Towards Affordable Disclosure of Spoken Word Archives

Roeland Ordelman, Willemijn Heeren, Marijn Huijbregts, Djoerd Hiemstra, Franciska de Jong

University of Twente, Enschede, The Netherlands

Abstract. This paper presents and discusses ongoing work aiming at affordable disclosure of real-world spoken word 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, the least we want to be able to provide is 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 still far from 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. To apply even the most basic form of archiving is hardly feasible for some content owners. Others need to make selective use of their annotation capacity. To guarantee content-based access to and exploitability of large amounts of potentially rich content the automation of semantic annotation seems necessary in this context.

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 [4, 7, 6]. However, success stories of the application of speech-based annotation for real-world archives lag behind. In our view, this may be 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 –often learned along with annual benchmark evaluations, as is the case with broadcast news or meeting data–, (ii) form a relatively homogeneous collection, and (iii) are annotated in quantities that are sufficient for speech recognition training and evaluation purposes. In actual practice however, the exact features of archival data are often unknown and the data may be far more heterogeneous in nature than those usually seen in the laboratory.

Annotated sample data resembling the audio conditions in these archives is typically not available via the usual channels[1], 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 [13], and archival data. We therefore 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.

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

Even when speech recognition can provide sufficiently accurate annotations, the question is how could these 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 artistic content, e.g., reflecting feelings or atmospheres. 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 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.

The development and tuning of collection-specific, automatic annotation tools is still far from 'affordable' or 'feasible' in real-world applications. Take for example automatic speech recognition (ASR). The introduction of ASR in a multimedia archive involves both fixed costs, regardless of the size of the collection, and variable costs that accrue depending on the size of the collection. Presently available ASR techniques require the investment of effort in several kinds of pre-processing, such as manual transcription of substantial quantities of representative speech. For the automatic transcription in the MALACH project for example, a large corpus (65-84 hours reported in [4]) was created for training the ASR system. Moreover, for non-static collections (e.g., news or regularly recorded meetings) system adaptation to the dynamics of the content (e.g., changing topics, new speakers) is critical. It is as yet not clear how to leverage these investments across diverse collections.

In this paper we present and discuss ongoing work that aims at the affordable disclosure of spoken word archives within the context presented above, related to a real-world use case: 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[2] 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 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 (see [8, 15] for a more detailed description). The 'Radio Oranje' project is an outstanding real-world example of how the use of automatic content-based indexing – with a special role for speech recognition technology – can boost the exploitability of a spoken word archive. Since its launch in the beginning of 2007 it has been visited over 1500 times.

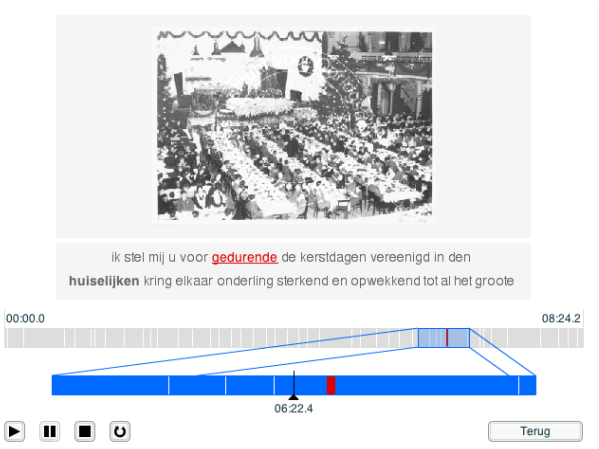


Fig. 1. Screen shot of the 'Radio Oranje' application showing bars visualizing the audio (entire speech and current segment), subtitles (keyword in bold-face, current word underlined) and an image that relates to the content of the speech (in this case: Christmas).

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[3] 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 a 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),

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

such as photos from the Picture Bank World War II[4] and maps and floor plans of the camp, or even related data available from the World Wide Web.

