Towards Affordable Disclosure of Spoken Heritage Archives

Roeland Ordelman (1) Willemijn Heeren (1) Franciska de Jong (1)
Marijn Huijbregts (2) Djoerd Hiemstra (3)*

December 10, 2009

Abstract

This paper presents and discusses ongoing work aiming at affordable disclosure of real-world spoken heritage archives in general, and in particular of a collection of recorded interviews with Dutch survivors of World War II concentration camp Buchenwald. Given such collections, we at least want to provide search at different levels and a flexible way of presenting results. Strategies for automatic annotation based on speech recognition – supporting e.g., within-document search– are outlined and discussed with respect to the Buchenwald interview collection. In addition, usability aspects of the spoken word search are discussed on the basis of our experiences with the online Buchenwald web portal. It is concluded that, although user feedback is generally fairly positive, automatic annotation performance is not yet satisfactory, and requires additional research.

1 Introduction

Given the quantity of digital spoken word collections, which is growing every day, the traditional manual annotation of collections puts heavy demands on resources. For some content owners, to apply even the most basic form of archiving is hardly feasible. Others need to make selective use of their annotation capacity. In order to enable content-based access to and exploitability of large amounts of potentially rich content the automation of the semantic annotation task is necessary.

There is common agreement that automatic annotation of audiovisual archives based on the automatic transcription of the spoken words therein may boost the accessibility of these archives enormously ([Byrne et al., 2004], [Goldman et al., 2005], [Garofolo et al., 2000]). However, success stories of the application of speech-based annotation for real-world archives lag behind. In our view, this is due in particular to (i) the (expected) low *accuracy* of automatic, speech-based, metadata generation, (ii) the uncertainty about how existing technology fits in given collection characteristics on the one hand, and often still quite unclear user needs on the other hand -roughly referred to as *usability*-, and (iii) uncertainty about the *affordability* of integrating the technology in existing workflows.

In the laboratory the focus is usually on data that (i) have well-known characteristics, (ii) form a relatively homogeneous collection, and (iii) are annotated in quantities that are sufficient for speech recognition training and evaluation purposes. Such data sets have often been created for (benchmark) evaluation purposes such as the NIST Rich Transcription series[1] that for a number of years in succession provided example data and ran evaluations on broadcast news data and meeting data. For researchers that are particularly interested in evaluating the performance of a low level analysis component in isolation, such as speech recognition, benchmarks or evaluation fora are more or less a sine qua non. In the archival practice however, the exact features of data are often unknown and the data may be far more heterogeneous in nature than those usually seen in the laboratory.

*(1) Human Media Interaction, University of Twente, Enschede, The Netherlands, (2) Centre for Language and Speech Technology, Radboud University, Nijmegen, The Netherlands, (3) Database Group, University of Twente, Enschede, The Netherlands

[1] http://www.nist.gov/speech/tests/rt

Annotated sample data resembling the audio conditions in these archives is typically not available via the usual channels[2], especially for less common languages, and in addition there is little match between the available sample data, such as the Spoken Dutch Corpus for The Netherlands ([Oostdijk, 2000]), and archival data. Because of their heterogeneity and lack of example data, we often refer to real-world data as '*surprise*' data. As a result, the accuracy of automatic transcription may often be on the low side so that the use of speech-based annotations for indexing needs to be carefully evaluated.

Even when speech recognition can provide sufficiently accurate annotations, the question is how these could serve the needs of potential users of a collection. Moreover, because human interpretation is lacking from automatically generated annotations, certain abstractions cannot be made easily. For instance, it will be easier to retrieve relevant documents that were automatically annotated when users look for factual content, e.g., topics or events, than when users look for 'appealing' content, e.g., reflecting some type of affect or atmosphere. A similar problem is the fact that a mismatch may be expected between the actual words that are being spoken and the more abstract semantic concepts that are being talked about. In addition, the value of an automatically annotated collection may become especially apparent when placed in context: cross-connected with related multimedia data, such as other collection items, collateral text data or even items from other collections. Attaching web-style keyword search to a speech-based index may not always be the best solution, especially because of the unstructured nature of audiovisual documents.

Provided that content owners are confident enough about accuracy and usability of automatic annotation tools, the affordability or feasibility of the implementation of such tools in a real-world application still remains an unknown quantity. Proofs-of-concept, typically developed from laboratory studies on relatively small and often homogeneous data sets, tend to disregard non-functional requirements of a tool operating in a real-life environment, such as robustness, reliability, scalability, interoperability, resource consumption, and failure management. Awareness of these requirements for deploying automatic annotation tools, and insight in the processes that are involved is crucial to be able to define strategies for the implementation of such tools.

Presently available ASR techniques for example, require the investment of effort in several kinds of pre-processing, such as speech/non-speech detection, speaker segmentation, and language detection to be able to select the proper AV parts for processing and guarantee robustness on the input level. Errors related to attempts to do Dutch speech recognition on music or Spanish speech will often only show up when it is too late: at search time. When pre-processing cannot be (fully) automated, additional manual labour might be calculated in, depending on the characteristics of the collection.

Another example relates to the level of accuracy that is required. In order to adapt an ASR system to the characteristics of a collection, the manual transcription of substantial quantities of representative speech data is prefered to improve the quality of the speech transcripts. For the automatic transcription in the MALACH project for example, a large corpus (65-84 hours reported in [Byrne et al., 2004]) was created for training the ASR system. This approach is not scalable however and hardly feasible when the aim is to automatically annotate a multitude of collections, either within a large archive or coming from various (cultural heritage) institutes. Manual transcription work is estimated to take between 8-15 times realtime depending on skill of the annotator (expensive professionals versus cheap non-professionals) and the required accuracy level of the annotation. For certain domains, exploiting available metadata and textual resources is critical to ensure that domain specific content words (jargon, named entities) are included in the recognition vocabulary. For non-static, regularly growing collections (e.g., news or regularly recorded meetings) it may even be required to automate this vocabulary update process to allow system adaptation to the dynamics of the content (e.g., changing topics).

Finally, next to the effort that is required for the set-up of the annotation system itself, additional actions are needed to connect the system to an archival work-flow (interoperability). Operating a speech recognition system requires at least the digitization of the data, selection of collections that are appropriate for automatic annotation, a functionality to provide audio data (and metadata) to the system, a suitable metadata model for incorporating time-labeled transcripts, a search facility for searching these, and finally a user interface that is able to exploit the benefits for having time-labeled transcripts available (usability).

[2]e.g., The Linguistic Data Consortium (LDC) or the Evaluations and Language resources Distribution Agency (ELDA)

In this paper we present and discuss ongoing work focussing on how affordable disclosure of spoken word archives can be put into practice. We study a real-world use case for a Dutch cultural heritage institute: the development of a multimedia web-portal for accessing recorded interviews with Dutch survivors of World War II concentration camp Buchenwald and related text data. Here, the conditions are the same as for many spoken-word archives: on the one hand there are audiovisual data (interviews), some descriptive metadata and a potentially large set of related, secondary information sources, and on the other hand some fine, freely available open-source tools for content source analysis such as ASR, indexing and search. The question is how to maximize the potential of the collection while minimizing the development costs. Features we would like to be able to provide are search at different levels (entire document, within a document and cross-media) and a flexible way of presentation of results.

