INTERDISCIPLINARY SCIENCE REVIEWS, Vol. 34 No. 2–3, 2009, 236–252

Easy Listening: Spoken Document Retrieval in CHoral

Willemijn Heeren, Laurens van der Werff, Franciska de Jong, Roeland Ordelman, Thijs Verschoor and Arjan van Hessen

Human Media Interaction Group, University of Twente, Enschede, The Netherlands

Mies Langelaar

Rotterdam Municipal Archives, Rotterdam, The Netherlands

Given the enormous backlog at audiovisual archives and the generally global level of item description, collection disclosure and item access are both at risk. At the same time, archival practice is seeking to evolve from the analogue to the digital world. CHoral investigates the role automatic annotation and search technology can play in improving disclosure and access of digitized spoken word collections during and after this transfer. The core business of the CHoral project is to design and build technology for spoken document retrieval for heritage collections. In this paper, we will argue that in addition to solving technological issues, closer attention is needed for the work-flow and daily practice at audiovisual archives on the one hand, and the state-of-the-art in technology on the other. Analysis of the interplay is needed to ensure that new developments are mutually beneficial and that continuing cooperation can indeed bring envisioned advancements.

KEYWORDS Spoken document retrieval, Speech indexing, Spoken word archives, Speech recognition, User interfaces

Introduction

Although oral culture has been part of our history for thousands of years, we have only fairly recently become able to record and save that part of our heritage. Over the past century, we have collected millions of hours of audiovisual data. A recent report on European collections, for instance, gave estimates of over nine million hours of audio and over 10 million hours of video (Klijn and de Lusenet 2008). Audiovisual (A/V) archives maintain these

DOI 10.1179/174327909X441135

collections, a significant part of which contains spoken word materials, such as interviews, speeches and radio broadcasts. These materials have great potential, for example, for new creative productions such as documentaries, research on social and historical questions, and educational purposes. It is generally acknowledged, however, that many A/V collections are poorly disclosed, as well as poorly accessible. The goal of the NWO-CATCH project CHoral (2006–2010)[1] is to investigate and develop technology with the goal of improving accessibility and disclosure for digital spoken word collections. CHoral is a cooperation between archiving professionals at the Rotterdam Municipal Archives and multimedia retrieval researchers from the University of Twente.

To illustrate the difficulties regarding disclosure and access that are exemplary for A/V archives in the cultural heritage domain, we present the archive of the 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. It spans several genres including local news reports, talk shows, interviews, and music programmes. Only a small amount of the collection has been disclosed, i.e. manually annotated through standardized archival description of the recordings' content and context (production date, producer, etc.). We will also refer to this annotation as 'metadata'. Each hour of broadcast material has been annotated with at least some information on its context, and for content description, keywords have been assigned. In some cases, a description of a few sentences was added. The majority of this collection, however, remains in the deposits, mainly 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 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.

To move towards more exploitable collections in general, first data storage needs to be adapted to the requirements of automatic analysis and online access. Secondly, the quantity and quality of the annotations, widely acknowledged key factors for (re-)usability of materials, have to be improved. How to proceed, however, when there is a large backlog of undisclosed materials that needs to be annotated, and when the information density in those annotations that are available is fairly low? Moreover, the cost-benefit of possible solutions must obviously be taken into account.

To begin solving the problems of the Radio Rijnmond collection in particular and at the same time address issues that are relevant for A/V collections and archives around the world, the CHoral project combines two lines of work. On the one hand, there is ongoing digitization, and the

development of a standard metadata scheme and a trusted digital repository ensuring long-term preservation of the digital audio. This part of the work is carried out by the archiving professionals. On the other, there is research into automatic annotation schemes and index generation for fine-grained access to audiovisual data to deal with the existing backlog. This part of the project is carried out by researchers from the University of Twente.

In the remainder of this paper, we will explain the issues that set the agenda for CHoral's research and development, which can help to overcome the obstacles and problems described above. The field in which CHoral operates is commonly known as spoken document retrieval. In the next section, earlier research into spoken document retrieval for cultural heritage collections will be presented, and the overall approach taken in CHoral will be introduced. The work in CHoral will be presented more elaborately and the issue of how spoken document retrieval can be fitted into the current practice of A/V archives will be discussed.

