in the collection, [2]. Named entities are particularly at risk of becoming irretrievable, but they are at the same time very popular query terms. The standard way of reducing this problem is through query expansion (e.g., [10]). Assuming that the top results of running the query on this (or some external) collection are correct, those top documents are used to automatically expand the original query so as to include other terms that may be relevant to the information need. This technique has been especially successful in spoken document retrieval, because more query terms make a search more robust towards transcription errors.

To guide users in selecting fragments or documents of interest before they start playing audio, result listings should provide users with insight into the documents' contents through either textual or visual information. If a high-quality textual transcript is available, the existing paradigms of text search can be applied and snippets with

[9] http://www.multimatch.eu
[10] http://www.europeana.eu

sentences matching the user's query may be shown (as in the Radio Oranje demonstrator system). If a transcript's word error rate is over 30%, users have been found to discard textual content representations, (e.g., [21]). As an alternative to presenting low-quality transcripts, the use of keyword extraction approaches, (e.g., [6]), and content visualizations, (e.g., [11, 24]) has been proposed. HMI has explored suitable ways of presenting only selected keywords from the transcripts. Experience up till now has learned that the relation between the accuracy of automatically generated indexes and the user experience should be monitored very closely for future interface design in this domain.

5. From Laboratory Settings to the Support of Real-life Use Cases

On the basis of the options explored thus far the question whether automated transcription and/or annotation technology can form an alternative and/or complementary approach to current practices of disclosing audiovisual archives can be answered affirmatively. However, the number of success stories for such applications in real-world settings is still scarce. The conditions that have hampered a wider take-up and what can be done to increase the chances for its effective deployment will be the topic of this section.

We have identified two main bottlenecks for the take-up of spoken document retrieval technology in audiovisual archives: (i) automatic speech recognition performance and (ii) archival infrastructure and IT expertise. To start with the former: for many collection types found in audiovisual archives the performance is significantly lower than the results reported on benchmark collections: there is clearly a gap between laboratory and real-life conditions. Processing time, i.e. the amount of time it takes to automatically generate indexes, is not a problem, however. We have proposed some research tracks that could lead to improved system performance in Sections 3 and 4. But until acceptable system performance levels are reached and tests with different kinds of users establish its usability, archiving professionals will understandably remain hesitant to take up technology that their customers might not appreciate.

Transfer of knowledge on all kinds of organizational topics is also an issue. Archival infrastructure is a bottleneck, despite the fact that mass digitization is underway and standard metadata schemes as well as trusted digital repositories are being developed. In general, archiving of digital audio requires a deeper understanding of technology than is often available at archives, but to be able to deploy the more advanced technologies and to understand the inherent workflow is more than can be expected of the average collection keeper. There also seems to be a considerable knowledge gap in the smaller institutes. In the Netherlands this state of play has recently triggered a consortium of archivists and developers to set up a joint dedicated expertise center as part of the distributed digital libraries infrastructure. The consortium includes a diverse group of institutes including the Netherlands Institute for War Documentation[11], the Veterans Institute[12], the International Information Center and Archives for the Women's Movement[13] and the Rotterdam Municipal Archives. The initiative is called Verteld Verleden[14] and the plan is to collect, and disseminate best practises in the handling of A/V recordings of spoken word tracks and to launch an online network and service center where spoken word collections can be submitted for automatic annotation and indexing. Offering guidance on how to optimize chances for good automatic indexing performance during the various stages of the handling of audio data is seen as an important ingredient for a programme that in the end should lead to a rich digital library infrastructure for oral history collections.

6. Conclusions

The automated generation of annotation for A/V material is growing into a mature functionality for digital libraries. The ongoing cooperation between technology developers, archivists and researchers with an interest in spoken word content will in the near future result in online access to a large and heterogeneous collection of audiovisual materials. The continued search for possibilities to create a networked infrastructure with dedicated tools and services will enhance the access and reuse of spoken word materials and their integration into existing and future digital library initiatives. In the domain of cultural heritage the added value of new technologies is not so much in the cost-reduction as in the wider usability of the materials, and in the impulse this may bring for sharing collections that otherwise would too easily be considered as of no general importance. If this message is well disseminated the future for oral culture could be bright.

Acknowledgements This paper is partly based on research carried out in the project CHoral - Access to Oral History, which is funded by the NWO program CATCH[15], and on ideas developed within Croatian Memories[16], a project being carried out with funding from the Netherlands Ministry of Foreign Affairs. We want to thank the Netherlands Institute for War Documentation and the Netherlands Institute for Sound and Vision for their cooperation in the development of some of the demonstrator systems described here.

References

[1] W. Byrne, D. Doermann, M. Franz, S. Gustman, J. Hajic, D. Oard, M. Picheny, J. Psutka, B. Ramabhadran, D. Soergel, T. Ward, and W-J. Zhu. Automatic recognition of spontaneous speech for access to multilingual oral history archives. *IEEE Trans. Speech Audio Proc.*, 12(4):420–435, 2004.

[11] http://www.niod.nl
[12] http://www.veteraneninstituut.nl/
[13] http://www.iiav.nl/
[14] http://www.verteldverleden.org
[15] http://www.nwo.nl/catch
[16] http://www.eur.nl/erasmusstudio/projects/matra/

[2] J. Carmichael, P. Clough, E. Newman, and G. Jones. Multimedia retrieval in multimatch: The impact of speech transcription errors in search behaviour. In *Proceedings of the ECDL 2008 Workshop on Information Access to Cultural Heritage*, 2008. Aarhus, Denmark.

[3] F.M.G. de Jong, D.W. Oard, W.F.L. Heeren, and R.J.F. Ordelman. Access to recorded interviews: A research agenda. *ACM Journal on Computing and Cultural Heritage*, 1(1):3–29, 2008.