The 'surprise data' problem for speech recognition constitutes a major obstacle towards forms of access that require time-labeled annotations. In this paper, we discuss ongoing work on the two strategies we have adopted to deal with the surprise data problem affordably: (i) developing a robust speech recognition system that can be deployed in unknown domains without the need for expensive manual annotation work for system training and without extensive manual tuning, and (ii) making smart use of available resources, such as descriptive metadata or collateral data, to provide useful annotations based on the speech in collections.

Section 2 describes the data collections we are focusing on and provides a global description of the retrieval framework that is used 2.1. Next, we zoom in on generating time-labeled access points in section 3. In section 4 usability issues will be discussed and related to the Buchenwald application in general and its web logs in specific. Section 5 finally discusses the current status of our work and future work.

2 From spoken-word archive to multimedia information portal

The 'Buchenwald' project is the successor of the 'Radio Oranje' project[3] that aimed at the transformation of a set of World War II related mono-media documents – audio, images, and text – into an on-line multimedia presentation with keyword search functionality. The mono-media documents consisted of the audio of speeches of the Dutch Queen Wilhelmina, the original textual transcripts of the speeches, and a tagged database of WWII related photographs. The transcripts were synchronized on the word-level (aligned) to the audio using speech recognition technology to be able to (i) generate a time-labeled index for searching and immediate play-back of relevant audio segments, (ii) show the spoken words in the speeches as 'subtitling' during audio playback, and (iii) show sliding images related to the content of the speech by matching the index words with the image tags from the database, as shown in Figure 1. Since its launch in the beginning of 2007 it has been visited over 1500 times.

The 'Radio Oranje' project is an outstanding example of how the exploitability of a spoken word archive and the experience of a user interacting with the content, can be boosted with minimal means, deploying already available textual resources in combination with well-known and robust alignment technology ([Heeren et al., 2007, Ordelman et al., 2006]) that has a fairly low complexity.

The 'Buchenwald' project extends the 'Radio Oranje' approach. Its goal is to develop a Dutch educational multimedia information portal on World War II concentration camp Buchenwald[4] giving its user a complete picture of the camp then and now by presenting written articles, photos and the interview collection. The portal holds both textual information sources and a video collection of testimonies from 38 Dutch camp survivors with durations of between one half and two-and-a-half hours. For each interview, an elaborate description, a speaker profile and a short summary are available. In the long term the data collection could be extended with content available from the content provider, the *Netherlands Institute for War Documentation* (NIOD), such as photos from the Picture Bank World War II[5] and maps and floor plans of the camp, or even related data available from the World Wide Web.

[3] Radio Oranje: `http://hmi.ewi.utwente.nl/choral/radiooranje.html`

[4] Buchenwald: `http://www.buchenwald.nl`

[5] 'Beeldbank WO2': `http://www.beeldbankwo2.nl/`

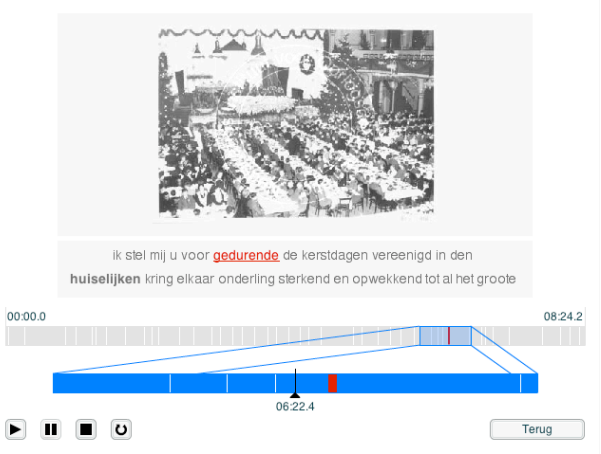


Figure 1: Screen shot of the 'Radio Oranje' application showing bars visualizing the audio (entire speech and current segment), subtitles (keyword in bold-face, current word underlined) and an image that relates to the topic of the speech (in this case: Christmas).

In addition to the traditional way of supporting access via search in descriptive metadata at the level of an entire interview, automatic analysis of the spoken content using speech and language technology also provides access to the video collection at the level of words and fragments. Having such an application running in the public domain allows us to investigate other aspects of user behavior next to those investigated in controlled laboratory experiments.

2.1 Spoken Heritage Access & Retrieval system description

The Buchenwald portal is implemented through the Twente Spoken Heritage Access & Retrieval framework consisting of flexible open-source components, developed with the aim to easy integration in real-life work-flows: (i) an analysis component that currently contains a module for audiovisual format conversion and demultiplexing, and the SHoUT speech processing toolkit, (ii) the PF/Tijah retrieval system, and (iii) an audiovisual search front-end.

The SHoUT speech processing toolkit consists of a set of applications for automatic speech recognition, speaker diarization and speech activity detection. SHoUT is a Dutch acronym[6] and it is available under the GNU General Public License[7]. For automatic speech recognition, a phone-based token-passing Viterbi decoder is used. The phones are modeled with three-state left-to-right Hidden Markov Models (HMM) with Gaussian mixture models as probability density functions. For speaker diarization, an agglomerative clustering method is used. Clusters are merged based on the Bayesian Information Criterion. Similar to the ASR module, the speech activity detection module is based on Viterbi search. SHoUT uses statistical methods that typically require statistical models that are created using example data. It is important that the conditions of the audio that is used for creating these models match the conditions of the audio that needs to be processed, as any mismatch will reduce the accuracy of recognition. When the conditions of the to-be-processed audio are unknown, the models and system parameters cannot be tuned to the domain. Therefore, the algorithms used in the SHoUT toolkit are designed to have as few parameters that need tuning as possible, so that the system is relatively insensitive to this mismatch problem (see also [Huijbregts, 2008]).

Tijah is a text search system integrated with the Pathfinder (PF) XQuery compiler ([Hiemstra et al., 2006]). It can be downloaded as part of MonetDB/XQuery, a general purpose XML database management system[8].