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., Byrne *et al.* 2004; de Jong *et al.* 2008). 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).

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 together with some pre- and post-processing are used to arrive at an index for an A/V collection.

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 (Garofolo *et al.* 2000), and has been developed for a larger set of domains, including voice-mail messages (Stark *et al.* 2000), webcasts (Munteanu *et al.* 2006), and meeting recordings (Wellner *et al.* 2004). In the rest of this section, we will focus on earlier efforts to develop spoken document retrieval technology for cultural heritage collections, including historical recordings of speeches and broadcasts (e.g., Hansen *et al.* 2005), and oral histories (e.g., Oard *et al.* 2002).

One of the early initiatives was the MALACH project (Multilingual Access to Large spoken ArCHives), a US NSF project (2001–2007) that aimed at the investigation of access technology for a vast collection of testimonies from survivors, witnesses and rescuers of the Holocaust (Oard *et al.* 2002). The goal of that project was to advance speech recognition for the oral history domain and to study how recognition could be best incorporated in further

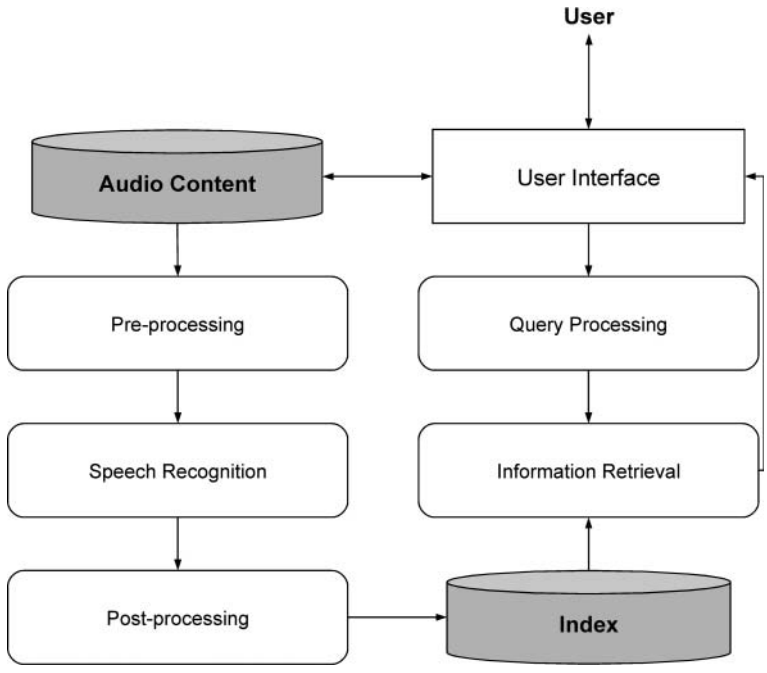

FIGURE 1 A generic spoken document retrieval system.

processing and retrieval steps (Gustman *et al.* 2002). Another project that contributed to advancing spoken document retrieval in the cultural heritage domain was The National Gallery of the Spoken Word project (Hansen *et al.* 2005). In that project, the SpeechFind spoken document retrieval system was developed: it automatically generated annotations for audio documents and made the audio searchable though a Web-interface. One of the lessons learned in that project was that it is not the age of the recording, but rather its quality and audio characteristics that determine the quality of the index. In the Netherlands, the WFH project[2] used automatic transcription to disclose a collection of interviews and lectures by a famous Dutch writer (Huijbregts *et al.* 2005).

The IST-FP5 project ECHO (European CHronicles Online) aimed at the realization of a searchable multilingual collection of video documentaries deploying speech recognition as one of the core technologies. The IST-FP6 project MultiMatch (2006–2008) aimed to develop an integrated search engine for multilingual and multi-media content from cultural heritage (Amato *et al.* 2007). For streaming media types they employed a commercial speech recognition system, the performance of which was judged sufficient for experimenting on search within videos (Carmichael *et al.* 2008b).

