Disclosing Spoken Culture: User Interfaces for Access to Spoken Word Archives

Willemijn Heeren
Human Media Interaction,
Faculty of Electrical Engineering, Mathematics and Computer Science,
University of Twente
+31102434619
w.f.l.heeren@ewi.utwente.nl

Franciska de Jong
Human Media Interaction,
Faculty of Electrical Engineering, Mathematics and Computer Science,
University of Twente
+31534894193
fdejong@ewi.utwente.nl

ABSTRACT
Over the past century alone, millions of hours of audiovisual data have been collected with great potential for e.g., new creative productions, research and educational purposes. The actual (re-)use of these collections, however, is severely hindered by their generally limited access. In this paper a framework for improved access to spoken content from the cultural heritage domain is proposed, with a focus on online user interface designs that support access to speech archives. The evaluation of the user interface for an instantiation of the framework is presented, and future work for the adaptation of this first prototype to other collections and archives is proposed.

Categories and Subject Descriptors
H.5.1 Multimedia Information Systems, H.5.2. User Interfaces.

General Terms
Performance, Design, Experimentation, Human Factors.

Keywords
Spoken document retrieval; Speech browsing support; User evaluation; Audio transcripts.

1. INTRODUCTION
Although oral culture has been part of our history for thousands of years, we have only fairly recently become able to record and keep that part of our heritage. Over the past century alone, we have collected millions of hours of audiovisual data. Audiovisual archives in the cultural heritage (CH) domain maintain these large numbers of spoken word collections with great potential for e.g., new creative productions, research and educational purposes. The actual (re-)use of these collections, however, is severely hindered by their generally limited access, e.g., [20]. Firstly, whereas catalogs are in an increasing number of cases suitable for online search, only document descriptions of audiovisual objects, but not the recordings themselves can be retrieved and accessed. Secondly, once users gain access to the recordings (here called documents) – by visiting the archive and requesting a copy – their exploration remains cumbersome as descriptions are often insufficiently specific and entire documents instead of fragments are retrieved. As a result, researching audiovisual sources is an effortful undertaking for e.g., the creative industry and scholars.

It has become increasingly apparent that fully manual indexing techniques for providing access to spoken word collections from CH at the level of fragments can affordably cover only a small fraction of the potential need. Therefore, automatic indexing using e.g., speech and language technology has increasingly received attention. In 2002 the DELOS/NSF working group in Spoken Word Audio Collections met to discuss 'an agenda in the area of spoken word archives and collaborative research projects' [7]. A number of items from that agenda concerning the use of technology have since been taken up in research projects.

One initiative was the MALACH (Multilingual Access to Large spoken ArCHives) project, a US NSF project (2001-2007), that investigated access to a vast collection of testimonies from Holocaust survivors, witnesses and rescuers [26]. The goal of that project was to advance Automatic Speech Recognition (ASR) for the oral history domain and to study how recognition can be best incorporated in further processing and retrieval steps [8]. Another project that contributed to advancing spoken document retrieval in the cultural heritage domain was The National Gallery of the Spoken Word project. In that project the SpeechFind spoken document retrieval system was developed: it automatically generates metadata for audio documents by segmenting the audio and generating ASR transcripts, and also makes the audio searchable though a Web-interface [9].

There are also a number of initiatives in spoken document retrieval in the cultural heritage domain for languages other than English. For example, the IST-FP5 project ECHO (European CHronicles Online) aimed at the realization of a searchable multilingual collection of video documentaries deploying speech recognition as one of the core technologies. The main focus of most earlier research projects on spoken document retrieval in the CH domain, however, has been the development of indexing technology. Now that automatically generated indexes are becoming of sufficient quality to support search, and the digital audio can be directly linked to online search results, a user interface should be developed that supports online search in speech archives.

In the CHoral project (http://hmi.ewi.utwente.nl/choral), part of the CATCH programme funded by the Netherlands

Organisation for Scientific Research (NWO), the goal is to provide users of speech archives with *online* access to relevant *fragments*. In the framework we propose for improved access to spoken CH-content three dimensions are distinguished: (i) automatic metadata generation, metadata enrichment and indexing, (ii) spoken document retrieval, and (iii) a user interface. For index generation, the framework relies onspeech-to-text conversion techniques from ASR and audio processing. The spoken document retrieval aspect (SDR) is addressed by research aiming at selecting the optimal unit of retrieval, and by developing proper performance evaluation metrics for SDR.

In this paper, we present issues related to the development of a user interface that supports users during the search process. Attention is given to both interface design and to the evaluation of user interfaces for access to speech archives with cultural heritage content. We will first present related work on user interfaces for access to spoken word documents. Next, we present the prototype online search engine for a spoken word collection from Dutch CH that is an instantiation of a speech retrieval system that meets the framework's goals of *online, within-document* access. Then, an evaluation of the functionalities of the prototype's user interface is presented. Next, we will discuss how the prototype can be adapted to encompass other (types of) collections.

2. USER INTERFACES FOR ACCESS TO SPOKEN WORD DOCUMENTS

In related work spoken document retrieval systems have been developed for different domains, such as broadcast news [37], voicemail messages [31], webcasts [25], meeting recordings [35], historical recordings of speeches and broadcasts [9], and oral histories [26]. Some of these systems have been subjected to user evaluations. Their interface solutions and evaluations are discussed in the rest of this section.

We will present this discussion according to three main stages in access functionality: collection access through searching or browsing, selection of a possibly relevant result, and playback of the selected result.

2.1 Searching and Browsing

Online search engines that offer the possibility of audio search, typically have a user interface that is structured in the same way as their counterparts built for text search: after typing a number of keywords, results are shown in ranked lists of about ten titles per page. Most queries to these engines, however, are meant for the searching of music instead of spoken word documents [15]. For accessing spoken word documents, the search engines currently available online were mostly developed in research projects, e.g., [9].