[6] 'Spraak Herkennings Onderzoek Universiteit Twente' ('Speech recognition research University of Twente')
[7] `http://sourceforge.net/projects/shout-toolkit/`
[8] http://dbappl.cs.utwente.nl/pftijah

PF/Tijah includes out-of-the-box solutions for common tasks such as index creation, document management, stemming, and result ranking (supporting several retrieval models), but it remains open to any adaptation or extension. For the Buchenwald interviews, PF/Tijah was used as a general purpose tool for developing end-user information retrieval applications, using XQuery statements with text search extensions. The transcripts generated by automatic speech recognition are stored directly as (verbose) MPEG-7 files in PF/Tijah. Since PF/Tijah is an XML database system, a simple *add document* command suffices. Similarly, the metadata is added as XML. The PF/Tijah system has a number of unique selling points that distinguish it from other information retrieval systems. Firstly, PF/Tijah supports retrieving arbitrary parts of the XML data, unlike traditional information retrieval systems for which the notion of a document needs to be defined up front by the application developer. So, PF/Tijah supports finding a complete interview that contains the query words, but it also supports searching for MPEG-7 *AudioSegment* elements, to retrieve the exact point in the interview where the query words are mentioned. Secondly, PF/Tijah supports text search combined with traditional database querying, including for instance joins on values. Metadata fields such as *name* (of the person being interviewed) and *date* are easily combined by a join on *interview identifiers*. Thirdly, PF/Tijah supports ad hoc result presentation by means of its query language. For instance, after finding the matching segment and the video's metadata, we may choose to show the interviewee's name, the date of the interview, and the duration of the interview. This is done by means of XQuery element construction. All of the above can be done in a single XQuery statement using a few lines of code.

The audiovisual search front-end shown in Figure 1 was developed as part of the CHoral (access to oral history) project, which is funded by the CATCH program of the Netherlands Organization for Scientific Research. This front-end was re-used and modified in the Buchenwald system. Some of its components have been used recently for other interview collections[9]. In the next section, we zoom in on the automatic speech processing component within the framework.

3 Automatic annotation using automatic speech recognition

Strategies for automatic annotation based on speech that in principle can be pursued may differ depending on the data that are available with spoken word collections, such as audio, some form of metadata, a thesaurus and collateral data. These data can be used as a starting point for providing document level, within-document level and cross-media search. Three main speech annotation strategies have been identified: (i) synchronization of metadata with the interview audio (alignment), (ii) linking terms from a thesaurus or metadata directly to the audio (spoken term detection), and (iii) full-scale large vocabulary speech recognition.

The question is which strategy suits best given on the one hand, the characteristics of the data and the *accuracy* levels that can be obtained and, on the other, the consequences with respect to *affordability* of pursuing such a strategy. Below the three options are surveyed in view of the Buchenwald case that besides audio has descriptive metadata, a carefully created list of thesaurus terms from the content provider, and collateral data. Note that in this paper we will use the term 'collateral data' to refer to data that is somehow related to the primary content objects, but that is not regarded as metadata.

3.1 Alignment

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. Alignment is a well-known procedure used frequently in ASR, for example when training acoustic models (see e.g., [Young et al., 2000]). It applies best when available transcripts closely follow the speech as it was found in the data, such as can be the case with accurate minutes from a meeting, although it holds for lower text-speech correlation levels as well. When the available data allows for the successful application of the alignment strategy, alignment has a number of benefits: it saves the development and tuning of collection-specific automatic

[9]Rotterdam Municipal Archives: `http://www.brandgrens.nl`

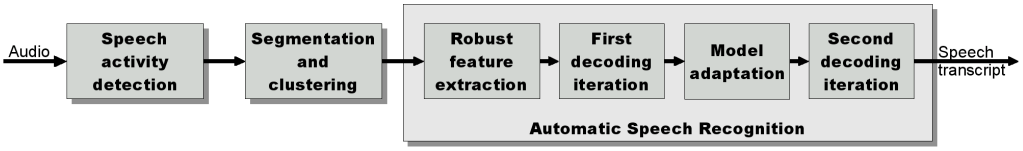


Figure 2: Overview of the decoding system. Each step provides input for the following step. There are three subsystems: segmentation, diarization and ASR.

speech recognition, it is accurate (e.g., with respect to collection-specific words that are especially important for search), resource efficient and fast. The 'Radio Oranje' project mentioned in section 2 made successfully use of the alignment strategy.

In the Buchenwald project the metadata consists of an outline of the interview, mentioning the highlights in on average 2500 words per interview. The correlation with the actual speech is however too low for the application of the alignment strategy to be successful. Often, lower speech-text correlations can be overcome by 'anchoring' the collateral data globally to the audiovisual data. This is done by an alignment on the text level, using a speech recognition transcript generated by a non-optimized (out-of-the-box) decoder. Parts that can be aligned (overlap) can be used as anchors, or segmentation points, that subsequently can be used as input for a standard acoustic alignment procedure that focusses on the unaligned parts (see also [De Jong et al., 2006]). In the case of the Buchenwald application, the outlines of the interview are not verbatim and do not follow the storyline closely (as interviewees often go astray). Added to this that speech recognition accuracy was not very high (as will be discussed below), the anchoring approach did not work.

We therefore concluded that the alignment of transcripts with a very low text-speech correlation requires an extension of the approaches used thus far and/or speech recognition transcripts with a higher accuracy level than we had available. As an alternative, we investigated an approach that uses the words from the collateral data to find relevant access points to the collections by deploying word spotting or spoken term detection. In section 3.2.1 below we will report on preliminary attempts in this direction. As we used a large vocabulary speech recognition set-up for the spoken term detection task we first discuss the large vocabulary speech recognition approach.

3.2 Large vocabulary speech recognition

When the alignment strategy is not an option or cannot provide useful synchronizations, mobilizing a full-blown speech-to-text system becomes inevitable when the aim is to enable within-document search. Available metadata and collateral data should then be used as a source for domain tuning (minimizing out-of-vocabulary rate) or even as a strong bias during recognition ('informed speech recognition').

Figure 2 is a graphical representation of the three-component SHoUT speech recognition system workflow that starts with speech activity detection (SAD) in order to filter out the audio parts that do not contain speech. The SAD subsystem filters out all kinds of sounds such as music, sound effects or background noise with high volume (traffic, cheering audience, etc.) in order to avoid processing audible non-speech portions of the data that would introduce noise in the transcripts due to assigning word labels to non-speech fragments.

After SAD, the speech fragments are segmented and clustered (diarization: "who speaks when?", see also [Tranter & Reynolds, 2006] for an overview). In this step, the speech fragments are split into segments that contain only speech from one single speaker with constant audio conditions. Speech from a single speaker under different audio conditions will be separated. Each segment is labeled with a speaker ID.

The ASR subsystem flow consists of feature extraction, a first decoding iteration, a model adaptation step and a second decoding iteration.

3.2.1 Speech processing for surprise data

The results of the publicly accessible Dutch broadcast news (BN) spoken document retrieval system[10] developed at the University of Twente ([Ordelman, 2003]) indicate that the quality of the speech recognition transcripts is high enough for retrieval purposes. A formal evaluation of Dutch speech recognition in the BN domain was done recently in the context of the Dutch N-Best benchmark evaluation[11] [Kessens & van Leeuwen, 2007]. Depending on the exact characteristics of the data, word error rates (WER) fluctuate between $25 - 35\%$ (N-Best also incorporated interviews and discussions in the evaluation set) ([Huijbregts et al., 2009]).