Though several research projects have been concerned with developing well-performing speech recognition for a number of languages, noisy indexes due to imperfect transcripts will remain a problem. This has led researchers

to question the suitability of the currently standard type of automatic speech recognition output, so-called 1-best ('one-best') transcripts, for indexing purposes (e.g., Siegler 1999). We share this point of view, and therefore CHoral is in the first instance targeting the improvement of index quality, and in the second trying to better understand the relation between the quality of automatically generated transcripts and index quality. 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. This calls for well-designed search interfaces, and therefore a second topic in CHoral is the development of tools that help users in the retrieval of and interaction with spoken word documents. These topics are discussed further in the following two sections.

Research topics in speech recognition and retrieval

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 automatically generated transcripts 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 usually is a mismatch to the type of speech a speech recognition system was developed for. In the following two subsections, the challenges will be discussed in more detail.

Challenges at the acoustic level

A typical speech signal consists of a combination of clean speech (containing natural variation in, for example, 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. It 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 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, he or she has a compelling reason for it: ignoring the occurrence of this word 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 is the most likely one 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 traditional dictation task did. It is no longer 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 a 10 per cent word

error rate (Matsoukas *et al.* 2006), whereas spontaneous speech typically results in an error rate of over 40 per cent (e.g., Hansen *et al.* 2005). In the case of additive noise or speech channel mismatches, we found that this may even rise to over 60 per cent (Ordelman *et al.* 2008). When applying speech recognition output in spoken document retrieval systems, the consensus seems to be that an error rate of less than 30–40 per cent renders a system usable (Garofolo *et al.* 2000).

Advancing speech recognition in a non-dictation context

Two topics have been taken up by the CHoral project to further spoken document retrieval for cultural heritage collections: (i) finding a suitable evaluation measure for establishing the performance of spoken document retrieval systems, and (ii) the use of a different output format, lattice structures, to compensate for certain types of speech recognition errors.

Evaluation issues

Most modern speech recognition engines were developed with dictation as their primary goal. Evaluating their performance is done by calculating the number of errors that the system makes, expressed as the word error rate. Spoken document retrieval is in fact a variation of Information Retrieval, where performance is determined by the amount of *information* that can be found, expressed as (Mean) Average Precision. Although previous work has shown that the Word Error Rate and Mean Average Precision measures may be closely correlated in certain conditions (Garofolo *et al*. 2000), this does not imply that optimization of word error rate is equivalent to optimization of mean average precision.

Word error rate can only be calculated for a single, literal transcription of the speech: the 1-best transcript. When an index is built not from this 1-best transcript, but from a non-deterministic set of hypotheses, many variables can come into play that the word error rate cannot capture. Examples of such variables are confidence scores and transcription alternatives (see following subsection). These variables may have a profound impact on mean average precision.

Optimization of speech recognition for spoken document retrieval is therefore a matter of optimizing the performance of the retrieval component through improvement of the index. Although it is possible to measure retrieval performance with mean average precision (MAP), this cannot be easily done for any given collection: to establish a MAP-score it is necessary to go through the rather labour-intensive process of selecting queries and judging documents. This prohibits evaluation on collections that are dissimilar to the typical benchmark collections.

However, given a reasonably large set of reference transcriptions for a new collection, it is possible to directly evaluate the quality of the index based on speech recognition output (van der Werff and Heeren 2007). In the vector space model of information retrieval, indexing can be seen as the creation of a multi-dimensional vector representing a document. For retrieval, the angle

between a vector representation of a user's query and each document vector is calculated. Our proposal is that in order to evaluate the similarity between speech recognition output and the content of the spoken word document, the angle between the vector(s) representing a reference transcription and the vector(s) representing the speech recognition output can be used.

Lattice-based indexing

For each term, an index contains a list of documents that are deemed relevant for that term, along with scores that represent the relevance of each document for that term. Such an index is traditionally built on plain text or, in the case of spoken document retrieval, on a 1-best, automatically generated transcription. When a non-deterministic input, such as a lattice structure (see Figure 2), is used for making an index, confidence scores must be introduced into the index. Calculating these scores is a trivial task (Wessel *et al.* 2000), but integrating them into an index is less straightforward.