Search in spoken word documents is often also done through text search by matching the query to the available, textual metadata. In addition to high-level semantic information such as a document's title or recording date, more elaborate content descriptions can sometimes be exploited. The metadata may be either manually generated, in the form of keyword assignments, short descriptions, summaries or full transcripts as in many sound archives and oral history collections, but it may also be in the form of automatically generated transcripts as in a number of research projects. Additional information to improve search performance can be drawn from thesauri or ontologies, and from related documents, such as so-called collateral data [5].

To support searching for relevant *fragments* in spoken word documents a time-stamped index to a textual representation of the words spoken is needed. That index can be based on either ASR transcripts, or on aligned manual transcripts. The disadvantage of ASR transcripts is that they are not error-free, but in many cases full, manual transcriptions are not available since it is extremely time-consuming – and therefore costly – to make them.

2.2 Presentation of Results for Selection

In search environments, links to retrieved documents are usually listed, and presented together with at least some basic metadata information (e.g., title, production date). More extensive metadata presentation may be both textual (e.g., summaries, excerpts) and visual (e.g., content visualizations, use of color). A paradigm underlining this is the *what you see is almost what you hear* principle that prescribes the use of a visual analogue to the speech content [37]. For the SCAN system, which gave access to a broadcast news archive, it was shown that multimodal presentation of the audio content (ASR transcripts and a visual content representation showing the occurrence of query terms in time) helped users to find facts and also decreased the overall search time [37].

An exploratory study of alternative relevance criteria that users employ to judge search results for audio indicated that topics and summaries were found helpful [8]. Also, information on the genre of the audio (e.g., interview, debate, report) was judged relevant as well as time information: both the time frame of the audio and its recency [17]. The relative usefulness of document characteristics, however, is expected to depend on the type of data. For instance, a document's recency seems more relevant in news than in oral histories.

Furthermore, automatically generated links to related items within or outside the collection, or to related collections, may enhance the users' experience and deepen their understanding of the result list, e.g., [23]. Another possibly useful addition to result lists are ratings given by previous users (e.g., YouTube or GoogleVideo) or even personalized preferences in the case of frequent users. These information sources aim to guide the user in selecting possibly relevant spoken word documents before they have even listened to their content.

2.3 Audio Playback

Once a user has selected a particular spoken word document, (s)he typically wants the fragment of interest to be played. Basic playback options (start–stop–pause) do not seem to suffice for navigation in spoken word documents. Therefore, more elaborate and more interactive user interfaces have been developed, that on the one hand help the user in building a mental model of the spoken content, and that on the other hand facilitate navigation in audio.

To give users a grasp of an audio document's content, visual representations have been developed that indicate speaker turns (e.g., [18,30]), or the location(s) of query terms in time (e.g., [37]). Next to visual representations, often transcripts are presented. Low-quality transcripts, i.e. with Word Error Rates over 30%, may be discarded by users [28,31], whereas high-quality transcripts can support them to perform tasks faster and was found to reduce the amount of audio that was played in the browsing process [31]. Moreover, the usability of high-quality transcripts was rated higher than of low-quality transcripts (e.g., [25,31]).

Tools developed for faster browsing allow users to speed up audio playback, since time-compressed speech remains intelligible up to double its original speed, e.g., [12,28]. Moreover, users have been found to prefer to take control of

audio playback over predetermined play durations, since restricted playback would stop at unpredictable places [36]. A more minor problem that may be encountered with playback options is when query terms occur right at the beginning of the retrieved fragment and it has been played before users are well-aware of it. In the W.F. Hermans project [13], this problem was overcome by allowing users to select the size of the fragment's context before playback.

3. RADIO ORANJE SEARCH ENGINE

The 'Radio Oranje' search engine is an instantiation of the access functionality envisioned in the CHoral project. It is used for user testing, and for further development and adaptation to other collections and users. The prototype search engine supports access to the collection of radio speeches that Queen Wilhelmina (1880-1962) addressed to the Dutch people during World War II. It is one of very few historical collections fully accessible via the Web. The rest of this section will introduce the search system (see [33] for a more elaborate explanation).

3.1 Metadata Generation and Indexing

The collection consists of 37 speeches with lengths varying between 5 and 19 minutes. Their style is very formal and language use is complex. Moreover, the audio quality was generally poor, since the old recordings contained much pops and hiss. The recordings as well as their 1940s transcripts have been digitized by the Netherlands Institute for War Documentation (NIOD) and the Netherlands Institute for Sound and Vision, which is one of Europe's largest audiovisual archives. Since both these sources were available, a word-level index could be generated through alignment of the text and the audio.

Alignment is the process of using an ASR system to recognize an utterance, where the words occurring in the utterance, but not their timing, are known beforehand. The result is a set of time-aligned word labels. Due to the low complexity of the task – everything is known except for the time labels – alignment is relatively robust against mismatches in acoustic conditions. This procedure has also been used in other systems, e.g., [2,3]. Another approach to link spoken content and text was taken by [19], who added bar-codes to paper transcripts to create a way of linking interview transcripts to direct video access on a PDA.

As a preprocessing step, all speeches were manually segmented at the sentence level, giving a total of 853 sentence-sized segments with an average length of 15.7 seconds. For evaluation purposes, two full speeches were segmented at the word level yielding 2028 manually aligned word boundaries. Optimal alignment performance was obtained using speaker-dependent, monophone acoustic models, trained from gender- and speaker-independent models optimized for broadcast news [27]. Performance was adequate for the generation of a word-level index into the collection: >90% of all word boundaries were found within 100 ms of the reference, i.e. within the correct syllable.