For the broadcast news domain, large amounts of newspaper data and audio data are available for training the language models and acoustic models respectively. However, a system trained using this data (and therefore referred to as broadcast news system) performs rather poorly outside the broadcast news domain. Using different broadcast news systems for indexing a multimedia archive with interviews of the Dutch novelist Willem Frederik Hermans in various audio qualities did not yield very accurate speech transcripts (WERs between $65-80\%$). More recently, we also found a substantial gap between automatic speech recognition performance on broadcast news data and performance on the (very heterogeneous) Academia collection from The Netherlands Institute for Sound and Vision that is currently used in the TRECVID 2007/2008 evaluations and consists of hundreds of hours of Dutch news magazines, science news, documentaries, educational programs and archival video. We found that this gap was to a large extent due to the mismatch between training and testing conditions ([Huijbregts et al., 2007a]).

For the SAD and diarization components we experienced the same kind of problem. Generally state-of-the-art SAD and diarization systems perform very well in controlled audio conditions. However when the training data do not match the evaluation data, performance tends to drop significantly. So, the system needs to be re-trained when the task domain changes. Typically it is difficult to find an annotated data set for training that matches the task data conditions. In addition for the SAD system, it is not known beforehand which models are needed to filter out unknown audio fragments such as music or sound effects and to collect training data for those models.

The approach taken in this paper for processing audio where conditions potentially do not match the training data conditions, is to reduce the need for training data and to employ techniques that make the system less sensitive to this mismatch. For SAD and speaker diarization, methods were developed that do not need any training or tuning data at all. For the ASR component a number of well known techniques were implemented for robust speech recognition.

3.2.2 System description

The speech activity detection (SAD) component retrieves all segments in the audio recording that contain speech. All other sound fragments need to be discarded. As mentioned before, for surprise data it is not possible to create accurate statistical non-speech models prior to decoding. Instead, a two pass SAD approach is taken ([Huijbregts et al., 2007b]). First, an HMM-based broadcast news SAD component is employed to obtain a bootstrapping segmentation. This component is only trained on silence and speech, but because energy is not used as a feature and most audible non-speech will fit the more general silence model better than the speech model, most non-speech will be classified as silence. After the initial segmentation the data classified as silence is used to train a new silence model and a model for audible non-speech, the 'sound' model.

The silence model is trained on data with low energy levels and the sound model on data with high energy levels and both models are trained solely on data of the recording that is being processed. After a number of training iterations, the speech model is also re-trained using solely the recording. The result is an HMM-based SAD subsystem with three models (speech, audible non-speech and silence) that are trained solely on the data under evaluation, solving the problem of mismatching training and evaluation conditions.

The SHoUT *speaker diarization* system is inspired by the system described in [Anguera et al., 2007] that was developed to have no tunable parameters or models that are created using a training set. The system uses an agglomerative clustering algorithm in which the speaker clusters are each represented by a GMM. Initially, too

[10] Dutch Broadcast News SDR system: http://hmi.ewi.utwente.nl/showcases/broadcast-news-demo

[11] N-Best: Northern and Southern Dutch Benchmark Evaluation of Speech recognition Technology

many HMM states are created. The number of states is then iteratively decreased and the GMMs are slowly trained on speech from a single speaker until the correct number of GMMs is reached. For more information on the exact algorithm, see [Anguera et al., 2007, Huijbregts, 2008]. The aim is not to use tunable parameters or models and the high performance of the system make this a very good approach for clustering data with unknown audio conditions.

The ASR subsystem consists of four steps. First, feature extraction is performed in a way that normalizes the features for speaker and audio variations as much as possible. The results of the first decoding pass are used to adapt the acoustic model for each cluster. These cluster dependent models are then used in the final decoding iteration.

For feature extraction, two existing techniques were chosen that aim to normalize the features as much as possible to variations in the audio due to speaker and audio characteristics. A first, simple but effective, technique that is applied is Cepstrum Mean Normalization (CMN). Vocal Tract Length Normalization (VTLN) is used to normalize variations in vocal tract length of the various speakers in both the training set as the evaluation set.

The ASR decoder applies a time synchronous Viterbi search in order to retrieve its hypothesis. The Viterbi search is implemented using the token passing paradigm. HMMs with three states and GMMs for its probability density functions are used to calculate acoustical likelihoods of context dependent phones. Up to 4-gram back-off language models (LMs) are used to calculate the priors. The HMMs are organized in a single Pronunciation Prefix Tree (PPT) and instead of copying PPTs for each possible linguistic state (the LM N-gram history), each token contains a pointer to its LM history.

The clustering information obtained during segmentation and clustering is used to create speaker dependent acoustic models. The SMAPLR adaptation method was chosen to adapt the means of the acoustic model Gaussians. This method was chosen because it requires hardly any tuning and it automatically determines to what extent the models can be adapted according to how much adaptation data is available. This procedure prevents the models from being over-fitted on the adaptation data when only small amounts of adaptation data are available while it adapts the model parameters as much as possible.

3.2.3 Evaluation

For the evaluation of speech recognition performance for the Buchenwald interviews, multiple speech segments from four interviews with a total duration of two hours (14810 words) were selected and annotated on the word level. The set consists of four aged, male interviewees, one female interviewer and a male narrator who appears in the beginning of every interview and introduces the interviewee. The interviews were processed by the speech processing system in one piece (SAD, diarization and ASR), although only parts of the interviews were used for evaluation, in order to be able to measure true system performance in which SAD and diarization errors are reflected. The different evaluation conditions described below all use the SAD and Diarization configuration as described above. Only speech recognition configuration parameters were altered.

The following ASR configurations were used:

1. Broadcast news configuration (SHoUTBN2008)
   This system corresponds to the system used in the N-Best benchmark evaluation described in section 3.2.1. It has acoustic models trained on 75 hours of speech from the Spoken Dutch Corpus (CGN, [Oostdijk, 2000]) and language models trained on the 1999-2004 newspaper corpus (485 Mw) of the Twente News Corpus ([Ordelman et al., 2008a]) plus speech transcripts available from the CGN corpus. The system has a $65K$ pronunciation dictionary that was manually checked.

2. Vocabulary adaptation system (SHoUTBW2008a)
   For this system a vocabulary and language model were created specifically for the domain to incorporate frequent words from the domain such as “Buchenwald” and “SS-er” (*SS person (Schutzstaffel)*) using descriptions available with the interviews and web data on Buchenwald and related topics. The language model was created by interpolating TwNC newspaper data, CGN transcripts, Wikipedia texts, and the collected text data on Buchenwald. Interpolation parameters were estimated using a small interview reference

of 9000 words that was not used for ASR evaluation. Roughly 25% of the $67,8K$ words in the pronunciation dictionary were generated automatically by a grapheme-to-phoneme converter ([Ordelman, 2003]) and checked manually.