The confidence score of a particular word depends upon the words immediately preceding it. An index that is based on single words and their individual confidence scores will therefore not be able to fully exploit the properties of its underlying lattice structure. One solution is to index all lattice n-grams, where an n-gram is a series of n subsequent words in the lattice structure and n depends on the size of the language model used for scoring. Such an index has to contain all 1-grams, 2-grams, 3-grams, etc. that are found in the lattice, each with their own confidence score. For large databases, Position Specific Posterior Lattices (Chelba and Acero 2005) may be used to reduce the index size, but for CHoral this may not be beneficial, since the Radio Rijnmond collection is a closed set of modest size.

Conclusion

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 per cent). This means that the standard approach — modelling 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 should be 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.

The use of lattices to create an index is a promising way of solving some of the current issues. However, this requires specific optimizations on both the retrieval and the speech recognition part. Evaluation of a speech recognition system must therefore be based on its suitability for making an index (van der Werff and Heeren 2007), and at some point, when the quality of automatically generated transcripts is no longer a bottleneck, on the suitability of this index to answer queries.

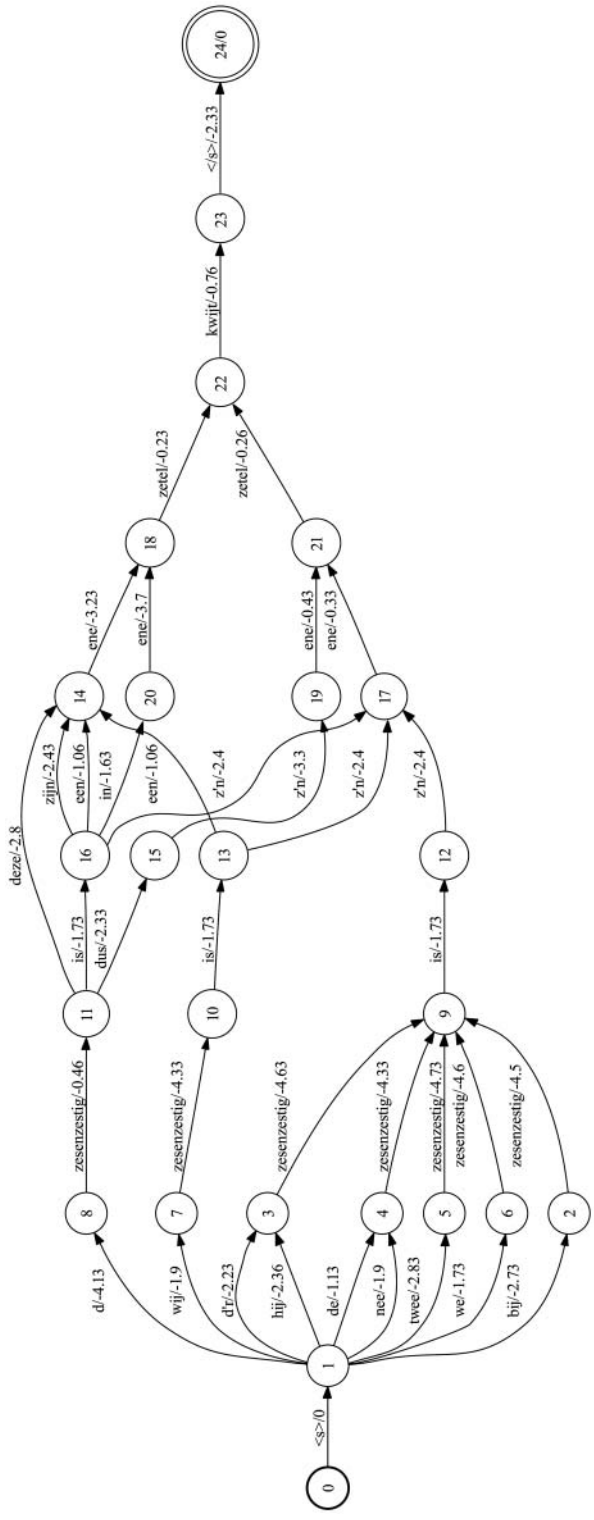

FIGURE 2 An example of a lattice structure for a sentence.