3.2 User Interface

To support exploration of this spoken word collection, the user interface (UI) allows entry to the collection at two levels: an entire speech or a speech fragment (see Figure 1). Users unfamiliar with the collection can ask for a list of all speeches by hitting a button labeled 'Show speeches', and can subsequently explore the audio by following links to entire speeches. Users with specific search questions can enter search terms, and then a list of fragments containing those search terms is returned, with the links to the beginning of the

Figure 1. Screenshot of the search page: (a) enter a search term or show all speeches, (b) optionally set date restrictions, and (c) examine the result list.

fragment. Moreover, users can optionally restrict the period within which they want to search documents or fragments by setting the temporal interval they are interested in.

Retrieval is currently supported by standard (Boolean) search and query results are ranked by date showing the speech's title, broadcast date and duration, as well as an excerpt of the relevant sentence fragment (in case of a query) or of the beginning of the speech (in case the 'show speeches' button was used).

Once the user selects a particular result, the playback page is shown (see Figure 2). It has (i) an interactive timeline that also shows locations of relevance, (ii) playback buttons, (iii) an extra button for restarting the fragment from the entry point, since a query term may have been played before users are well-aware of it, (iv) running subtitling, and (v) a running photo presentation.

The audio starts playing at the beginning of the sentence containing the query term, while the entire speech is given as context. In the case of spoken word documents from cultural heritage – but also in many other cases – providing context is imperative for allowing the listener to correctly understand the meaning of the speaker's utterance.

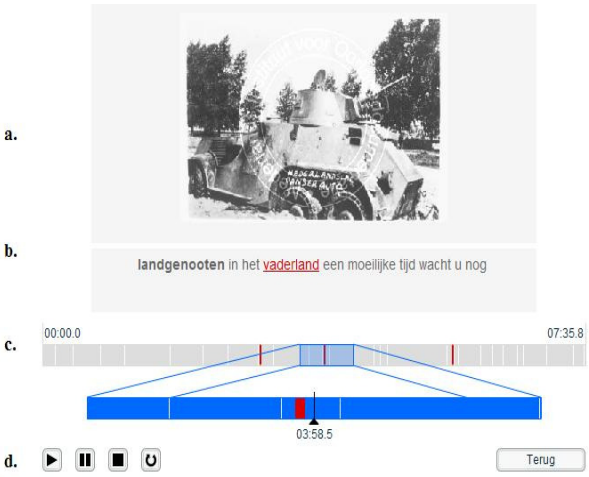

Figure 2. Screenshot of the playback page: (a) related photos, (b) subtitling, (c) interactive timeline, (d) playback buttons with an additional fragment-restart button.

Subtitling is shown while the audio is playing: it highlights the word being spoken, and shows the query term(s) in bold. Our way of transcript presentation, i.e. subtitling, differs from those used in earlier user studies (e.g., [37]) in that we only show the current sentence instead of the full text or a paragraph. We thus aim to support perception of the old-fashioned speech, while keeping the information focused to the current fragment a user is listening to.

The *interactive timeline* gives an overview of the entire speech as well as a zoomed-in view of a 45s window around the cursor. The exact locations of query terms and sentence boundaries are shown in both bars. Furthermore, both bars are fully clickable for within-document navigation; the lower one provides the used with more precise control as it visually stretches the interval that is being played.

Finally, the speeches were semi-automatically linked to *photos on the same topic*. These are downloaded in real time from a collection of over 55,000 photos maintained by NIOD, to provide the user with a more attractive and more informative experience.

4. EVALUATION

The evaluation of our prototype´s user interface is divided into two parts. Since the search engine is online and publicly available, as part of a section on Queen Wilhelmina at the NIOD's website, we could log data from actual users. The log was restricted, however, to users' interactions with the search interface (figure 1). To study the users' interaction with the playback functionality (figure 2) we designed user experiments. The following subsections will present these evaluations.

4.1 Use of the Search Option

We analyzed the user logs of the prototype search engine that was online available to anyone via the website of NIOD. The analysis presented here was run on data collected between July 1 2007 and January 10 2008. We filtered out days on which we demonstrated the system during presentations, and also excluded requests coming from our own department's IP-addresses.

4.1.1 Results

A total of 1243 requests were made, spread over 325 sessions. This means that there were about 11 sessions per week. This number is comparable to the numbers of requests that weekly reach keepers of Dutch audiovisual archives [11]. Among those 325 sessions we counted 273 different IP addresses, meaning that some users revisited the search website.

When users started a search session, they used the 'Show speeches' button in 205 cases (63%), and typed in a query in 120 cases (37%). This indicates that the addition of that button to allow users to explore the collection was needed. Over all requests, 479 queries were typed in and the 'Show speeches' button was used 764 times, only minimally changing the ratio of use for the two manners of entering the collection. As for the queries that were formulated, the median length was 1 word, with a range of 1 to 6 words. The most popular search terms were fairly general war-related terms such as 'jew', the Dutch nickname for the German invader, names of members of the Dutch royal family, and Dutch city names. Users only 21 times changed the date settings to restrict the search space.

After about one third of the requests, 483 times, a user clicked a link to an audio file. The listening duration was estimated from the durations of those listening actions that were not the end of a session, since session ending times were not registered. We had already excluded listening durations under 10 seconds (314 cases), since those are considered too short for a listener to actually examine the audio file. We expect those to be mainly due to listeners clicking the link twice, or becoming impatient while the audio file was being loaded. The distribution of listening times (in minutes) is given in Figure 3. Listening sessions longer than one hour are not shown in the figure (17 cases).

In about half of the listening actions, users examined the audio for just a few minutes, i.e. they listened to a fragment of a speech. As for the other half of the listening actions, they listened for the duration of an entire speech (5-19 minutes). We cannot know however, whether listeners replayed a fragment several times, or simply let the audio play to the end of the speech.