3. Interview-specific language models (SHoUTBW2008b)
   In this configuration, interview-specific language models were created by interpolating the Buchenwald LMs from BW2008a with bigram language models that were created individually for every interview based on the descriptions in the metadata (resembling informed speech recognition using a strong bias).

4. Acoustic adaptation using 'enrollment'(SHoUTBW2008c)
   To evaluate performance gain from speaker-specific acoustic models based on manual intervention, either in a so called enrollment session (typically having interviewers read from prepared text) or by annotating a small part of the interview, we annotated parts (on average 2 minutes) of the speech of the main speakers and used this for acoustic model adaptation. The adapted models were then used for the recognition of an entire interview containing this speaker.

Table 1 shows the substitutions, insertions, deletions, word error rates (WERs), and out-of-vocabulary rates on the Buchenwald evaluation data for the different system configurations. The labels behind the system-IDs refer to (1) the first recognition pass without automatic acoustic adaptation, and (2) the second pass with acoustic adaptation. Note that especially when the test set is relatively small, it is possible that a decrease in WER is simply due to chance. In [Gillick & Cox, 1989], two significance tests are proposed for ASR: the McNemar's test and the matched-pairs test. It is stated that especially the matched-pairs test is suitable for significance testing of connected speech. With this test, it can be calculated what the probability *p* is that two hypothesis are the same. If *p* is very small (typically 0.05, 0.01 or even 0.001), the two hypothesis are considered significantly different. An application that calculates *p* for two hypothesis is described in [Pallet et al., 1990]. This application will be used in the remainder of this work for performing significance tests.

We see that word error rates significantly improve ($p < 0.001$,) after acoustic adaptation in the second pass (73.7% to 71.7%, and 72.0% to 69.1%), by bringing down the out-of-vocabulary rate and using domain specific language models (71.7% to 69.1%), by using topic specific LMs (69.1% to 67.9%) and by using manual AM adaptation (69.1% to 67.3%). Note that WERs may exceed 100% as WER is defined as the number of insertions, deletions and substitutions, divided by the number of words in the reference transcript.

Table 2 provides the error rates and number of words for the interviewees only for the broadcast news run, the BW2008a-2 run (LM adaptation) and the BW2008c-2 (manual AM adaptation). An upward trend in performance can be observed.

Configuration	main feature	%WER	%Sub	%Del	%Ins	%OOV
BN2008-1	BN models	73.7	50.4	16.0	7.3	2.65
BN2008-2	BN models + AM adapt	71.7	48.9	13.0	9.7	2.65
BW2008a-1	LM adapt	72.0	48.9	15.6	7.6	1.77
BW2008a-2	LM adapt + AM adapt	69.1	46.8	14.1	8.2	1.77
BW2008b-2	int. specific LM + AM adapt	67.9	45.6	14.8	7.5	na
BW2008c-2	LM adapt + manual AM adapt	67.3	45.6	14.1	7.6	1.77

Table 1: Speech recognition results of the different system configurations

The speech recognition evaluations show that the transcription quality for this collection is very low. There are difficulties on the acoustic level and the language model level that might explain the results. On the acoustic level, the sometimes mumbled speech of the aged men does not match the speech models used well; these were not trained specifically for speech from the elderly. This has clearly different spectral and pronunciation patterns as a result of degradation of the internal control loops of the articulatory system and changes in the size and periodicity of the glottal pulses. As a result speech recognition performance for seniors is known to degrade

Speaker	#wrds	BN2008-2	BW2008a-2	BW2008c-2
int06-male	2524	77.0	75.3	73.1
int07-male	4878	89.7	88.6	84.4
int08-male	3240	115.0	109.0	104.8
int09-male	2701	49.9	46.6	45.3

Table 2: Speech recognition results of the main three system configurations for each of the interviewees in the set.

Termset	#wrds	occ. in ref	correct	false alarm	miss	ATWV
thesaurus	1394	136	43	80	93	0.3202
queries	275	236	67	55	169	0.2417
metadata	619	1117	322	337	795	0.2677

Table 3: For each term list, the number of words that occur in the reference transcripts, the number of correctly recognized words, false alarms, misses, and ATWV.

considerably (see e.g., [Anderson et al., 1999]). Moreover, some speakers (e.g., from interview-7 and interview-8) have a strong accent resulting in pronunciations that deviate a lot from the canonical pronunciations in the recognizer's dictionary. Another acoustic problem was observed in interview-8: as it was recorded in a garden, it has twittering of birds in the background that could have had a dramatic influence on ASR performance. On the level of the language model, there is a clear mismatch between the models trained using mostly newspaper text data and the spontaneous nature of the speech with hesitations and grammatical errors from the interviews.

Although efforts to improve baseline system performance by adapting it to the task domain with widely used approaches –LM adaptation using available metadata and collateral data, automatic AM adaptation– and a minimum amount of manual labor (manual AM adaptation) yielded some performance gain, the transcript accuracy of the best performing system configuration is still very poor. It is questionable whether it is useful as an indexing source. To get a better picture of this without an extensive user evaluation, the usability of the speech recognition transcripts for a search task were evaluated by looking at the speech recognition performance from a spoken term detection (STD) perspective. In 2006 NIST initiated a Spoken Term Detection evaluation benchmark where the task was to find all of the occurrences of a specified 'term' in a given corpus of speech data. Typically, detection performance is measured via standard detection error trade-off (DET) curves of miss probability versus false alarm probability. In the STD evaluation the overall system detection performance is measured in terms of 'actual term-weighted value' (ATWV) which is computed on the basis of the miss and false alarm probabilities (see [NIST, 2006]). For the Buchenwald collection we used the STD paradigm to compute STD performance using three different 'term lists' representing user queries: (i) the thesaurus from the content provider dedicated to 'war documentation' with 1394 terms, (ii) a list of 275 single word terms extracted from the query logs, filtered using a stoplist of 1500 words (see section 4.1 below), and (iii) all words with a frequency of 10 and above from the metadata, filtered using the same stoplist. In Table 3 the number of words in the reference transcripts, the number of correctly recognized words, false alarms, misses, and ATWV for each term list are shown. It is clear that the number of false alarms and misses are off balance with the number of hits resulting in low actual term-weighted values.

4 The Buchenwald user interface

The user interface (UI) developed for access to this interview collection allows users to search and browse the videos and the corresponding texts. Search in the texts (elaborate descriptions, speaker profiles, short summaries) is comparable to the traditional way of accessing interview collections, and audiovisual archives in general. In such archives, however, the lack of a direct link between the text and the corresponding fragment of audio or video makes search relatively slow and effortful. The addition of a time-stamped word-level index generated through

ASR enables search *within* the videos and supports direct retrieval of video fragments.

Figure 3: Screen shot of the 'Buchenwald' application showing the search field and a list of results given a 'provide overview' request.