Research topics in user interface development

When searching for fragments of speech, what do end users need? Which kinds of questions do they pose? How can we help them to find the fragments of their interest efficiently? At the start of the CHoral project, there were hardly any studies into user requirements for searching audiovisual archives in the cultural heritage domain. Nor were there any working examples of online access to spoken word documents using automatically generated content representations. To get started with the development of the CHoral user interface, we therefore carried out an initial requirements analysis, and launched a first demonstrator system of online, word-level search in a spoken word collection: the 'Radio Oranje' search engine. It was the starting point for the development of CHoral's spoken document retrieval framework and our first testbed for a series of usability studies.

The Radio Oranje demonstrator system

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[3] (NIOD) and Sound and Vision[4] (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[5] provides an example of how state-of-the-art visualization and indexing technology can boost the accessibility and enliven the perception of such collections (see Figure 3).

The text versions of the speeches were synchronized with the audio using an alignment tool. Alignment uses an automatic speech recognition system to recognize where which utterance occurs, while the grammar and dictionary

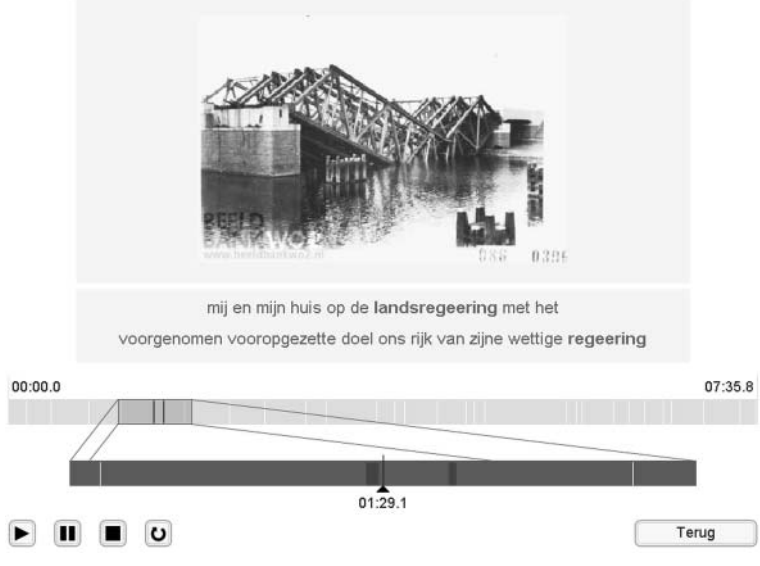

FIGURE 3 Audio playback page for the Radio Oranje system.

are restricted by the words from the transcript. The speeches were aligned at the word level and evaluation showed that over 90% of the word boundaries were found within a span of 100 ms of the reference. 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., Whittaker *et al.* 1998), and facilitate navigation in audio (e.g., Ranjan *et al.* 2006). The CHoral timeline visualization may contain different types of information (including segment boundaries and query term locations). This visualization 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 students from the Humanities Department 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 the 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.

The Buchenwald demonstrator system

A successor of the Radio Oranje demonstrator is the interface developed for the 'Buchenwald' website,[6] a Dutch multimedia information portal on World War II concentration camp Buchenwald (Ordelman *et al.* 2008). The collection, maintained by the Netherlands Institute for War Documentation, holds both textual information sources and a video collection of testimonies from 38 Dutch camp survivors. For each interview with a duration of 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 4), and a link to the video file and the elaborate description (see Figure 5). For video navigation, the CHoral timeline visualization was used, and the latter functionality is currently being improved to differentiate between markings relevant at the document level and those relevant within the document, at the fragment level.

A first 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 websites 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

FIGURE 4 Result listing for the Buchenwald collection.

