AFFORDABLE ACCESS TO MULTIMEDIA BY EXPLOITING COLLATERAL DATA

Willemijn Heeren, Roeland Ordelman and Franciska de Jong

Human Media Interaction
University of Twente
{w.f.l.heeren,ordelman,fdejong}@ewi.utwente.nl

ABSTRACT

In addition to multimedia collections and their metadata, there often is a variety of collateral data sources available on (parts of) a collection. Collateral data – secondary information objects that relate to the primary multimedia documents – can be very useful in the process of automated generation of annotations for multimedia archives in that they reduce both costs and effort in annotation and access. Furthermore, they can be used to enhance result presentation in retrieval engines. To optimally exploit collateral data, methods for automatic indexing as well as changes in the current archiving workflow are proposed.

1. INTRODUCTION

Content-based indexing of multimedia is a prerequisite for conceptual search by both professional users and the general public, and it plays an important role in the exploitability of content. Given the sheer quantity of digital multimedia collections – which is growing by both the hours of materials that are being created on a daily basis and the digitization of existing analog collections – the traditional manual annotation of collections puts heavy demands on resources. Some content owners do not have the resources to apply the most basic form of archiving, while others need to make selective use of their annotation capacity. Hence, from a cost perspective the automation of semantic annotation seems necessary to guarantee some basic form of access to large amounts of potentially rich content.

Although there is a common agreement that speech-based, automatic annotations of audiovisual archives may boost the accessibility of these archives enormously [1, 2, 3], success stories of the application of speech-based annotation for real-life archives lag behind. This seems to be due to a mismatch between laboratory and real-life conditions. In the laboratory (to some extent represented by benchmark evaluation activities such as TREC and CLEF) the focus is usually on data that (i) have well-known characteristics, often learned along with annual benchmark evaluations (e.g., NIST Rich Transcription series[1]), as is the case with broadcast news or meeting data, (ii) form a relatively homogeneous collection, and (iii) are annotated in large quantities for speech recognition training purposes. Moreover, only a handful of languages (typically American-English, Spanish, Arabic, and Chinese) are addressed. In real-life however, the exact characteristics of archival data are often unknown, and are far more heterogeneous in nature than those found in laboratory conditions. Moreover, in addition to the fact that annotated 'example' data is typically not available (especially for less common languages) via the usual channels such as The Linguistic Data Consortium (LDC) or the Evaluations and Language resources Distribution Agency (ELDA), the match between potential example data and the overall data characteristics is small. We therefore refer to real-life data as 'surprise' data.

The Academia collection from The Netherlands Institute for Sound and Vision is as a typical example of such a heterogeneous collection. It is currently being 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 a large gap between automatic speech recognition (ASR) performance on broadcast news data and performance on this collection, which to a large extent was due to the mismatch between training and testing conditions, see [4].

1.1. The Triple-A approach

On the basis of our experience with collections such as the TRECVID collection we propose a 'triple-A approach' towards speech-based annotation of surprise data. Our triple-A approach tries to maximize the potential benefit from applying ASR by focusing on (i) the *automation* aspect that sets the level of automatic processing that is required or possible given the collection, (ii) the *accuracy* aspect that aims at an optimal performance under the circumstances, given available resources and related to the last focus of attention, and (iii) the *affordability* aspect that sets the boundaries for manual effort. We will shortly address these three aspects.

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).

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

978-1-4244-2044-5/08/$25.00 ©2008 IEEE.

Automation Whereas traditional manual annotation is often necessarily limited to assigning terms from a thesaurus, automatically created annotations may yield annotations with a wider coverage and more access points into the collection. They range from full-text transcripts of the spoken word tracks generated via ASR, e.g., [5], to audiovisual source information on the location of music and on who is talking, see [6], and on the presence of objects and events detected on the basis of image analysis, e.g., [7].

Accuracy The quality of automatic analyses is dependent on various data characteristics and the match between the models and the data. Surprise speech data typically results in high Word Error Rates, as it often contains spontaneous and even non-standard speech, interrupted by musical intervals and other audio events, and is recorded under suboptimal circumstances. Depending on the users' access requirements and the resources available to alleviate errors, the degree of accuracy may fluctuate: e.g., a higher degree of accuracy is required for locating specific fragments in documents than for locating documents in a collection.

Affordability The development and tuning of collection-specific, automatic annotation tools is still far from 'affordable' or 'feasible' in real-life applications. To illustrate this, let's zoom in on the affordability of 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. 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.