The first phase of the project was dedicated to the development of an online browse and search application for the disclosure of the testimonies. 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 Twente Spoken Heritage Access & Retrieval system description

The Buchenwald portal is implemented through the Twente Spoken Heritage Access & Retrieval framework consisting of freely available open-source components: (i) an analysis component that currently contains a module for video format converting 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[5] and it is available under the GNU General Public License[6]. 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 [11]).

PF/Tijah is a text search system (Tijah) that is integrated with the Pathfinder (PF) XQuery compiler [9]. It can be downloaded as part of MonetDB/XQuery, a general purpose XML database management system[7]. 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

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

[5] 'Spraak Herkennings Onderzoek Universiteit Twente' ('Speech recognition research University of Twente')

[6] `http://hmi.ewi.utwente.nl/~huijbreg`

[7] http://dbappl.cs.utwente.nl/pftijah

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 and will be developed further for use in interview collections. In the next section, we zoom in on the automatic speech processing component within the framework.

3 Automatic annotation using automatic speech recognition

Given the data that are generally available with spoken word collections –audio, some form of metadata, a thesaurus, collateral data– that can be used as a starting point for providing document level, within-document level and cross-media search, there are three strategies for automatic annotation based on speech that in principal can be pursued: (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., [18]). 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 surprisingly low text-speech correlation levels as well, especially when some additional trickery is applied. 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 speech recognition, it is accurate (e.g., with respect to collection-specific words) 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. Unfortunately, the correlation with the actual speech is too low for the application of the alignment strategy to be successful. As speech

recognition accuracy was not very high (as will be discussed below), an approach that tries to find global alignments using the time-labeled speech recognition transcripts, 'anchoring' (see also [5]), 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 higher accuracy than we had available. As even metadata with a low text-speech correlation can be used to select collection-specific terms that should have access points to the collection, deploying a word-spotting or spoken term detection approach seems obvious. In section 3.2 below we will report on preliminary attempts in this direction. In the next section 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 extensive 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.

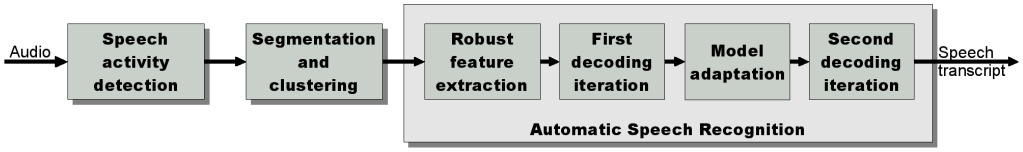


Fig. 2. Overview of the decoding system. Each step provides input for the following step. There are three subsystems: segmentation, diarization and ASR.

After SAD, the speech fragments are segmented and clustered (diarization: "who speaks when?", see [17] 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. Decoding is done using an HMM-based Viterbi decoder. In the first decoding iteration, triphone acoustic models and trigram language models are used. For each speaker, a first best hypothesis aligned on a phone basis is created for unsupervised acoustic model adaptation. The second decoding iteration uses the speaker adapted acoustic models to create the final first best hypothesis aligned on a word basis. Also, for each segment, a word lattice is created that can be used for re-scoring.

Speech processing for surprise data The results of the publicly accessible Dutch broadcast news (BN) spoken document retrieval system[8] developed at the University of Twente [14] 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[9][12]. 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).

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 [10].

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.

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. 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

[8] Dutch Broadcast News SDR system: http://hmi.ewi.utwente.nl/showcases/broadcast-news-demo
[9] N-Best: Northern and Southern Dutch Benchmark Evaluation of Speech recognition Technology

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 [3] 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 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 [3, 11]. 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 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 overfitted on the adaptation data when only small amounts of adaptation data are available while it adapts the model parameters as much as possible.