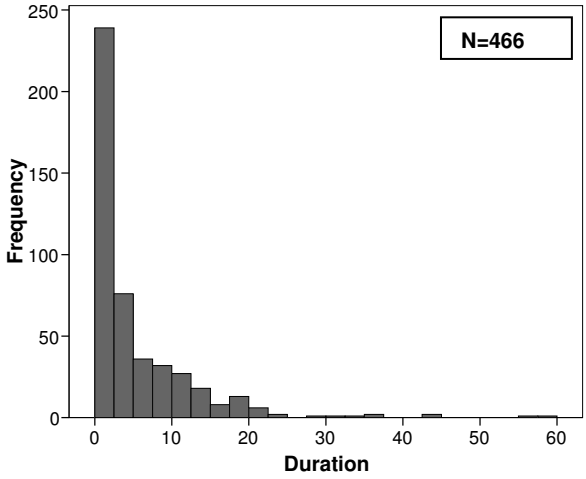

Figure 3. Histogram of the listening durations per speech, given in minutes. Listening actions less than ten seconds and over one hour were left out.

4.1.2 Discussion

Our interpretation of these log statistics is that most users of the online prototype did not use the search engine to fulfill a preset information need, but used it to get general information on the collection as a whole. We base this interpretation on the frequent use of the 'Show speeches' button instead of the formulation of a query, and also on the fact that the queries that were formulated were in most cases fairly general, as users typed only one term in the majority of cases. However, using only one or two words that are likely to have been mentioned during speeches from the war period is a safe strategy for users to get initial results, and explore the collection from that starting point.

4.2 Use of the Playback Option

The evaluation of the playback option was addressed by running small-scale experiments to test the intended use of its main components (see Figure 2): subtitling, and the insertion of location marks in the interactive timeline visualization. Based on these first experiments, larger scale user evaluations will be planned.

4.2.1 Participants

Ten students (5 males, 5 females; age range 19-27) from the departments of History and Linguistics of Leiden University in The Netherlands took part in 50-minute, individual sessions for

which they received 10 euro. These users were chosen as they are taken to represent one of our target user groups, i.e. researchers in the humanities.

4.2.2 Procedure

Participants first filled out a demographic questionnaire. Next, they completed two tests in the order as they are described below. Both tests were run in within-subjects designs and the order of conditions was counterbalanced across subjects. We measured both performance and perception. Tests were run in a quiet classroom using a standard PC setup (17" monitor, keyboard, mouse) and Sennheiser HD 477 headphones. During testing, the running photo presentation was replaced by a constant photo of the queen in order to avoid influences from the changing photos on the interface components being tested.

4.2.3 Task 1: Within-document Navigation

Location marks were added to the interactive timeline to support easy navigation within a spoken word document by showing the exact locations of search terms and sentence boundaries, and it was also intended to give an overview of relevant locations throughout a document (see the timeline in Figure 2). We wanted to examine whether *within-document* navigation is supported by such marks along the timeline, without prior explanation of its functionality. Based on earlier work on visual overviews for speech documents by [37], who however did not integrate the overview and the audio player, we hypothesized that users would be more successful at navigating within a document if marks are shown, in comparison with the normal playback situation in which no relevant locations are given. We also expected them to perceive the task as being easier with than without location information.

To address the use of showing the locations of search terms and sentence boundaries, subtitling was hidden in either condition, and user performance was measured on within-audio search tasks either with or without the location marks present. The participants' task was to transcribe an entire, spoken sentence which they located within the collection by searching for a term specific to the sentence. The assumption was that more successful within-document navigation while locating a particular sentence would lead to more correct transcripts within a restricted amount of time.

Seven terms were chosen such that they occurred only once in the collection and in long sentences (22–27s) that cannot be transcribed correctly after having been played once. Participants were thus forced to navigate through the audio and to restart the fragment several times. Users got maximally 6 times the sentence duration for transcription plus 20s to retrieve the correct audio file. The longest sentence was used for practice in both conditions, the rest was used for testing: three tasks were performed with location marks, and three were done without. On-screen activity was recorded by means of screen-capture software. Task duration was about 20 minutes.

4.2.4 Task 2: Perception of Old-fashioned Speech

Subtitling was added with the intention of aiding the perception of the old-fashioned speech, while keeping the information focused to the current fragment a user is listening to. Perception was difficult due to the poor audio quality of the old recordings in combination with the old-fashioned, formal language use in the speeches of Queen Wilhelmina. For instance, a typical sentence from this collection –translated from Dutch– was: "Never has there been a more confident condemnation of a system and its barbarous application by a tyrant that tries to impose his will on millions".

Given these semantically complicated sentences, we had decided not to address comprehension by measuring gist recall, but by measuring direct recall, a measure to assess the listeners' short term memory for the literal spoken content: the better it has been coded, the better recall. However, even though subtitling may be considered necessary to avoid misperception of the noisy formal speech, it may also increase the user's cognitive load [e.g., 21,22]: it has been shown that simultaneous presentation of the same information as both text and audio increased cognitive load in learners, since they had to divide their attention between the two modes [16]. Therefore, recall might suffer from the addition of subtitling as listeners must split their attention.

We hypothesized that the presentation of subtitling text while the audio is being played helps users to process the spoken content better, despite a possible increase in cognitive load. This will be reflected by higher percentages of correctly remembered words in the subtitling condition.

Twenty-four sentences with lengths of 8 to 12 seconds were played to users in the playback UI. Subjects were shown the subtitling for the first half, but no subtitling for the second half of the sentences, or the other way around. After a sentence had been played, a sequence of words from the sentence was printed on screen and participants were asked to recall the next three to six words. After writing down their response, they were also asked to indicate for every sentence how well they were able to hear the speaker. The answer sheet indicated the required number of words to be filled in per task. Half the tasks asked for word sequences from the beginning of a sentence, the other half asked about the end. Moreover, sentence order was pseudo-randomized and task duration was about 14 minutes.

4.2.5 Results

The pretest questionnaire showed that all participants used computers on a daily basis, had experience with well-known, commercial search engines and regularly searched catalogues of CH institutes, mostly libraries and archives. Three out of ten participants indicated to also search for audio and video.

Since the number of ten users used in this study is relatively small, despite the within-subjects design, we only used non-parametric statistical tests to study effects of location marks in the timeline and of subtitling.