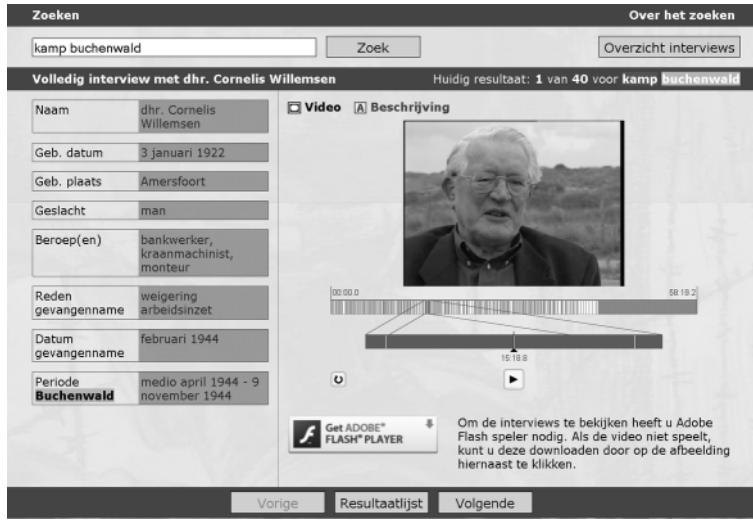

FIGURE 5 Video playback for the Buchenwald collection.

or two fairly general, but topic-related terms. Moreover, when a user typed a query, it contained a named entity in almost 60 per cent of the cases. We also found that the functionality to access the interviews and the related texts is 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 and archivists as to the use, possibilities and restrictions of such technology for disclosure of audiovisual archives.

Next steps for CHoral's user interface

The next step in development of the user interface within CHoral is scaling up from relatively small, homogeneous collections to large and heterogeneous collections such as Radio Rijnmond's archive. Query logs showed that users often do not formulate a query, but ask for a list of documents to begin their exploration of a collection (e.g., Ordelman *et al.* 2008). Whereas content listing is a feasible approach for a collection of up to several dozens of documents, it is not useful in the setting of an entire audiovisual archive. Therefore, instead of a complete list of documents, document clustering (by, for example, year of production, creator, topic) could provide users with a way of exploring an archive's contents. This paradigm is already being used by archiving institutes, such as the Netherlands Institute for Sound and Vision, 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 may be incomplete. This can result in users being unable to find fragments that are present in the collection (Carmichael *et al.* 2008a). Named entities are particularly at risk of becoming irretrievable, but, at the same time, they are very popular query terms. The standard way of reducing this problem is through query expansion (e.g., Jourlin *et al.* 1999). 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 to them 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 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 per cent, users have been found to discard textual content representations (e.g., Stark *et al.* 2000). As an alternative to presenting low-quality transcripts, it has been proposed to use keyword extraction approaches (e.g., Haubold and Kender 2004), and content visualizations (e.g., Kimber *et al.* 1995; Whittaker *et al.* 1999). In CHoral, suitable ways of presenting only selected keywords from the transcripts are currently being explored. Experience up till now has taught that user assessments of the various options for content representations are needed, and that the relation between the accuracy of automatically generated indexes and the user experience should be monitored very closely for interface design in this domain.

How to fit spoken document retrieval into the archiving workflow

On the basis of the experiments and assessments conducted in CHoral, the question of whether spoken document retrieval 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 question of which conditions 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 taking up of spoken document retrieval technology in audiovisual archives: automatic speech recognition performance and archival infrastructure. 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 though. We have proposed some research tracks that could lead to improved system performance, however, 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.

Archival infrastructure is a bottleneck as well, despite the fact that mass digitization is under way and standard metadata schemes as well as trusted digital repositories are being developed. Apparently, the archiving of digital, audiovisual documents requires the integration of existing and new technology, a process that is not yet fully understood. Moreover there seems to be a knowledge gap, especially in the smaller institutes. In general, archiving of digital audio already 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 from the average collection keeper. This could be taken as an argument for setting up joint web-based services as part of the archival infrastructure and have these services run by experts that can take over the tasks that are not easily taken up in archival institutes.

Although evidence suggests that the two disciplines are at work on the bottlenecks separately, experts from the various fields involved in the life cycle of spoken documents are increasingly in touch with each other. This may well lead to further improvements in disclosure and access. And of course, the emerging collaboration can be reinforced by well-chosen concerted action, as is demonstrated by IST project Prestospace[7] and Coordination Action CHORUS.[8]

Looking for synergies on a smaller scale can produce results. Archives are continuously developing tools and methods in cooperation with third parties in order to offer wider and more coherent access to their collections. The take-up of technology could therefore best be stimulated by a strategy that on the one hand incorporates this model of collaboration, and on the other hand stimulates the dissemination of best practices in spoken document retrieval. The latter should help archives to assess if a certain technology is suited for the kind of content they want to make available to their users, and whether the cost-benefits involved suit their resources. This can help monitoring expectations and could prevent disappointment and the inherent risk of permanent loss of interest. Also guidance on how to optimize chances for

good automatic indexing performance during the various stages of the handling of the data is crucial.