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. Although only parts of the interviews were used for evaluation, the entire interviews were processed by the speech processing system as a whole (SAD, diarization and ASR) 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. It has acoustic models trained on 75 hours of speech from the Spoken Dutch Corpus (CGN, [13]) and language models trained on the 1999-2004 newspaper corpus (485 Mw) of the Twente News Corpus [16] 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 [14] 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 an enrollment (typically created by 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, 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. 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%).

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 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

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.

explain the results. On the acoustic level, the sometimes mumbling 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 considerably (see e.g., [2]). 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 speech recognition'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 [1]). 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 logs, filtered using a stoplist of 1500 words (see section 4.2 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 amount of false alarms and misses are off balance with the number of hits resulting in low actual term-weighted values.

4 Usability aspects

4.1 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,

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.

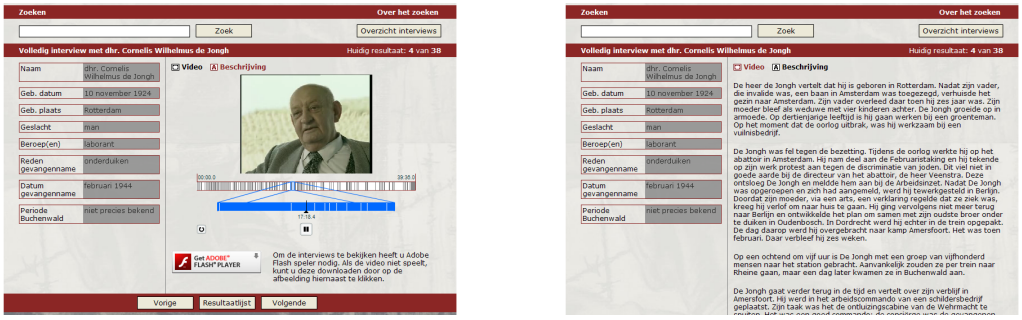
speaker profiles, short summaries) is comparable to the traditional way of accessing interview collections, and audiovisual archives in general. In such archives the lack of a direct link between the text and the corresponding fragment of audio or video makes search relatively slow and effortful. The addition of a time-stamped word-level index generated through ASR enables search *within* the videos and supports direct retrieval of video fragments.

From the user logs of the 'Radio Oranje' system we learned that many users want to browse the collection without a specific question. 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. The 'Show all' button is found to the right of the general search field (see Fig. 3).

Fig. 3. Screen shot of the 'Buchenwald' application showing the search field and a list of results.

Users can type a query in the search field; advanced search –allowing 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.



(a) Video playback

(b) Elaborate description

Fig. 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.

A user selects a result by clicking the link to the video and the elaborate description. In the case that a full interview was retrieved (text-match) the video starts playing at the beginning, in the case that a video fragment was retrieved (ASR-match) the video starts playing at the beginning of the segment in which the 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. 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.

4.2 Use of the Buchenwald UI

We conducted a preliminary analysis of the user logs for the Buchenwald UI. The website has been on-line since April 11th of this year, Buchenwald remembrance day, and we have analyzed the user logs of the interview search system from April 12 until July 14 to get a first impression of the system's use. We logged all button and link clicks and search requests.

The data consist of 1096 sessions from 699 different IP addresses[10]. As for the 'Radio Oranje' search system, we found a 2:1 ratio for the use of the 'Show all' button versus typing a query. If a user typed a query, which occurred 539 times, it consisted of or contained a named entity in almost 60% of the cases; most requests were for names of interviewees, other people and cities. Furthermore, most queries were one to three words long and for about 30% of the queries, no results were found. Table 4 shows some statistics on the query words found in the session logs. The decomposed queries produced 300 unique words of which 20.3% did not occur in the speech recognition vocabulary. Only 59% of the query words occurred in the manual metadata and about one third of the query words occurred in the speech recognition transcripts. Note, however, that (a part of) these occurrences may be false alarms.