1.2. Contents

In this paper we discuss the effect of the exploitation of so-called *collateral data* on spoken document retrieval (SDR) according to the triple-A approach. We will argue that collateral data can be exploited to (i) produce highly accurate, automatic indexes in an affordable way (section 3.1), (ii) tune speech and language processing tools to the focus domain (section 3.2), and (iii) enhance presentation of SDR search results by adding extra layers of information (section 4). Collateral data should therefore be the main focus of attention for all parties interested in increased access for media archives based on minimum-effort approaches. In section 5 we will present advances that can be made over the current archiving practice by the use of indexing technology, and propose which steps could be taken to facilitate the exploitation of collateral data in real-life archiving processes. First, we explain in more detail what we mean by collateral data.

2. THE POTENTIAL OF COLLATERAL DATA FOR INDEXING

The Webster online English dictionary defines collateral as "parallel, coordinate, or corresponding in position, order, time, or significance" and in this paper we will use the term to refer to data that is somehow related to the primary content objects, but that is not regarded as metadata. The term metadata will be used to refer to the description of documents or collections as found in a catalog or index. Metadata may consist of content descriptors that reflect the coverage of the audiovisual document, such as summaries and keywords, and of contextual descriptors, also called surface features, that specify e.g., document length, the document's location, and its production date. In contrast to metadata, collateral data are not describing a primary media object. They can be documents by themselves, produced either as byproduct in the pre-production or post-production stage (e.g., scripts, program guide summaries, reviews), or independently of the primary object (e.g., related newspaper articles).

Despite the fact that metadata and collateral text data can be formally separated, collateral text data may show great overlap with content descriptions that are part of the metadata. They may also be used to generate metadata descriptions, but once these have been created, the multimedia documents and those collateral data sources become separate objects again. Take, for instance, subtitling information for the hearing-impaired (e.g., CEEFAX pages 888 in the UK) that is available for the majority of contemporary broadcast items, at least for news programs. Subtitles contain a nearly complete transcription of the words spoken in the video items, and provide an excellent information source for automatic indexing. They, however, are usually not part of the archival description of broadcast material.

With respect to spoken word collections, there are many cases where metadata is not or only sparsely available, which is why those collections were only searchable by linear exploration. From the results of the TREC Spoken Document Retrieval (SDR) tracks it can be inferred that speech recognition technology can be used for indexing purposes: as a result of massive research efforts along with the TREC SDR tracks from 1997 (TREC6) until 2000 (TREC9), the SDR problem was declared 'solved' [8], at least for the broadcast news domain, for the American-English language.

A fact that was evidently not missed by the researchers involved, but that might have been underestimated by non-specialists, is that for other domains and for other languages solving the access problem using ASR technology can be complicated, or may at least require large investments in both the pre-production stage (e.g., doing domain-specific modeling) and the production stage itself (e.g., adaptation to longitudinal dynamics). More generally, SDR is not yet part of a well-established toolbox. Against this background, the potential use of collateral data as an indexing source or as a valuable source of information for collection-specific ASR development will be discussed in section 3.

As mentioned earlier, collateral data in the broadcast news domain can be found in the form of subtitling information for the hearing impaired. It is an obvious and cheap solution for indexing. The number of words in teletext subtitling transcripts is typically cut down drastically due to a minimum of available space on a screen, mixing up phrases in an attempt to convey the same message with less and often other words. Nevertheless, already in 1995 the feasibility of using subtitling for indexing was demonstrated [9]. In that case, subtitles where recorded using a teletext decoder (nowadays available with most video capture TV cards) and synchronized with the video by adding the time at which each line of text appeared during the broadcast. This was accurate to within a few seconds.

Broadcast news subtitles in The Netherlands even incorporate topic boundaries that can be used to segment a news show into subdocuments. In order to demonstrate the added value of links that are automatically generated across media types via collateral data, the 'linguistic annotations' of the news items (subtitles) were linked to an up-to-date database of Dutch newspaper articles made available for demonstration purposes by PCM publishers. The links from broadcast news fragments to related newspaper articles were generated by (i) using a stopped version of the textual video annotation as a query for a search in the newspaper archive, (ii) matching the query with the content in the newspaper archive using Okapi term weighting, and (iii) presenting the top-n results in a clickable list, ordered by date or by relevance.

Textual sources that can play a similar role are teleprompter texts –also referred to as auto-cues– read from a screen by an anchor person. Although teleprompter texts are usually an accurate representation of the anchor person's speech (with an accuracy measured on a Dutch collection of around 90%, see [10]), they often do not include transcripts of interviews and dialogs with on-site reporters.