A few examples may underline this: at the stage of data capture, producers of content should be made aware of the fact that the circumstances under which the audiovisual material is being produced is of crucial importance for the quality of the automatic indexes, and thus the results returned during subsequent search. At the data acquisition stage, archivists should check if newly received material meets their standard metadata scheme and plan updates where needed. The problem of out-of-vocabulary terms has led the speech researchers to propose to archivists that they incorporate the names of places and persons associated with multimedia documents in the manually produced annotation. Alternatively, collateral data such as scripts and notes from producers of multimedia documents may provide this information, which means that links between related documents should be incorporated in the metadata (Heeren *et al.* 2008). Also, related textual resources may be used to train better models for automatic speech recognition.

An important caveat is that this additional workflow is likelier to increase the amount of manual work than to reduce it. This holds in particular for the kind of collections that are kept and being indexed for non-commercial purposes, and for which human dedication and the level of sophistication of the available background knowledge are important features of the indexing capacity. In such contexts, the added value of new technologies is not so much in the cost-reduction but 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.

Interaction between researchers and archivists, and even better, collaborative projects have proven a vital instrument for increasing the chances for the take-up of advanced technologies. At least in the case of CHoral, the CATCH-model has helped the Rotterdam Municipal archiving staff, as well as the associated parties such as the Netherlands Institute for War Documentation, the Veterans Institute[9] and the International Information Centre and Archives for the Women's Movement,[10] to develop new ideas about how to present their content to their users. In return, the archivists have informed researchers of the crucial role of archival standards and values, including the importance of carefully presenting the context of search results to end users.

Limitations and challenges inherent to the handling of spoken word collections along the lines advocated above are still likely to be discovered in the collaboration that we see ahead, but among the parties collaborating in the context of CHoral, a greater awareness of the chances and possibilities for (re-)use of the materials has clearly been established.

Conclusion

The continuing cooperation between the researchers and archivists in the NWO-CATCH project CHoral will, in the near future, result in online access to a large and heterogeneous collection of audiovisual materials, i.e. the Radio Rijnmond collection mentioned in the Introduction. In addition, we will seek

ways to deploy the tools and best practices developed for this particular broadcast archive to other spoken word collections, and for possibilities to contribute to the realization of an infrastructure that can enhance the access and reuse of spoken word materials. It will be a test case for further development and fine tuning of the instruments developed by both parties. The end goal is to better serve the archive's existing clients and to attract new ones by offering innovative services.

Acknowledgements

This paper is based on research carried out in the project CHoral — Access to Oral History, which is funded by the NWO programme CATCH (http://www.nwo.nl/catch, 4/2/09). We would like to thank the Netherlands Institute for War Documentation and the Netherlands Institute for Sound and Vision for their cooperation in the development of the CHoral demonstrator systems.

Notes

[1] http://hmi.ewi.utwente.nl/choral/ (5/3/09)
[2] http://wwwhome.cs.utwente.nl/~huijbreg/demo-pages/hermans/hermans.php (5/3/09)
[3] http://www.niod.nl (5/3/09)
[4] http://portal.beeldengeluid.nl/ (6/3/09)
[5] http://hmi.ewi.utwente.nl/choral/demo (5/3/09)
[6] http://vuurvink.ewi.utwente.nl:8080/Buchenwald (5/3/09)
[7] http://prestospace.org/ (4/2/09)
[8] http://www.ist-chorus.org/ (4/2/09)
[9] http://www.veteraneninstituut.nl/ (4/2/09)
[10] http://www.iiav.nl/ (4/2/09)

Bibliography

Amato, G., J. Cigarrán, J. Gonzalo, C. Peters, and P. Savino. 2007. MultiMatch — Multilingual/Multimedia access to cultural heritage. *Proceedings of the 11th European Conference on Research and Advanced Technology for Digital Libraries*, 505–8. Berlin, Heidelberg: Springer-Verlag. Budapest, Hungary.