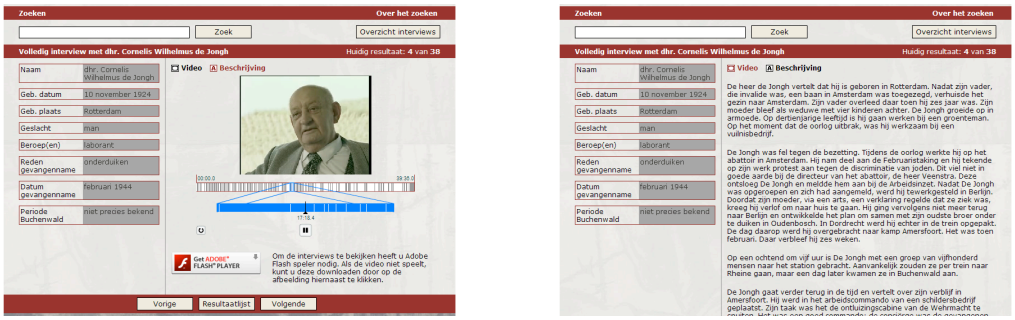
Users can type a query in the search field; search in specific data types or metadata fields is not (yet) supported. The retrieval engine matches the query against the ASR-based index and the different types of texts. Search results are listed below the search field (see Fig. 3) and contain context information (interview duration, location, and date), links to content information (speaker profile, short summary), and a link to the video file and the elaborate description. The speaker profile and summary are shown as extensions of the result list when the user requests this information. The video and the description are shown on a separate page, see Fig. 4(a) and Fig. 4(b). When a match with the query is found in the speech track of the video, the result list indicates that a fragment was retrieved. When the match was found in the texts, it indicates that an entire interview was retrieved. In both cases, query terms are highlighted so that the user can see why the result was presented.

From the user logs of the 'Radio Oranje' system we learned that many users want to browse the collection without a specific question ([Heeren & de Jong, 2008]). We had created the opportunity to do this by presenting a button that generates a list of all radio speeches. Since users started a session by clicking that button about 2/3 of the time over six months of use, we made a similar function for exploring the Buchenwald collection. This 'Show all' button is found to the right of the general search field (see Fig. 3).

A user selects a result by clicking the link to the video. In the case that a full interview was retrieved (text match) the video starts playing at the beginning of the interview, in the case that a video fragment was retrieved (ASR match) the video starts playing at the beginning of the first fragment in which a match occurs. The user can navigate through the video by clicking the timeline visualization consisting of a file overview (upper bar) and a zoomed-in view of 30 sec around the cursor (lower bar). The file overview shows the distribution of query terms over the interview. In both bars, the locations of segment boundaries and query terms are shown. The user also has a play/pause button available, and a button that restarts the video at the position where it first began playing, i.e., at the beginning of either the interview or the fragment.

4.1 Evaluation of the Buchenwald UI

To evaluate the use and usability of the Buchenwald UI two types of analyses were run. First, the use of the Buchenwald UI was analyzed through its user logs. Second, a heuristic evaluation of the website was done.



(a) Video playback (b) Elaborate description

Figure 4: Screen shot of the 'Buchenwald' application showing (a) the playback environment with speaker details and video, controlled by both playback buttons and an interactive timeline showing segment boundaries and query locations and (b) the elaborate description of the same interview.

4.1.1 Log analysis

The website has been on-line since April 11th 2008, Buchenwald remembrance day of that year. A first log analysis was done after the portal had been in use for only two months ([Ordelman et al., 2008b]), and here a comparable analysis of 14 months of use is presented (April 12 2008 through June 12 2009). Log analysis can be a valuable source of information on the use of a website, as it provides insight into interactions between the user and the service ([Jansen, 2006]). Log analyses cannot, however, provide insight into the types of users, their intentions or actual information needs.

The gathering of interaction logs was intended to obtain information on the frequency of user interactions with specific functions (e.g., hyperlinks into interview or speaker profiles, playback buttons). Through this information further development of these functions could be motivated. The use of all fields, buttons and hyperlinks with which the user could interact was logged. The log data consisted of 5,063 sessions from 3,083 different IP addresses[12].

Comparable to results found for the 'Radio Oranje' search system, see [Heeren & de Jong, 2008], we found that the 'Show all' button was used 77% of the time, whereas a query was typed 23% of the time. If a user typed a query, which occurred 2,340 times, it consisted of or contained a named entity in the majority of cases; most requests were for names of interviewees, other people and cities. Furthermore, most queries were fairly short, i.e., one to three words, and for about 39% of the queries, no results were found.

The links to the personal details and the short summary (see Fig. 3) were used regularly, 6,168 and 4,728 times respectively. Users also often chose to follow a link into an interview (video and elaborate summary); this occurred over 5,100 times. Across users we found over 10,000 clicks on the timeline visualization showing that they interacted with the video during playback. This moreover was the preferred way for video playback control, as the control buttons were used much less: 2,600 times. Especially the playback button was used minimally (250 times). The frequency of visits to the Buchenwald search system was high directly after its release, but after the first month visit frequency lowered and seemed to stabilize to about 40-60 visits per week (see Figure 5). The frequency increased again in May 2009, which is explained by the yearly, Dutch remembrance days for World War II in the beginning of that month.

The average user session had a duration of about 12 minutes (excluding sessions under one minute from the analysis). Most user sessions were under 10 minutes (66%). About 31% of users spent between 11 and 60 minutes at the website, and an other 3% stayed over an hour. The majority of sessions seems indicative of users 'having a look around' as they are relatively short. However, the website also appears to attract users who are trying to answer more elaborate information needs.

[12]Our department's IP addresses were filtered out before analysis.

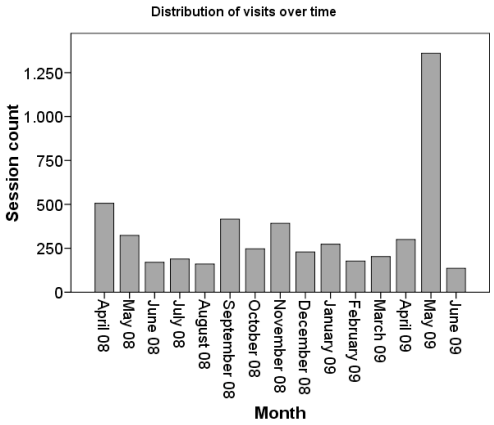


Figure 5: Distribution of visits to the Buchenwald interview search website over time.

4.1.2 Heuristic evaluation

The heuristics evaluation was intended to assess the general usability of the system, ([Nielsen & Molich, 1990]). Using both Nielsen's heuristics[13] and a short set of questions targeting this particular system's functionality a group of students from the University of Washington's Information School, who visited the Netherlands during an exchange program, gave their opinion of the strong and weak points of the design. During interaction with the system, they filled out a form on which the questions were formulated in English. Even though the native language of these evaluators, i.e. English, differed from the language used in text and recordings of the web portal, i.e. Dutch, they were well able to perform the task; many of the crucial terms for understanding results are identical or similar in English and Dutch (e.g., 'interview', 'fragment' or 'personalia'), date and duration formats can be recognized without knowledge of the language, and hyperlinks and playback controls are language-independent. Participants were furthermore not asked to judge the relevance of retrieved interview fragments.