Also outside the broadcast sector collateral data that represents the speech in a collection can be found. A collection of recorded lectures may have presenter notes attached to it, speeches may be accompanied with the written text version, and in the meeting domain there may be minutes available, or at least an agenda. Furthermore, lectures may be accompanied by notes, text books and slides, and interviews recorded for research purposes are often extensively summarized of even fully transcribed.

In sum, for many spoken word documents there is some kind of collateral text data available. Two remarks have to be made however. First, availability is often a relative concept. Making several metadata and/or collateral data streams available for a proof-of-concept demonstrator as in the broadcast news cross-media browser mentioned above is one thing, restructuring a complex real-life workflow for a running application, possibly involving commitment from multiple branches or even companies, is something else. Second, the level of similarity between the collateral data and the speech, may differ. At one end of the spectrum lie full transcripts of the spoken content, and via extensive summaries or documents laying out the linear structure, such as slides or agendas, we arrive at textual documents that only relate to the speech semantically. Even when collateral data and spoken content diverge, but still correlate on the topic level, collateral data can be of added value as will be presented in the following sections.

3. INDEXING MULTIMEDIA EXPLOITING COLLATERAL DATA

Especially when the nature of the multimedia collection demands for a search functionality that allows searching *within* documents, a *time-coded* index needs to be generated. From the triple-A perspective this is preferably done fully based on available collateral data, but in practice the index is often generated using multiple resources. The similarity between the collateral data and the actual speech in the collection determines for a large part which method suits the most for indexing.

If the collateral data at hand more or less follows the speech in the collection, making it available for indexing requires the synchronization of the text data to the audiovisual source. Except for broadcast-related sources such as subtitling, collateral data usually does not contain time information. This process of labeling the text with time-codes is called the 'alignment' of text and speech, a well-known procedure used frequently in ASR, for example when training acoustic models (see e.g., [11]). The alignment of collateral data holds for surprisingly low text-speech correlation levels, especially when some additional trickery is applied. However, when the collateral data only correlates with the speech on the topic level, full-blown speech recognition must be called in, using the collateral data as a strong prior ('informed speech recognition') or source for extensive domain tuning. In section 3.1 we will first discuss alignment in more depth, followed by a discussion on using collateral data for domain adaptation in section 3.2.

3.1. Indexing through 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. Due to the low complexity of the task – everything is known except for the time labels – alignment is relatively robust against mismatches in acoustic conditions. In the following sections, the alignment strategy is discussed in more detail on the basis of two scenarios. The first one explains the standard procedure given ideal circumstances: an accurate transcript with the audio. The second scenario gives an example of how can be dealt with incomplete, collateral transcripts.

3.1.1. Alignment of accurate transcripts

When full-text transcripts are available for a multimedia collection, generating a time-stamped index at the word level is done by aligning the spoken word document with its transcription. This scenario applies in the case of e.g., speeches that were fully written out, and oral history collections or other interview collections gathered for research purposes.

An example of an archive for which full-text alignment has been applied is the so-called 'Radio Oranje' collection. Within the NWO-CATCH project CHoral a demonstrator search system was developed for this collection of radio speeches that Queen Wilhelmina of the Netherlands addressed to the Dutch people during World War II. On the basis of alignment an index was built which turned this historical collection into on online, searchable asset[2].

The collection consists of 37 speeches with lengths varying between 5 and 19 minutes. Their style is very formal and language use is complex. The recordings as well as their 1940s transcripts have been digitized by the Netherlands Institute for War Documentation (NIOD) and the Netherlands Institute for Sound and Vision, one of Europes largest audiovisual archives. The audio quality be considered poor; the recordings are noisy and contain artifacts (e.g., hiss, pops).

The alignment tool from an off-the-shelf multi-mixture, Gaussian HMM-based speech recognition engine was used, [12]. This produces Viterbi-optimized, word-based alignments. Optimal alignment performance was obtained using speaker-dependent, monophone acoustic models, trained from gender- and speaker-independent models optimized for broadcast news (see also [13]). Performance was adequate for this task: $>90\%$ of all word boundaries were found within 100 ms of the reference, i.e. within the correct syllable.

3.1.2. Alignment of incomplete transcripts

In the case that the match between the collateral text and the spoken content is incomplete, the transcripts, such as meeting minutes, can be automatically enhanced via the generation of time-stamps. This approach was taken in two pilot projects in the domain of e-Government. In the first case the meeting minutes were the so-called *Handelingen*, i.e., the minutes of the Dutch Parliament. The second case concerned minutes from city council meetings. Due to the difference in accuracy of these two types of minutes, two different approaches were developed for indexing.