Byrne, W., D. Doermann, M. Franz, S. Gustman, J. Hajic, D. Oard, M. Picheny, and J. Psutka. 2004. Automatic recognition of spontaneous speech for access to multilingual oral history archives. *IEEE Transactions on Speech and Audio Processing* 12(4): 420–35.

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

Carmichael, J., M. Larson, J. Marlow, E. Newman, P.D. Clough, J. Oomen, and S. Sav. 2008b. Multimodal indexing of electronic audio-visual documents: A case study for cultural heritage data. *Proceedings of CBMI 2008*. London, UK.

Chelba, C., and A. Acero. 2005. Position specific posterior lattices for indexing speech. *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*. Ann Arbor, USA.

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

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

Gustman, S., D. Soergel, D. Oard, W. Byrne, M. Picheny, B. Ramabhadran, and D. Greenberg. 2002. Supporting access to large digital oral history archives. *Proceedings of the Joint Conference on Digital Libraries*, 18–27. Portland, Oregon, USA.

Hansen, J.H.L., R. Huang, B. Zhou, M. Deadle, J.R. Deller, A. R. Gurijala, and M. Kurim. 2005. SpeechFind: Advances in spoken document retrieval for a national gallery of the spoken word. *IEEE Transactions on Speech and Audio Processing* 13(5): 712–30.

Haubold, A., and J.R. Kender. 2004. Analysis and visualization of index words from audio transcripts of instructional videos. *Proceedings of the IEEE Sixth International Symposium on Multimedia Software Engineering*, 570–3. Miami, Florida, USA.

Heeren, W.F.L., R.J.F. Ordelman, and F.M.G. de Jong. 2008. Affordable access to multimedia by exploiting collateral data. *Proceedings of CBMI 2008*, 542–50. London, UK.

Huijbregts, M.A.H., R.J.F. Ordelman, and F.M.G. de Jong. 2005. A spoken document retrieval application in the oral history domain. *Proceedings of 10th International Conference Speech and Computer*, 2. 699–702. Patras, Greece.

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

Kimber, D.G., L.D. Wilcox, F.R. Chen, and T.P. Moran. 1995. Speaker segmentation for browsing recorded audio. *Proceedings of CHI 1995*, 212–13. Denver, Colorado, USA.

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

Matsoukas, S., J-L. Gauvain, G. Adda, T. Colthurst, C.L. Kao, O. Kimball, and L. Lamel. 2006. 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): 1541–56.

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

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

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

Ranjan, A., R. Balakishnan, and M. Chignell. 2006. Searching in audio: the utility of transcripts, dichotic presentation and time-compression. *Proceedings of CHI 2006*, 721–30. Quebec, Canada.

Siegler, M. 1999. Integration of continuous speech recognition and information retrieval for mutually optimal performance. PhD diss., CarnegieMellon University.

Stark, L., S. Whittaker, and J. Hirschberg. 2000. ASR satis-ficing: the effects of ASR accuracy on speech retrieval. *Proceedings of the International Conference on Spoken Language Processing*, 1069–72. Beijing, China.

van der Werff, L.B., and W.F.L. Heeren. 2007. Evaluating ASR output for information retrieval. *Proceedings of the ACM SIGIR Workshop on Searching Spontaneous Conversational Speech*, 7–14. Amsterdam, The Netherlands.

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

Wessel, F., R. Schluter, and H. Ney. 2000. Using posterior word probabilities for improved speech recognition. *Proceedings of the 2000 IEEE International Conference on Acoustics, Speech, and Signal Processing* 3: 1587–90.

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

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

Notes on Contributors

Correspondence to: Willemijn Heeren, University of Twente, Department of Electrical Engineering, Mathematics and Computer Science, Cluster: Human Media Interaction, P.O. Box 217, NL-7500 AE Enschede, The Netherlands.
Email: w.f.l.heeren@ewi.utwente.nl