The analysis revealed which functionality the users liked and also gave suggestions for improvement. The users were content with (i) the marking of query term locations in the timeline visualization, (ii) the fact that the entire audio document is presented as context during playback, (iii) the presentation of interview screen shots along with the result list, (iii) the color coding of multiple search terms in the results, (iv) the collapsable presentations of e.g., 'personalia', and (v) the 'please be patient' message that is shown when video loading takes several seconds.

Suggestions for improvement were (i) to make it more clear why a particular result is retrieved when the search term cannot be found in the result snippet, (ii) to make the screen shot a clickable link to the video, (iii) to more clearly present the distinction between the options of video playback and description text on the playback page, (iv) to make the distinction between fragments as results or entire interviews as results more clear, (v) to cluster results per interview instead of list each fragment separately, and (vi) to make playback control more like that found in commercial players, as that is what users are used to.

4.2 Ongoing development

The findings reported in the previous subsection formed the basis for ongoing design changes to the Buchenwald UI. For instance, to better show why a particular audio fragment is retrieved, keywords extracted from the ASR transcript are shown together with the fragment (see also [Haubold & Kender, 2004]). We chose to present keywords by extracting the words with the highest *tf.idf* s per fragment, and showing these words clustered by fragment together with the video player. The full ASR text was not used as users have been found to discard such

[13] `http://www.useit.com/papers/heuristic/heuristic_list.html,lastvisitedNov.24,2009`

texts when transcripts' word error rates lie over 30% (e.g., [Stark et al., 2000]), which is the case for this collection. However, even if the ASR output would have been absolutely correct, a literal transcript of spontaneous speech is very difficult to read due to ungrammatical sentences, hesitations, frequent use of interjections such as 'uh', and repetitions or restarts.

In addition, the playback controls are in the process of being further developed, mainly based on the results from the log analysis. As the restart button was minimally used, it was removed from the new design. Thirdly, as the timeline visualization was being used extensively, it was developed further. In the next version, the overview bar will only show the distribution of query terms over time. The segment boundaries will be left out. The more detailed view, including the segment boundaries, will be given by the zoom bar. Through this change we intend to make a better use of the timeline visualization by reducing redundancy in the individual bars and locking information to the interval size at which it is most relevant. Finally, the other suggestions given by evaluators at the end of section 4.1, e.g., result clustering and making the photo a clickable link into the video, were taken over and implemented.

5 Discussion and conclusion

We have presented a study focussing on the affordable application of speech-based annotation in creating access functionality to a video archive: the recorded interviews with Dutch survivors of World War II concentration camp Buchenwald. We outlined a retrieval framework consisting of open-source components, developed with the aim to ease integration in real-life work-flows: (i) a speech processing work-flow that minimizes manual intervention and can smoothly follow up on available resources (manual annotations, metadata, collateral data), (ii) a highly flexible search system that supports the retrieval of arbitrary document fragments and combinations of metadata fields, and allows to change ad hoc result presentation, and (iii) components for search front-ends typically designed for browsing and searching audiovisual content.

Concerning the accuracy of the automatic annotation, we have shown that domain adaptation helps, but that the data characteristics of the Buchenwald collection (spontaneous, elderly speech with sometimes strong accents), without the help of a substantial amount of example data for training, forces us to temper expectations regarding usability of the speech recognition transcripts. Although elaborate descriptive metadata was available that could potentially be used for synchronization with the speech and the creation of pointers at the within-document level, we saw (i) that the descriptions deviated too much with respect to time-synchronicity from the speech in the interviews to allow direct synchronization using forced alignment, and (ii) that the performance of full-scale large vocabulary speech recognition was too noisy te be of any help finding global alignment. As a result, the ideal case of using a lightweight ASR configuration to time-align the available metadata to support multi-level search had to be abandoned. Instead, we had to fall back to sub-optimal speech recognition transcripts generated by deploying all available adaptation strategies in the large vocabulary speech recognition setup.

In spite of the low transcription accuracy, users indicated that they appreciated the search functionality, although this cannot be supported by a formal evaluation. The log analysis showed that the web portal is being used regularly, and users seem to find their way around the website rather successfully. Some usability issues were identified with the help of students who assessed its use, but questions and improvement issues remain. One of the questions we cannot answer from analyzing the logs is which kinds of people search the interviews and why they do so. From feedback addressed to the NIOD we know of one type of use: users visit the website to learn more about their family histories by listening to stories of relatives and/or stories of people who went through the same ordeals. The peak in the number of visits in the month May of 2009, associated with national remembrance days, furthermore suggests that the website is used as a reference for knowledge about camp Buchenwald and World War II. As for improvement to the interview access website we have already introduced some ongoing development in section 4.2; a soon to be published update of the user interface will include the changes that were designed to overcome most of the shortcomings identified in the evaluation phase.

We addressed some out of a wide range of issues that are related to the affordability of the implementation of automatic annotation tools in a real-world application. We argued that proofs-of-concept tend to disregard

the feasibility of implementation, and that awareness of the interrelation between technology aspects, collection requirements, and available resources, should take shape in order to let content owners be able to define appropriate strategies for disclosing their collections.

The application was built out of open-source components that have the advantage to be generally less expensive then commercial packages (as far as these are available). It must be noted however that the consultation of experts may often be required for optimization. In certain cases, costs for the use of language resource, such as models and thesauri, have to be considered. For example, acoustic models and language models are typically created with a lot of manual effort and/or by purchasing expensive annotated corpora and usually not freely available, or only in sub-optimal versions. In general, the components are designed to reduce both the need for manual intervention and the dependency on expensive language resources. Examples provided in this paper are unsupervised model adaptation schemes for acoustic modeling and language modeling, and the full exploitation of available textual resources. We showed that in cases, a minimal amount of manual effort can help to improve accuracy and/or usability, such as by manually annotating small data portions of speakers, or manually assigning speaker labels to speaker IDs coming from speaker diarization. Finally, we stressed the importance of investigating what is already available in advance. In our experience, collateral data sources are often neglected as valuable input for the automatic disclosure of audiovisual content.

Acknowledgments

This paper is based on research funded by the NWO program CATCH (`http://www.nwo.nl/catch`) and by bsik program MultimediaN (`http://www.multimedian.nl`).

References

[Anderson et al., 1999] Anderson, S., Liberman, N., Bernstein, E., Foster, S., E., C., Levin, B., & Hudson, R. (1999). Recognition of elderly speech and voice-driven document retrieval. In *Proceedings of the ICASSP* Phoenix.