The 'Handelingen' are stenographic minutes that closely follow the discourse of a parliamentary session; they are only corrected for slips of the tongue and ungrammatical sentences. Given the close match with the actual speech, a so-called forced alignment procedure could be used. This is a technique commonly used in acoustic model training for ASR. To be able to train phone models, an acoustic model is used to align words and phones in pre-segmented sentences to their exact location in the speech. In the first iteration an 'averaged' bootstrap model is used. The alignment and the model should improve iteratively. Given the sequence of words in a sentence the acoustic model tries to find the optimal distribution of these words in the corresponding audio signal. This is done on the basis of the speech sounds the words are composed of. When alignment is used for indexing, pre-segmented sentences are evidently not available. However, as long as the text follows the spoken content well enough, the word-level alignment can be found by using relatively large windows of text. This alignment procedure works well even if some words in the minutes are not actually present in the speech signal.

If the textual content does not match the speech too well, as was the case with the city council meetings, the alignment procedure may fail to find a proper alignment. In order to produce a suitable index, we used a two-pass strategy, similar to the one proposed in [14]. A baseline large vocabulary ASR system[3] is used to generate a relatively inaccurate transcript of the speech with word-timing labels. This transcript is referred to as 'hypothesis'. Next, the hypothesis is aligned to the minutes at the word level using a dynamic programming algorithm. At the positions where the hypothesis and the minutes match so called 'anchors' are placed. A match is defined as three correctly aligned words in a row. Using the word-timing labels provided by the speech recognition system, the anchors are used to generate segments. Individual segments of audio and text are accurately synchronized using forced alignment.

3.2. Indexing through Automatic Speech Recognition

If collateral text matches the speech only poorly, e.g., if there is only a correlation at the topic-level, the generation of a time-labeled index typically depends on accurate annotations from automatic speech recognition. To make accurate, automatic annotation using speech recognition technology affordable, any available textual resource, including metadata and collateral data, must be deployed to adapt the system semantically and acoustically to the task domain. Moreover, collateral data can be helpful to correct speech recognition errors in retrospect (e.g., correct "James Frown" to "James Brown" based on a parallel corpus), or used as input for query or document expansion [15, 16].

For the broadcast news domain, huge quantities of available training data allow acoustic models and language models to be trained adequately. Language models and recognition

[2] The demonstrator can be found at `http://hmi.ewi.utwente.nl/choral/demo.html`.

[3] Optionally the speech recognition is adapted to the task, for example by providing it with a vocabulary extracted from the minutes

vocabularies, however, are usually created using fixed training corpora that are often outdated. Vocabularies are based on word frequencies derived from these corpora, whereas the linguistic properties of broadcast news are continuously changing: names of places and people that were previously mentioned only infrequently may become part of current affairs without prior indication, people who dominated the news for a certain period of time disappear from the headlines, jargon may be adopted by the general public, new words are invented, and there are words that are likely to (co-)occur in one period of the year but that are highly improbable in others (e.g., *hurricane* or *Christmas*).

The properties of spoken content in other domains than broadcast news (e.g., corporate, cultural heritage) may also mismatch standard language models. For instance, oral history collections may focus on a specific period such as a war, and as a result are packed with names, keywords, or euphemisms specific to that time, but generally absent in models built from current news texts. To limit the number of out-of-vocabulary (OOV) words, the ASR engine employed in a multimedia retrieval environment should use models that can deal with linguistic variation. First, a query consisting of an OOV word, a so-called QOV (query-out-of-vocabulary), will never match terms in an ASR transcript, even if the QOV occurs in the speech. Second, the word occurring in the transcript at the position of the OOV may induce a false alarm to another query.

Several solutions to this problem have been proposed, ranging from the use of larger recognition vocabularies, to dynamic adaptation of vocabularies based on temporal information in parallel corpora or metadata that is available for a document [17, 18, 19]. Another strategy for dealing with QOV words is to avoid speech recognition vocabulary restrictions by creating audio document representations based on phones or sub-word units, instead of words [20, 21].