4.2.5.1 Interactive Timeline

Performance: To study the effect of adding term location marks to the interactive timelines, we computed the percentage of correctly transcribed words per search task, and subsequently determined an average across tasks per participant, and per condition.

Looking at figure 4, the main performance difference is found when no location marks are shown ($Z=-2.0$, $p<.05$). If that is the first condition listeners are confronted with, they perform moderately; they get 70% of the words correct. If it is their second condition, they reach 86% correct. With location marks present, there is no significant performance difference (86% v. 87% correct). These results seem to indicate that users can efficiently use the timelines with annotations for navigation, and they do so with minimal training. When using the control condition, however, users first need to get used to the system to use it successfully.

Perception: After completing each condition participants indicated their agreement with statements on the UI's perceived usefulness, and the ease of navigation. Agreement was measured on a scale of 1 (low) to 5 (high). The usefulness of

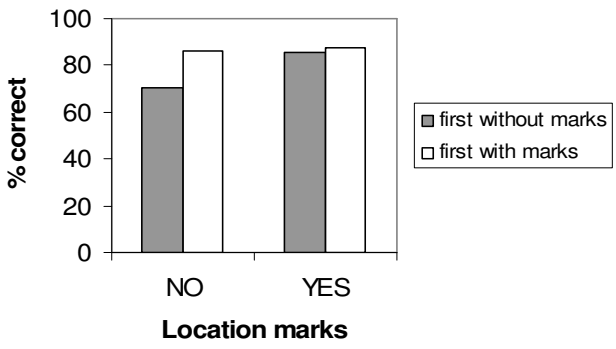

Figure 4. Percentages of words correct as a function of the presence of location marks and task order.

the timelines including location marks was, as expected, rated higher than the one without marks, Z=-2.2, p<.05. The median scores were 4.5 with and 3.0 without marks. Though the median of the navigation ease score was higher with marks, 5.0, than without marks, 3.5, the difference was not found to be significant.

After the task was completed two users reported being confused about not being able to drag the cursor from one point to another in the timelines. They explained their confusion as being due to the difference in functionality between our timeline (click only) and that of the audio players they are used to (click or drag), such as Windows Media Player and iTunes' player.

In the users' sentence transcriptions a number of misperceptions were found, supporting our assumption that textual support is needed to correctly perceive the speech in the old recordings. In the next section the results on the subtitling task are presented.

4.2.5.2 Subtitling

Performance: To examine the effect of subtitling on the recall of the spoken content the average proportions of correctly recalled words were computed per participant and per condition.

An effect of sentence position, i.e. whether the words to be recalled came from the begin or the end of a sentence, was present, Z=-2.8, p<.01, see Figure 5. Participants better recalled words from sentence endings (95%) than from beginnings (44%); this might be explained by recency, but predictability of sentence endings given the beginnings might also have contributed to this difference. However, no effect of subtitling on users' performance in recall was found.

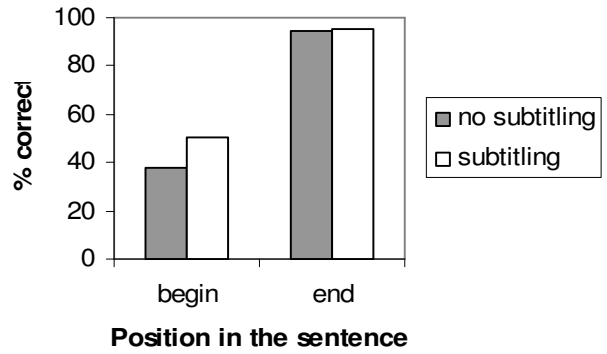

Figure 5. The percentage of correctly recalled words as a function of sentence position and the presence of subtitling.

Perception: Perceived audibility was measured on a scale of 1 (poor) to 5 (good) for each sentence. With subtitling its median was 4, whereas it was 3 without subtitling. This difference was not significant, however, which corresponds to the lack of a performance effect.

Participants' reactions after the task were both that "textual support is very important" and that "it was more difficult with text because you are trying to attend to two things at the same time", although that user indicated to "prefer the text condition".

4.2.6 Discussion

The results of the timeline test were partly in line with expectations: location marks helped listeners to transcribe more words correctly, which we take as evidence for faster and more successful navigation, but with practice users could show similar performance using the control system. The timelines in our prototype search engine support users while navigating within documents right from the first use, even without a prior explanation. In the control condition, however, users could only reach high performance after they had practiced with the system. Moreover, participants preferred to use the annotated timelines for listening to historic, spoken word documents. In line with earlier research, e.g., [25,37] we therefore found that users are supported by the presentation of visual information on the spoken content.

Subtitling did not show a significant advantage or disadvantage in the recall of spoken content. We expect that the potential facilitation offered by subtitling may be evened out by the higher cognitive load it causes, which seems to be confirmed by the listeners' reactions. However, since we cannot rule out a null-effect given these results or a non-significant effect given the relatively small number of participants, more research is needed to confirm this expectation.

Since subtitling did not obviously hinder users we think that a textual content representation should be shown on-screen, as it at least helps users in other ways. First, for other tasks than recall high-quality transcripts have been shown to aid users in performing certain tasks [25,28,31]. Second, there is support for adding subtitling from the timelines task: the number of misperceptions found in that task – reflected by incorrect, but phonetically similar transcriptions – can be reduced by showing the text.

A remaining question is whether the choice for running subtitling was the optimal way of presenting textual transcripts. The fact that subtitling is refreshed and does not remain on screen, as a transcript would, for instance does not allow users to glance forward or backward over larger stretches of time. Whether our choice for a focused way of presenting transcripts should be reconsidered will be addressed in future research. Future work in multimedia retrieval should also address the question of how we can design UIs that support searchers with visual information on the spoken content while minimizing the user's cognitive load.