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

single query words	300
OOV rate (ASR dct)	20.3%
terms in manual metadata	59%
terms in ASR hyps	35%

Table 4. Query word analysis

The links to the personal details and the short summary (see Fig. 3) were used regularly, over 1000 times each, to extend the information shown in the result list. Users also often chose to follow a link into an interview (video and elaborate summary); this also occurred over 1000 times. Across users we found about 1000 clicks on the timeline showing that they interacted with the video during playback, and did not just listen form where the video started playing. The play/pause button and the restart button were also used, but less often.

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 visits per week. The average user session had a duration of about 7 minutes (excluding sessions of 10 seconds or less from the analysis). Figure 5 shows that most user sessions were relatively short and therefore more indicative of 'having a look around' than of 'trying to answer an information need'.

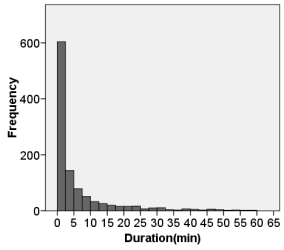


Fig. 5. Histogram for session duration.

Even though we estimate from the relatively short average for session duration that many of our visitors so far have only been 'looking around' the Website, we found that the functionality we included for access to the interviews and the related texts is being used fully and – in comparison with traditional audiovisual archives – frequently. Future monitoring of search activity will allow for a more detailed analysis of how the access functionality is being used.

5 Discussion and conclusion

We have presented a real-life use case of the 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. A description of the retrieval framework was given as well as an evaluation of ASR performance and a preliminary evaluation of the user interface.

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 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 low for using the ASR transcripts for a more global alignment. As a result, the ideal case of using a cheap and simple ASR configuration to time-align the available metadata to support multi-level search had to be abandoned. The speech recognition evaluation showed that extensive domain adaptation helps, but that the data characteristics of the Buchenwald collection (spontaneous, elderly speech with sometimes strong accents) do not yet allow for a satisfactory level of speech recognition accuracy without the help of a substantial amount of example data for training.

As for the user interface we found that its functionality is fully being used by Website visitors. One of the questions we cannot answer from analyzing the logs is why people search the interviews. From feedback addressed to the NIOD, however, 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.

Plans for further development of the Buchenwald UI include the following. In addition to the basic search field offered presently, the option of advanced search in specific metadata fields or media types may help users, as well as the option to rank results according to such characteristics. The presentation of search results might benefit from clustering the results per interview; currently each result is presented separately, which, with frequently occuring words, lead to a long list of results from a single interview. Furthermore, the first view of a result does not reveal which query terms were matched in which context and how often. For the text results it would be feasible to show this information, but for results found in the ASR-based index it is not so straightforward, as the high word error rates make presentation of the ASR output to users basically pointless. However, even if the ASR output would be absolutely correct, a literal transcript of spontaneous speech is very difficult to read due to ungrammatical sentences, hesitations and frequent use of interjections such as 'uh'. One way of representing the spoken content would be the use of key words or key phrases that we will experiment with for use in future search systems.

The timeline visualization shown during playback shows both segment boundaries and query term locations. Since the density of segment boundaries may be high, their representation in the upper bar of the timeline may obscure the location of query terms. In a next version therefore we will experiment with showing the segment boundaries in the zoomed-in view only, as well as with the presentation of other information layers, such as the speaker turns (interviewer-interviewee).

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

1. Nist spoken term detection (std) 2006 evaluation plan. http://www.nist.gov/speech/tests/std/-docs/std06-evalplan-v10.pdf, 2006.
2. S. Anderson, N. Liberman, E. Bernstein, S. Foster, Cate E., B. Levin, and R. Hudson. Recognition of elderly speech and voice-driven document retrieval. In *Proceedings of the ICASSP*, Phoenix, 1999.
3. Xavier Anguera, Chuck Wooters, and J. Pardo. 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, October 2007. Springer Verlag.
4. W. Byrne, D. Doermann, M. Franz, S. Gustman, J. Hajic, D. Oard, M. Picheny, J. Psutka, B. Ramabhadran, D. Soergel, T. Ward, and W-J. Zhu. Automatic recognition of spontaneous speech for

access to multilingual oral history archives. *IEEE Transactions in Speech and Audio Processing*, 12(4):420–435, 2004.