4. EXPLOITING COLLATERAL DATA FOR ENHANCED RESULT PRESENTATION

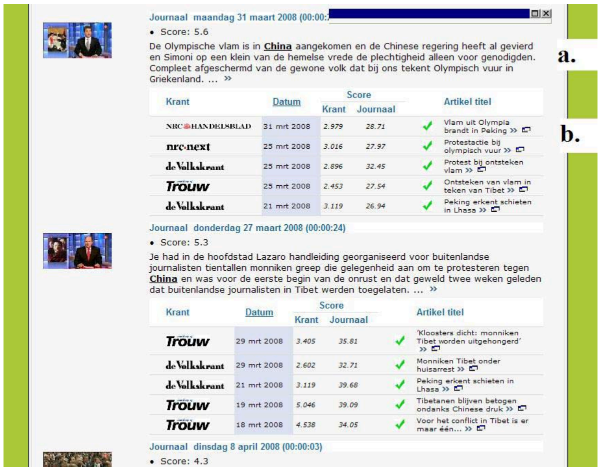
In addition to the use of collateral text data to generate annotations for multimedia collections, collateral data can be employed to enhance the presentation of multimedia search results. Instead of matching user queries to a collection's semantic representation, semantic representations of the content from different media types and collections can also be associated: cross-media linking, [22]. Through cross-media linking relations between collateral documents from different media types may be detected and the various items can be presented as one cluster. This allows users to retrieve multiple perspectives given a single query. The following subsections present demonstrator systems that incorporate some version of cross-media linking in two different domains: broadcast news and cultural heritage.

4.1. Broadcast news cross-media browsing

The system described in this section was initially developed as a demonstrator for on-line access to an archive of Dutch news broadcasts ('NOS 8 uur Journaal'). In addition to listing video results, it generates links to related newspaper articles (see figure 1).

For the generation of time-coded indexes to search within news shows, the system illustrates the exploitation of (i) subtitling information for the hearing-impaired, and (ii) ASR transcripts. Subtitling information is acquired using a teletext capturing card and synchronized with the video content using a manually determined off-set value. The speech recognition system is based on decision-tree state-clustered acoustic models trained on approximately 20 hours of speech from the Spoken Dutch Corpus [23], a vocabulary of 65K words extracted from a newspaper collection and a 3-gram language model trained on some 300M words of newspaper text data.

Currently, the speech recognition system is static; it does not update its vocabulary and language model, nor does it perform any acoustic adaptation schemes. The incorporation of such procedures is scheduled for a new version of the demonstrator. As the subtitling information provides information on topic boundaries, validated topic boundaries can be used for the segmentation of the news show into 'subdocuments'. In the speech recognition mode, these boundaries cannot be used and segmentation of the news show is done on the basis of acoustic information such as speaker changes, speech/non-speech occurrences and silences.



Journaal maandag 31 maart 2008 (00:00:

- Score: 5.6

De Olympische vlam is in **China** aangekomen en de Chinese regering heeft al gevierd en Simoni op een klein van de hemelse vrede de plechtigheid alleen voor genodigden. Compleet afgeschermd van de gewone volk dat bij ons tekent Olympisch vuur in Griekenland. ... »

Krant	Datum	Score			Artikel titel
		Krant	Journaal		
NRC HANDELSBLAD	31 mrt 2008	2.979	28.71	✓	Vlam uit Olympia brandt in Peking »
nrc·next	25 mrt 2008	3.016	27.97	✓	Protestactie bij olympisch vuur »
de Volkskrant	25 mrt 2008	2.896	32.45	✓	Protest bij ontsteken vlam »
Trouw	25 mrt 2008	2.453	27.54	✓	Ontsteken van vlam in teken van Tibet »
de Volkskrant	21 mrt 2008	3.119	26.94	✓	Peking erkent schieten in Lhasa »

Journaal donderdag 27 maart 2008 (00:00:24)

- Score: 5.3

Je had in de hoofdstad Lazaro handleiding georganiseerd voor buitenlandse journalisten tientallen monniken greep die gelegenheid aan om te protesteren tegen **China** en was voor de eerste begin van de onrust en dat geweld twee weken geleden dat buitenlandse journalisten in Tibet werden toegelaten. ... »

Krant	Datum	Score			Artikel titel
		Krant	Journaal		
Trouw	29 mrt 2008	3.405	35.81	✓	'Kloosters dicht: monniken Tibet worden uitgehongerd' »
de Volkskrant	29 mrt 2008	2.602	32.71	✓	Monniken Tibet onder huisarrest »
de Volkskrant	21 mrt 2008	3.119	39.68	✓	Peking erkent schieten in Lhasa »
Trouw	19 mrt 2008	5.046	39.09	✓	Tibetanen blijven betogen ondanks Chinese druk »
Trouw	18 mrt 2008	4.538	34.05	✓	Voor het conflict in Tibet is er maar één... »

Journaal dinsdag 8 april 2008 (00:00:03)