5. ADVANCED ACCESS TO SPEECH

In the previous sections we have presented the 'Radio Oranje' prototype search engine and an evaluation of its user interface. This prototype was a starting point for development of a framework for enhanced access to spoken word documents. Even though it adheres to our goal of online access to

fragments, and as such simplifies access to this particular collection, there is a lot of work to be done before such a framework can cope with the variety of collections found in audiovisual archives.

A challenge in designing user interfaces for access to spoken word archives is that the various collections kept in an archive may differ quite a lot from each other, resulting in a heterogeneous set of collections. As opposed to the (relative) homogeneity of most individual collections for which access applications have been developed and evaluated – such as the royal speeches at the 'Radio Oranje' website, but also meeting recordings and broadcast news (see section 2) – audiovisual archives contain the full range of possible recording types, i.e. from eloquent monologues to noisy amateur recordings. To support users in their exploration of this diversity we at least need a rich index that codes as much relevant information on a spoken word document as possible. The next step then is the development of user interfaces that present the information such that users can find their way around.

In the rest of this section we will explain how our framework can be used or adapted to realize access to other types of spoken word collections in the CH domain. We will first discuss automatic annotation to arrive at a rich index, and next we will go into the design of user interfaces.

5.1 Automatic Index Generation

Automatic generation of a detailed index into the spoken content can be achieved in several ways, depending on the amount of metadata or collateral data that is available for a collection. At one extreme of the metadata dimension we see systems that can exploit a full manual transcript, but at the other extreme are the systems that cannot use any manual annotation at all.

Aligning a transcript to the audio will generate an accurate index of the kind illustrated by the 'Radio Oranje' case (see also e.g., [2]). A significant proportion of spoken word collections from the cultural heritage domain come with transcripts, such as interview collections gathered by oral historians or social scientists, and (historical) speeches. Improved access to those collections could also be realized – relatively straightforwardly – by alignment of these transcripts to the audio for index generation, and by making that index searchable through an interface.

Where instead of full-text transcripts there are text sources available that follow the audio linearly, such as summaries or agendas, multi-pass alignment approaches may be used to generate annotations, e.g., [23]. This indexing scenario for instance applies to meeting recordings with corresponding minutes, to broadcasts for with production schemes are available, and to oral history collections with elaborate summaries.

If these options are not available, ASR can be employed for generating a textual index into the spoken content. Especially in heterogeneous collections, pre-processing of the audio is required. This includes speech/non-speech detection – to prevent the ASR engine from indexing music and non-speech – and speaker segmentation. Next, the ASR engine processes the segments labelled as speech, and the output is used to build an index.

An example of a heterogeneous spoken word collection in the CH domain is the TRECVID 2007/2008 collection. It is a subset of the so-called Academia collection from the Netherlands Institute for Sound and Vision, one of Europe's largest audiovisual archives. This collection consists of hundreds of hours of Dutch news magazines, science news, documentaries, educational programs and archival video. As for pre-processing performance on the TRECVID data, the Speech Activity Detection error rate on these data was 10.4% after optimization, and the transcript's Word Error Rate (WER) was 64.0% [14]. ASR performance is somewhat lower than performance reported on comparable English CH collections, see [1,9]. A WER in the range of 30-40% is deemed sufficient for use in information retrieval [6], but ASR performance on CH spoken word documents is generally lower.

The match between the words spoken and the topic or concept that is being talked about is only partial. As a result it may be difficult to estimate what a particular document is about given an index based on an aligned transcript or ASR output. Since manual addition of keywords is extremely (time-)costly, automatic extraction of high-level semantic information through e.g., topic detection, could be employed instead (e.g., NIST TDT evaluations 1998-2004)[1]. ASR transcripts can be the basis for several kinds of Natural Language Processing (NLP) that are suited for the capturing of the semantic layers in speech. This secondary analysis can be used as a basis for the generation of semantic annotation layers. These more atomic pieces of information might then be searched for with structured queries (e.g., using a relational database) or they might be used to generate structured summaries. This has been done for a variety of element types, including names (e.g., for persons, organizations and locations), times, and (at a higher aggregation level) events. NLP techniques can also be used to link transcribed text to semantic units or concepts from a domain-specific ontology or a thesaurus to support search and/or navigation.

In addition to indexing words and topics, other levels of information present in the recordings could be exploited for indexing purposes. As was mentioned earlier, pre-processing can reveal the presence of non-speech segments that can be classified according to their acoustic characteristics, such as music or noise (see [32] for an overview). The recording's bandwidth may provide information on whether it concerns a phone conversation or studio speech. Furthermore, there is more information in the speakers' voices than just the speech sounds that were uttered; they for instance also hold information on the emotional state of the speaker (see [34] for an overview), and prosody may be used for sentence and topic segmentation [27]. We know that these examples are far from exhaustive, but the message is that much more information can be extracted from spoken word documents than is captured in a literal transcript.

When automatically extracted information layers such as the ones introduced above are taken together, and are merged with manual descriptions when available, a multi-faceted index into the audio can be generated that has the potential to meet more information needs than current manual descriptions can.

5.2 Guiding the User

The question we discuss in this section is how we can adapt the user interface of our framework to begin scaling up its functionalities to the challenges coming with the kind of heterogeneous speech archives described above, as well as the richness and multiple perspectives that may come with more advanced tools for automated content analysis.

[1] NIST TDT, see: http://www.nist.gov/speech/tests/tdt/

5.2.1 *Searching and Browsing*

Our user logs from the 'Radio Oranje' study showed that users often did not formulate a query, but asked for a list of all speeches to begin their exploration of the collection. Whereas content listing is a feasible approach for a collection of 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 could provide users with a way of exploring an archive's contents. Index characteristics that could be used for document clustering are for instance year of production and creator, taken from the manual metadata, and topic or speaker profile, taken from the automatic index. Moreover, in addition to clustering audio files, manual annotations can also form the basis for collection access through browsing. For instance, speaker profiles and written summaries of the audio files' contents could be offered to users as a way of exploring a collection[2], and direct them to the spoken word documents of their interest.