[Anguera et al., 2007] Anguera, X., Wooters, C., & Pardo, J. (2007). Robust speaker diarization for meetings: ICSI RT06s evaluation system. In *Machine Learning for Multimodal Interaction (MLMI)*, volume 4299 of *Lecture Notes in Computer Science* Berlin: Springer Verlag.

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

[De Jong et al., 2006] De Jong, F., Ordelman, R., & Huijbregts, M. (2006). Automated speech and audio analysis for semantic access to multimedia. In *Proceedings of Semantic and Digital Media Technologies, SAMT 2006*, volume 4306 of *Lecture Notes in Computer Science* (pp. 226–240). Berlin: Springer Verlag. ISBN=3-540-49335-2.

[Garofolo et al., 2000] Garofolo, J., Auzanne, C., & Voorhees, E. (2000). The TREC SDR Track: A Success Story. In *Eighth Text Retrieval Conference* (pp. 107–129). Washington.

[Gillick & Cox, 1989] Gillick, L. & Cox, S. (1989). Some statistical issues in the comparison of speech recognition algorithms. *Acoustics, Speech, and Signal Processing, 1989. ICASSP-89., 1989 International Conference on*, (pp. 532–535 vol.1).

[Goldman et al., 2005] Goldman, J., Renals, S., Bird, S., de Jong, F. M. G., Federico, M., Fleischhauer, C., Kornbluh, M., Lamel, L., Oard, D. W., Stewart, C., & Wright, R. (2005). Accessing the spoken word. *Int. Journal on Digital Libraries*, 5(4), 287–298.

[Haubold & Kender, 2004] Haubold, A. & Kender, J. R. (2004). Analysis and visualization of index words from audio transcripts of instructional videos. In *ISMSE '04: Proceedings of the IEEE Sixth International Symposium on Multimedia Software Engineering (ISMSE'04)* (pp. 570–573). Washington, DC, USA: IEEE Computer Society.

[Heeren & de Jong, 2008] Heeren, W. & de Jong, F. (2008). Disclosing spoken culture: user interfaces for access to spoken word archives. In *People and computers XXII. Culture, creativity, interaction. Proceedings of HCI 2008*, volume 1 (pp. 23–32). Swindon: The British Computer Society. ISBN=978-1-906124-04-5.

[Heeren et al., 2007] Heeren, W., van der Werff, L., Ordelman, R., van Hessen, A., & de Jong, F. (2007). Radio oranje: Searching the queen's speech(es). In C. Clarke, N. Fuhr, N. Kando, W. Kraaij, & A. de Vries (Eds.), *Proceedings of the 30th ACM SIGIR* (pp. 903–903). New York: ACM.

[Hiemstra et al., 2006] Hiemstra, D., Rode, H., van Os, R., & Flokstra, J. (2006). PF/Tijah: text search in an XML database system. In *Proceedings of the 2nd International Workshop on Open Source Information Retrieval (OSIR)* (pp. 12–17).

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

[Huijbregts et al., 2007a] Huijbregts, M., Ordelman, R., & de Jong, F. (2007a). Annotation of heterogeneous multimedia content using automatic speech recognition. In *Proceedings of Semantic and Digital Media Technologies, SAMT 2007*, volume 4816 of *Lecture Notes in Computer Science* (pp. 78–90). Berlin: Springer Verlag.

[Huijbregts et al., 2009] Huijbregts, M., Ordelman, R., van der Werff, L., & de Jong, F. (2009). Shout, the university of twente submission to the n-best 2008 speech recognition evaluation for dutch. In *Proceedings of Interspeech* Brighton, UK.

[Huijbregts et al., 2007b] Huijbregts, M., Wooters, C., & Ordelman, R. (2007b). Filtering the unknown: Speech activity detection in heterogeneous video collections. In *proceedings of Interspeech* Antwerp, Belgium.

[Jansen, 2006] Jansen, B. (2006). Search log analysis: what it is, what's been done, how to do it. *Library & Information Science Research*, 28, 407–432.

[Kessens & van Leeuwen, 2007] Kessens, J. & van Leeuwen, D. (2007). N-best: The Northern- and Southern-Dutch benchmark evaluation of speech recognition technology. In *Interspeech* Antwerp, Belgium.

[Nielsen & Molich, 1990] Nielsen, J. & Molich, R. (1990). Heuristic evaluation of user interfaces. In *CHI '90 Proceedings* (pp. 249–256).

[NIST, 2006] NIST (2006). Nist spoken term detection (std) 2006 evaluation plan. http://www.nist.gov/speech/tests/std/docs/std06-evalplan-v10.pdf.

[Oostdijk, 2000] Oostdijk, N. (2000). The Spoken Dutch Corpus. Overview and first evaluation. In M. Gravilidou, G. Carayannis, S. Markantonatou, S. Piperidis, & G. Stainhaouer (Eds.), *Second International Conference on Language Resources and Evaluation*, volume II (pp. 887–894).

[Ordelman, 2003] Ordelman, R. (2003). *Dutch Speech Recognition in Multimedia Information Retrieval*. Phd thesis, University of Twente, Enschede.

[Ordelman et al., 2006] Ordelman, R., de Jong, F., & Heeren, W. (2006). Exploration of audiovisual heritage using audio indexing technology. In L. Bordoni, A. Krueger, & M. Zancanaro (Eds.), *Proceedings of the first workshop on intelligent technologies for cultural heritage exploitation* (pp. 36–39). Trento: Universit di Trento. ISBN=not assigned.

[Ordelman et al., 2008a] Ordelman, R., de Jong, F., van Hessen, A., & Hondorp, H. (2008a). TwNC: a multi-faceted dutch news corpus. http://wwwhome.cs.utwente.nl/ ordelman/twnc/TwNC-ELRA-final.pdf.

[Ordelman et al., 2008b] Ordelman, R., Heeren, W., Huijbregts, M., Hiemstra, D., & de Jong, F. (2008b). Towards affordable disclosure of spoken word archives. In M. Larson, K. Fernie, J. Oomen, & J. Cigarran (Eds.), *Proceedings of the ECDL 2008 Workshop on Information Access to Cultural Heritage (IACH2008)* (pp.1̃5). Amsterdam, The Netherlands: ILPS, University of Amsterdam. ISBN=978-90-813489-1-1.

[Pallet et al., 1990] Pallet, D., Fisher, W., & Fiscus, J. G. (1990). Tools for the analysis of benchmark speech recognition tests. *Acoustics, Speech, and Signal Processing, 1990. ICASSP-90., 1990 International Conference on*, (pp. 97–100 vol.1).

[Stark et al., 2000] Stark, L., Whittaker, S., & Hirschberg, J. (2000). ASR satisficing: the effects of ASR accuracy on speech retrieval. In *Proceedings of International Conference on Spoken Language Processing, 2000*.

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

[Young et al., 2000] Young, S., Kershaw, D., Odell, J., Ollason, D., Valtchev, V., & Woodland, P. (2000). The HTK book version 3.0. Cambridge, England, Cambridge University.