[4] J.S. Garofolo, C.G.P. Auzanne, and E.M. Voorhees. The TREC spoken document retrieval task: A success story. In *Proceedings of RIAO*, 2000. Paris, France.

[5] J.H.L. Hansen, R. Huang, B. Zhou, M. Deadle, J.R. Deller, A. R. Gurijala, M. Kurimo, and P. Angkititrakul. Speechfind: Advances in spoken document retrieval for a national gallery of the spoken word. *IEEE Transactions on Speech and Audio Processing*, 13(5):712–730, 2005.

[6] A. Haubold and J.R. Kender. Analysis and visualization of index words from audio transcripts of instructional videos. In *Proceedings of the IEEE Sixth International Symposium on Multimedia Software Engineering*, pages 570–573, 2004. Miami, Florida, USA.

[7] W.F.L. Heeren, L.B. van der Werff, R.J.F. Ordelman, A.J. van Hessen, and F.M.G. de Jong. Radio Oranje: Searching the queen's speech(es). In C.L.A. Clarke, N. Fuhr, N. Kando, W. Kraaij, and A. de Vries, editors, *Proceedings of the 30th ACM SIGIR*, pages 903–903, New York, 2007. ACM.

[8] T. Hey, S. Tansley, and K. Tolle eds. *The fourth paradigm. Data-intensive scientific discovery*. Microsoft Research. Redmond, Washington, 2009.

[9] M. Huijbregts. *Segmentation, Diarization and Speech Transcription: Surprise Data Unraveled*. PhD thesis, University of Twente, November 2008.

[10] P. Jourlin, S.E. Johnson, K. Spärck Jones, and P.C. Woodland. General query expansion techniques for spoken document retrieval. In *Proceedings of the ESCA Workshop on Accessing Information in Spoken Audio*, pages 8–13, 1999. Cambridge, UK.

[11] D.G. Kimber, L.D. Wilcox, F.R. Chen, and T.P. Moran. Speaker segmentation for browsing recorded audio. In *Proceedings of CHI 1995*, pages 212–213, 1995. Denver, Colorado, USA.

[12] E. Klijn and Y. de Lusenet. *Tracking the reel world. A survey of audiovisual collections in Europe*. European Commission on Preservation and Access, Amsterdam, 2008.

[13] S. Matsoukas, J-L. Gauvain, G. Adda, T. Colthurst, C-L. Kao, O. Kimball, L. Lamel, F. Lefevre, J. Ma, J. Makhoul, L. Nguyen, R. Prasad, R. Schwartz, H. Schwenk, and B. Xiang. Advances in Transcription of Broadcast News and Conversational Telephone Speech within the Combined EARS BBN/LIMSI System. *IEEE Transactions on Audio, Speech and Language Processing*, 14(5):1541–1556, 2006.

[14] C. Munteanu, R. Baecker, G. Penn, E. Toms, and D. James. The effect of speech recognition accuracy rates on the usefulness and usability of webcast archives. In *Proceedings of CHI 2006*, pages 493–502, 2006. Montreal, Canada.

[15] D.W. Oard, D. Demner-Fushman, J. Hajic, B. Ramabhadran, S. Gustman, W.J. Byrne, D. Soergel, B.J. Dorr, P. Resnik, and M. Picheny. Cross-language access to recorded speech in the malach project. In *Proceedings of the 5th International Conference on Text, Speech and Dialogue*, pages 57–64, London, UK, 2002. Springer-Verlag.

[16] R.J.F. Ordelman, W.F.L. Heeren, M.A.H. Huijbregts, D. Hiemstra, and F.M.G. de Jong. Towards affordable disclosure of spoken word archives. In *Proceedings of the ECDL 2008 Workshop on Information Access to Cultural Heritage*, 2008. Aarhus, Denmark.

[17] B. Pellom. Sonic: The University of Colorado continuous speech recognizer. Technical report, University of Colorado, March 2001. Technical Report TR-CSLR-2001-01, University of Colorado.

[18] A. Ranjan, R. Balakishnan, and M. Chignell. Searching in audio: the utility of transcripts, dichotic presentation and time-compression. In *Proceedings of CHI 2006*, pages 721–730. Quebec, Canada.

[19] M. Siegler. *Integration of continuous speech recognition and information retrieval for mutually optimal performance*. PhD Thesis. Carnegie Mellon University, 1999.

[20] C. G. M. Snoek and M. Worring. Concept-based video retrieval. *Foundations and Trends in Information Retrieval*, 14(2):215–322, 2009.

[21] L. Stark, S. Whittaker, and J. Hirschberg. ASR satisficing: the effects of ASR accuracy on speech retrieval. In *Proceedings of the International Conference on Spoken Language Processing*, pages 1069–1072, 2000. Beijing, China.

[22] P. Wellner, M. Flynn, and M. Guillemot. Browsing recorded meetings with Ferret. In *Proceedings of Machine Learning for Multimodal Interaction*, pages 12–21, 2004. Martigny, Switzerland.

[23] S. Whittaker, J. Choi, J. Hirschberg, and C. Nakatani. What you see is almost what you hear: Design principles for accessing speech archives. In *Proceedings of the Fifth International Conference on Spoken Language Processing*, 1998. Sydney, Australia.

[24] S. Whittaker, J. Hirschberg, J. Choi, D. Hindle, F.C.N. Pereira, and A. Singhal. SCAN: Designing and evaluating user interfaces to support retrieval from speech archives. In *Proceedings of SIGIR99 Conference on Research and Development in Information Retrieval*, pages 26–33, 1999. Berkeley, USA.