- Score: 4.3

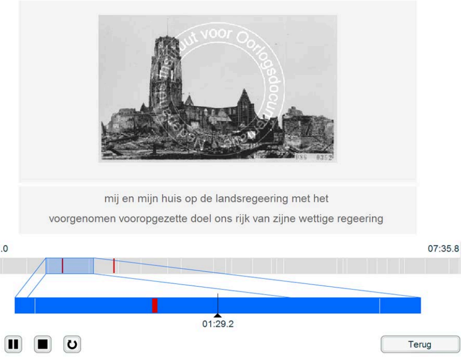
Fig. 1. Screen shot of the result page of the broadcast news search engine, listing news items together with ASR transcripts (a), and related newspaper articles (b).

The linguistic annotations of news items (based on either subtitles or ASR) were linked to an up-to-date database of Dutch newspaper articles. We can use that database for demonstration purposes by courtesy of PCM publishers, one of the largest publishers in the Dutch language region. For reasons of IPR, the public version of the demonstrator does not contain the links to these articles. The links from broadcast news fragments to related, i.e. collateral, newspaper articles are generated by (i) using a stopped version of the textual video annotation to query the newspaper archive, (ii) matching the query with the content in the newspaper archive using Okapi term weighting, [24], and (iii) presenting the top-n results in a clickable list, ordered by date or by relevance.

4.2. Cross-media linking of historical collections

Another domain in which the combination of multiple multimedia sources can enhance the quality of the user experience is cultural heritage. Now that more historical texts, images, pamphlets, photos, audiovisual materials etc. are being digitized it becomes possible (i) to automatically identify links between documents from a variety of modalities, and (ii) to present related documents in one multimedia presentation.

In the 'Radio Oranje' demonstrator (see also section 3.1.1), spoken word fragments and photographic material on the same topic, World War II, were linked (see figure 2). The spoken audio consists of speeches from Queen Wilhelmina of The Netherlands, the photographic material comes from a collection of around 120,000 photos maintained by the same institute: the photos are partly from the same period as the Radio Oranje broadcasts. Searching and browsing functionalities were developed for the spoken content, and as a preprocessing step sets of photos were semi-automatically linked to the speeches.



mij en mijn huis op de landsregeering met het
voorgenomen vooropgezette doel ons rijk van zijne wettige regeering

Fig. 2. Screen shot of the playback interface of the 'Radio Oranje' engine, with both the aligned transcripts and linked photos.

The photos were semantically described through keywords, available from their metadata database. Since the language used in the speeches was very formal and metaphorical, it was not possible to automatically match the spoken content at the sentence level to the photo keywords. Therefore, semantic representations for the speeches were generated by manually assigning one or more keywords from the photo thesaurus to each speech. This was done on the basis of its title and global content. The 19 speeches were described by (combinations) of 13 keywords such as Liberation, May 1940, Christmas or Netherlands Indies. These keywords referred to photo sets ranging in size from 2 to 200. Also, at the level of the whole collection of speeches, a number of relevant keywords was selected, containing keywords such as Illegal Press & Radio, Queen Wilhelmina and Dutch Street Scenes that can be used to select photos in the case that the speech-level sets are not large enough. The resulting multimedia presentation allows users to search the content of the speeches, or to browse through the speeches. While the audio is being played, a slide show of photos that relate to the topic is presented.

5. ENHANCED ACCESS TO MULTIMEDIA COLLECTIONS

Automatic indexing and SDR are hardly being used in real-life access applications for spoken word archives, despite extensive research investments. Exceptions can be found in research projects in the broadcast news and cultural heritage domains, such as MALACH [25], and systems such as Speechfind [5]. Numerous audio archives still depend on manual annotation. To what extent can automatic indexing help to improve the disclosure and accessibility of spoken word collections from an archivist's point of view? And secondly, how can archivists aid in the development and successful application of technology? These questions will be addressed in the following subsections.

5.1. The archivists' perspective

As part of the CHoral project, a requirements analysis was conducted among keepers of multimedia collections in the Netherlands, the largest of which was that of the Netherlands Institute of Sound and Vision (holding over 700,000 hours). Eight archiving professionals participated in semi-structured interviews to gather information on the bottlenecks in the current situation of disclosure of and access to spoken word collections, and on the collection keepers' opinion on the potential of automatic indexing technology.