5. F.M.G. de Jong, R.J.F. Ordelman, and M.A.H. Huijbregts. 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*, pages 226–240, Berlin, 2006. Springer Verlag. ISBN=3-540-49335-2.
6. J.S. Garofolo, C.G.P. Auzanne, and E.M Voorhees. The TREC SDR Track: A Success Story. In *Eighth Text Retrieval Conference*, pages 107–129, Washington, 2000.
7. J. Goldman, S. Renals, S. Bird, F. M. G. de Jong, M. Federico, C. Fleischhauer, M. Kornbluh, L. Lamel, D. W. Oard, C. Stewart, and R. Wright. Accessing the spoken word. *Int. Journal on Digital Libraries*, 5(4):287–298, 2005.
8. W.F.L. Heeren, L.B. van der Werff, R.J.F. Ordelman, A.J. van Hessen, and F.M.G. de Jong. Radio oranje: Searching the queen's speech(es). In C.L.A. Clarke, N. Fuhr, N. Kando, W. Kraaij, and A. de Vries, editors, *Proceedings of the 30th ACM SIGIR*, pages 903–903, New York, 2007. ACM.
9. Djoerd Hiemstra, Henning Rode, Roel van Os, and Jan Flokstra. PF/Tijah: text search in an XML database system. In *Proceedings of the 2nd International Workshop on Open Source Information Retrieval (OSIR)*, pages 12–17, 2006.
10. M.A.H. Huijbregts, R.J.F. Ordelman, and F.M.G. de Jong. Annotation of heterogeneous multimedia content using automatic speech recognition. In *Proceedings of SAMT 2007*, volume 4816 of *Lecture Notes in Computer Science*, pages 78–90, Berlin, 2007. Springer Verlag.
11. Marijn Huijbregts. *Segmentation, Diarization and Speech Transcription: Surprise Data Unraveled.* PhD thesis, University of Twente, November 2008.
12. Judith Kessens and David van Leeuwen. N-best: The northern- and southern-dutch benchmark evaluation of speech recognition technology. In *Interspeech*, Antwerp, Belgium, August 2007.
13. N. Oostdijk. The Spoken Dutch Corpus. Overview and first evaluation. In M. Gravilidou, G. Carayannis, S. Markantonatou, S. Piperidis, and G. Stainhaouer, editors, *Second International Conference on Language Resources and Evaluation*, volume II, pages 887–894, 2000.
14. R.J.F. Ordelman. *Dutch Speech Recognition in Multimedia Information Retrieval.* Phd thesis, University of Twente, Enschede, October 2003.
15. R.J.F. Ordelman, F.M.G. de Jong, and W.F.L. Heeren. Exploration of audiovisual heritage using audio indexing technology. In L. Bordoni, A. Krueger, and M. Zancanaro, editors, *Proceedings of the first workshop on intelligent technologies for cultural heritage exploitation*, pages 36–39, Trento, 2006. Universit di Trento. ISBN=not assigned.
16. Roeland Ordelman, Franciska de Jong, Arjan van Hessen, and Hendri Hondorp. Twnc: a multifaceted dutch news corpus. http://wwwhome.cs.utwente.nl/ ordelman/twnc/TwNC-ELRA-final.pdf, 2008.
17. S.E. Tranter and D.A. Reynolds. An overview of automatic diarization systems. *IEEE Transactions on Audio, Speech and Language Processing*, 14(5):1557–1565, 2006.
18. S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland. The HTK book version 3.0., 2000. Cambridge, England, Cambridge University.