As for the formulation of queries, users should be enabled to exploit the multiple annotation layers encoded in the index. Complex queries raise a number of important research questions, including how best to store and access multiple annotation layers, how best to combine evidence from multiple annotation layers for ranked retrieval, and how we can best support the development of appropriate mental models by the users of our systems. Though the concept of advanced search is not unknown to searchers, from a usability point of view the questions of how they will use it to access speech archives and which data characteristics are important to them in that situation are open issues.

5.2.2 *Result Selection and Playback*

In both the selection and the playback of results, visual and textual content representations may help users to locate and to navigate through relevant fragments of spoken word documents, which we also showed in our prototype's evaluation. Such representations help the user to build a mental model of the contents of an audio file, and we therefore expect them to play an important role when scaling up to larger archives.

The usefulness of content representations, however, depends of their accuracy [28,31]. For most CH collections automatically generated indexes will be based on ASR transcripts, and as yet their error rates are generally higher than those reported useful for user support [25,31]. If direct presentation of the transcripts may not support the user, an alternative way of presenting the content textually is by showing key words and key phrases extracted from low-quality ASR transcripts, as reported in [10]. In that study, however, the extraction technique relied on related text data for term selection, i.e. books and papers for recorded university courses. The question is whether external text sources or techniques from NLP, such as POS tagging, can also be deployed to extract alternative textual presentations from ASR output that can represent the spoken content for user support.

As for visual content representations, the multi-level indexes can be used to design more elaborate (timeline) visualizations that in addition to information on query term occurrence include information on various other annotation layers, such as intervals on certain topics, speaker turns, audio conditions etc. Some interfaces including multiple layers of information have been proposed (e.g., [30,36]), but extensive user testing, e.g., across multiple collections, has to our knowledge not yet been conducted.

As we learned from our user evaluation, the amount of information presented to users for building a mental model of the spoken word document's contents should be taken into account in future research. Possible solutions to this problem during result presentation and playback may lie in the dosage of information presented at one time. This could be realized by e.g., asking users to indicate their preferences by checking the characteristics of interest, or by including multiple tabs with different views on the data.

6. CONCLUSION AND FUTURE WORK

We have presented steps in the development of a user interface within a framework for access to speech archives with cultural heritage content. Through this framework, in which techniques from speech and language processing and interaction design are combined, spoken word collections from the cultural heritage domain can become more easily accessible for research and education in the arts and humanities, and for new creative productions. An evaluation of the functionalities of our prototype's user interface, the 'Radio Oranje' search engine, was presented as well as our ideas on how the framework can be developed further for use in heterogeneous speech archives.

Ongoing research within the CHoral project aims at improving the disclosure and access of Dutch spoken word collections from the cultural heritage domain. Successful retrieval starts with a good index into the spoken word documents. To improve the quality of automatic indexing acoustic and language models are being tuned for speech recognition purposes in this domain, and the question of how out-of-vocabulary terms can be handled is being addressed. Out-of-vocabulary (OOV) terms are words that are not in the vocabulary of the speech recognition engine, and therefore in principle can never become part of the index, which would make the fragment in which they occur irretrievable. Typically, OOV terms are named entities that form a significant proportion of query terms users would enter, which is why their retrieval is on our agenda.

To improve access functionality we are working on methods to visualize the contents of spoken word documents, both during the presentation of search results and during navigation within an audio document. Part of this work will be to reconsider our choice for a focused way of presenting high-quality transcripts by showing them as subtitling. Moreover, in the case of low-quality transcripts generated by ASR, ways of extracting the verbal essence instead of presenting full transcripts are being considered. Thirdly, since there often is manually generated metadata for heritage collections, varying from a few keyword assignments to full transcripts, ways of exploiting the combination of manually generated annotations and automatically generated representations for content presentation to users will be studied further. Finally, in order to enable scaling up to heterogeneous archives, research is needed into ways of presenting results in such a way that both similarities and differences between the results themselves, and between the results and the user's query can be quickly assessed by the user.

---

[2] An online example of full access to a spoken word archive through both ASR-based search and browsing of the audio and related documents can be found at http://www.buchenwald.nl (in Dutch). It discloses a collection of interviews with survivors of the concentration camp Buchenwald.

7. ACKNOWLEDGEMENTS

This paper is based on research funded by the CATCH programme (http://www.nwo.nl/catch) of the Netherlands Organisation for Scientific Research Research and the research programme MultimediaN (http://www.multimedian.nl/) supported by the Netherlands bsik-funding scheme.

8. REFERENCES

[1] Byrne, W., Doermann, D., Franz, M., Gustman, S., Hajic, J., Oard, D., Picheny, M., Psutka, J., Ramabhadran, B., Soergel, D., Ward, T., and Zhu, W-J. Automatic Recognition of Spontaneous Speech for Access to Multilingual Oral History Archives, *IEEE Transactions on Speech and Audio Processing*, 12(4), 2004, 420–435.

[2] Christel, M.G., Richardson J., and Wactlar, H.D. Facilitating access to large digital oral history archives through Informedia technologies, *Proceedings JCDL* '06, 2006, 194–195.

[3] Christel, M. and Warmack, A. The Effect of Text in Storyboards for Video Navigation, *Proceedings IEEE ICASSP*, Salt Lake City, UT, 2001.

[4] De Jong, F., Ordelman, R. and Huijbregts, M. Automated speech and audio analysis for semantic access to multimedia, *Proceedings SAMT 2006,* 226–240.

[5] De Jong, F.M.G., Westerveld, T. and de Vries, A.P. Multimedia search without visual analysis: the value of linguistic and contextual information. *IEEE Transactions on Circuits and Systems for Video Technology*, 17 (3), 2007, 365-371.