In the current practice of disclosure, the restricted resources for annotation form a bottleneck. Accurately describing spoken word documents costs five to ten times real time, which can be expensive. As a result, collections remain undisclosed or are annotated by means of short descriptions or keywords only. For successful retrieval, this is generally too unspecific. Given the vast size of several historical audio collections that have *not* been disclosed, and the amount of manual labor that disclosure would cost, archivists and collection keepers acknowledge the added value of automatic indexing to generate at least some form of annotation for these materials. Moreover, interviewees suggested that such a rudimentary index could be used to decide whether the collection should be disclosed more elaborately.

The second bottleneck in the current archiving practice is the slow access procedure. In the current practice of retrieving documents, archive users usually run text searches on catalog descriptions in databases, which results in a hit list with a number of possibly relevant audiovisual documents. In contrast to other media types, such as text or images, however, the multimedia files themselves are often not linked to the retrieved result lists. This forces searchers to visit the institute maintaining the collection if they want to have the selected content played. One example of problems resulting from this restricted and relatively slow access procedure is that content producers, such as news editors, have difficulty finding and gaining access to relevant items in broadcast archives, especially when time pressure is high, as is the case with reports on unforeseen events. This means that not only annotation, but also access places high demands on resources.

Technology for multimedia retrieval can change access to collection items. Automatic indexes generated on the basis of metadata or collateral data – and that may complement current manual metadata – can be used to increase the number of access point to collections. For instance, documents would not only be accessible at the fixed levels of radio/TV programs or tracks, but also at the exact fragment or word within such a document. Together with links from the catalog to the multimedia documents, the time needed for search and the evaluation of search results is expected to decrease significantly.

To conclude, the interviews are underpinning the expectation that automatic indexing and multimedia retrieval technology could substantially improve access to multimedia collections.

The requirements analysis, however, also pointed out a number of situations where automatically generated metadata currently does not suffice as a substitute for manual metadata. In the next section, these issues will be discussed, and we will propose how they can be solved by both the application of existing technology in the domain of multimedia archiving and also by changes in the workflow of archiving.

5.2. Changes in the multimedia archiving workflow

There are a number of important aspects, where automatic annotations fall short in accuracy. The first problem relates to the so-called semantic gap. Because human interpretation is lacking from automatically generated annotations, certain abstractions cannot easily be made. 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.

To reduce the semantic gap between user queries and indexes, there is a well-established range of generally applicable methods to turn text-based indexing into something that can be considered to support automatic semantic annotation, e.g., topic clustering and automatic classification. The application of these techniques in multimedia archives is also expected to improve accessibility (see [26] for an overview of techniques and performance figures).

The second problem concerns the fact that a large number of user queries to multimedia collections involve named entities, such as personal names and locations. As was explained in section 3.2, especially named entities run the risk of being out-of-vocabulary and may therefore be irretrievable if textual sources for tuning vocabularies to specific collections are not available. One way of reducing this problem is by carefully annotating at least the names of places and persons that are associated with a certain multimedia document during the description process. Usually, metadata models encompass standard fields to enter such information. Alternatively, collateral data such as scripts and notes from producers of multimedia documents may provide this information.

Thirdly, when collections are to be used for certain types of scholarly research automatic transcription may not be suitable at all. Researchers from these fields need manually checked indexes that often abstract away from the words spoken. To generate a first version of an index, however, speech processing seems a useful technology that can be employed to reduce the amount of work. Moreover, for fast and easy access to such collections the manual annotations generated during a first pass of research can be aligned with the audiovisual documents relatively straightforwardly. In the domain of oral history, for instance, full transcripts are often made.

Archivists can help to improve access to multimedia collections by describing the link in content between related sources, i.e. the primary document and the collateral data, so that these can be traced for automatic processing. Again, the makers of collections should be made aware of the added value of collateral text so that related documents are jointly transferred to archivists. In order to benefit from the use of collateral data for cost-effective annotation and access, we propose changes in the workflow of multimedia archiving: (i) primary and secondary information objects should already be identified by producers, (ii) related sources are preferably jointly transferred to archives, and (iii) links between related sources should be described.

6. CONCLUSION AND FUTURE WORK

In this paper we have argued that exploiting collateral data is a key step in building real-life access applications for multimedia archives. We have shown that collateral data can be used for multiple purposes. First, it can be used to automatically produce highly accurate, automatic indexes in an affordable way, i.e. through alignment of transcripts that are available for many types of collections such as oral history collections, interviews, and meeting recordings. Secondly, collateral data can be used to tune speech and language processing tools to the focus domain, which lowers the resources needed for generation of training data. And thirdly, it may be used to enhance the presentation of SDR search results to users. Links to collateral data may offer a broader and deeper response to the information need. Furthermore, we have discussed these uses for collateral data against the background of the triple-A approach for the speech-based annotation of surprise data: automation, accuracy and affordability. The right balance between these three dimensions maximizes the potential benefit of applying speech technology in SDR.