[6] Garofolo, J.S., Auzanne, C.G.P., and Voorhees, E.M. The TREC Spoken Document Retrieval Track: A success story, *Proceedings RIAO 2000.*

[7] Goldman, J., Renals, S., Bird, S., de Jong, F. M. G., Federico, M., Fleischhauer, C., Kornbluh, M., Lamel, L., Oard, D.W., Stewart, C., and Wright, R. Accessing the Spoken Word, *International Journal on Digital Libraries*, 5(4), 2005, 287—298.

[8] Gustman, S., Soergel, D., Oard, D., Byrne, W., Picheny, M. Ramabhadran, B., and Greenberg, D. Supporting Access to Large Digital Oral History Archives, *Proceedings of the Joint Conference on Digital Libraries*, 2002, 18-27.

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

[10] Haubold, A., and Kender, J.R. Analysis and visualization of index words from audio transcripts of instructional videos, *Proceedings ISMSE '04*, 2004.

[11] Heeren, W.F.L. User requirements for access to Dutch spoken audio archives, *CTIT Technical Report, University of Twente*, 2008.

[12] Hürst, W. User Interfaces for Speech-Based Retrieval of Lecture Recordings, *Proceedings ED-MEDIA*, 2004.

[13] Huijbregts, M.A.H., Ordelman, R.J.F., and de Jong, F.M.G. A Spoken Document Retrieval Application in the Oral History Domain, *Proceedings of SPECOM 2005*, Patras, Greece, 2005, 699-702.

[14] Huijbregts, M.A.H., Ordelman, R.J.F., and de Jong, F.M.G. Annotation of Heterogeneous Multimedia Content Using Automatic Speech Recognition, *Proceedings SAMT 2007* 2007 vol. 4816, *Lecture Notes in Computer Science*,78-90, Berlin, Springer Verlag.

[15] Jansen, B.J., Goodrum, A., and Spink, A. Searching for multimedia: analysis of audio, video and image Web queries, *World Wide Web*, 3, 2000, 249-254.

[16] Kalyuga, S, Chandler, P., and Sweller, J. Managing split-attention and redundancy in multimedia instruction, *Applied Cognitive Psychology*, 1999, 13, 351–371.

[17] Kim, J. Oard, D.W., and Soergel, D. Searching Large Collections of Recorded Speech: A Preliminary Study. *Proceedings of the Annual Conference of the American Society for Information Science and Technology*, 2003.

[18] Kimber, D.G., Wilcox, L.D., Chen, F.R., and Moran, T.P. Speaker segmentation for browsing recorded audio, *Proceedings CHI 1995*.

[19] Klemmer, S.R., Graham, J., Wolff, G.J., and Landay, J.A. Books with voices: paper transcripts as a tangible interface to oral histories, *Proceedings CHI* 2003, 89–96.

[20] Marsden, A., Nock, H., Mackenzie, A., Lindsay, A. Coleman, J., and Kochanski, G. ICT Tools for searching, annotation and analysis of audiovisual media, *AHRC ICT Strategy Project Report*, October 2006.

[21] Mayer, R.E., and Moreno, R. A split-attention effect in multimedia learning: evidence for dual processing systems in working memory, *Journal of Educational Psychology*, 90(2), 312-320.

[22] Mayer, R.E., and Moreno, R. Nine ways to reduce cognitive load in multimedia learning, *Educational Psychologist*, 38(1), 43-52.

[23] Morang, J. Ordelman,R., de Jong, F., and van Hessen, A. Infolink: analysis of Dutch broadcast news and cross-media browsing, *Proceedings IEEE ICME 2005*, 1582-1585.

[24] Moreno, P.J., Joerg, C., Van Thong, J.-M., Glickman, O. A Recursive Algorithm for the Forced Alignment of Very Long Audio Segments, *Proceedings ICSLP'98* 1998, Sydney, Australia.

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

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

[27] Pellom, B. SONIC: The University of Colorado Continuous Speech Recognizer, *Technical Report TR-CSLR-2001-01, University of Colorado*, 2001.

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

[29] Shriberg, E., Stolcke, A., Hakkani-Tür, D., and Tür, G., Prosody-based Automatic Segmentation of Speech into sentences and Topics, *Speech Communication*, 32, 2000, 127-154.

[30] Slaughter, L. Oard, D.W., Warnick, V.L., Harding, J.L., and Wilkerson, G.J. A Graphical Interface for Speech-Based Retrieval, *ACM Digital Libraries*, 305-306, 1998.

[31] Stark, L., Whittaker, S., and Hirschberg, J. ASR satisficing: the effects of ASR accuracy on speech retrieval, *Proceedings ICSLP 2000,* 1069–1072.

[32] Tranter, S.E., and Reynolds, D.A. An overview of automatic diarization systems, *IEEE Transactions on Audio, Speech and Language Processing*, 14,(5), 2006, 1557-1565.

[33] Van der Werff, L., Heeren, W., Ordelman, R, and De Jong, F. (2007). Radio Oranje: Enhanced access to a historical spoken word collection. In P. Dirix, I. Schuurman, V. VandeGhinste & F. Van Eynde (eds.): *CLIN 2006,* 207-218.

[34] Ververidis, D., and Kotropoulos, C. Emotional speech recognition: resources, features, and methods, *Speech Communication* 48, 2006, 1162-1181.

[35] Wellner, P., Flynn, M. and Guillemot, M. Browsing recorded meetings with Ferret. *Proceedings MLMI 2004*, 12–21.

[36] Whittaker, S. Choi, J., Hirschberg, J., and Nakatani, C. What you see is almost what you hear: Design principles for accessing speech archives, In *Proceedings ICSLP98*.

[37] Whittaker, S., Hirschberg, J., Choi, J., Hindle, D., Pereira, F.C.N., and Singhal, A. SCAN: Designing and Evaluating User Interfaces to Support Retrieval From Speech Archives, *Proceedings ACM SIGIR99*, 26–33.