In order to guarantee successful deployment of collateral data, however, changes in the archiving workflow are required. These come down to keeping primary sources linked in some way to their collateral data sources, so that the latter can be used to boost the performance of automatic annotation and access technology. This is, however, also expected to involve the adaptation of current metadata schemes, and standardization of formats and NLP pre-processing tools. These issues should be on the agendas of international archiving organizations such as the International Association of Sound and Audiovisual Archives (IASA[4]) and CHORUS[5], and should be addressed in future research. We aim to further quantify the added value of collateral data in ASR tuning, and to explore their use in result presentation to end users of multimedia archives.

7. REFERENCES

[1] 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. J. Dig. Lib.*, vol. 5, no. 4, pp. 287–298, 2005.

[2] J.S. Garofolo, C.G.P. Auzanne, and E.M Voorhees, "The TREC SDR Track: A Success Story," in *Eighth Text Retrieval Conference*, Washington, 2000, pp. 107–129.

[3] W. Byrne, D. Doermann, M. Franz, S. Gustman, J. Hajic, D. Oard, M. Picheny, J. Psutka, B. Ramabhadran, D. Soergel, T. Ward, and W-J. Zhu, "Automatic recognition of spontaneous speech for access to multilingual oral history archives," *IEEE Trans. Speech Audio Proc.*, vol. 12, no. 4, pp. 420–435, 2004.

[4] M.A.H. Huijbregts, R.J.F. Ordelman, and F.M.G. de Jong, "Annotation of heterogeneous multimedia content using automatic speech recognition," in *Proc. SAMT 2007*, Berlin, 2007, vol. 4816 of *Lecture Notes in Computer Science*, pp. 78–90, Springer Verlag.

[5] J.H.L. Hansen, R. Huang, B. Zhou, M. Deadle, J.R. Deller, A. R. Gurijala, M. Kurimo, and P. Angkititrakul, "Speechfind: Advances in spoken document retrieval for a national gallery of the spoken word," *IEEE Trans. Speech Audio Proc.*, vol. 13, no. 5, pp. 712–730, 2005.

[6] S.E. Tranter and D.A. Reynolds, "An overview of automatic diarization systems," *IEEE Trans. Audio Speech Lang. Proc.*, vol. 14, no. 5, pp. 1557–1565, 2006.

[7] A.F Smeaton, W. Kraaij, and P. Over, "Trecvid - an overview," in *Proc. TRECVID 2003*, USA, 2003, NIST.

[8] J.S. Garofolo, C.G.P. Auzanne, and E.M Voorhees, "The TREC SDR Track: A Success Story," in *Eighth Text Retrieval Conference*, Washington, 2000, pp. 107–129.

[9] M. G. Brown, J.T. Foote, G.J.F. Jones, K. Sparck Jones, and S. J. Young, "Automatic Content-based Retrieval of Broadcast News," in *Proc. the third ACM international conference on Multimedia*, San Francisco, November 1995, pp. 35–43, ACM Press.

[10] R.J.F. Ordelman, *Dutch Speech Recognition in Multimedia Information Retrieval*, Phd thesis, University of Twente, Enschede, Oct. 2003.

[11] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, "The HTK book version 3.0.," 2000, Cambridge, England, Cambridge University.

[12] B. Pellom, "Sonic: The university of colorado continuous speech recognizer," Tech. Rep., March 2001, TR-CSLR-2001-01, University of Colorado.

[13] L.B. van der Werff, W.F.L. Heeren, R.J.F. Ordelman, and F.M.G. de Jong, "Radio oranje: Enhanced access to a historical spoken word collection," in *Proc. CLIN 17*, 2007, pp. 207–218.

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

[15] P.C. Woodland, S.E. Johnson, P. Jourlin, and K. Spärck Jones, "Effects of Out of Vocabulary Words in Spoken Document Retrieval," in *2000 ACM SIGIR Conference*, Athens Greece, 2000, pp. 372–374.

[16] P. Jourlin, S.E. Johnson, K. Spärck Jones, and P.C. Woodland, "General query expansion techniques for spoken document retrieval," in *Proc. ESCA Workshop on Extracting Information from Spoken Audio*, Cambridge, UK, 1999, pp. 8–13.

[17] R. Rosenfeld, "Optimizing Lexical and N-gram Coverage Via Judicious Use of Linguistic Data," in *Eurospeech-95*, 1995, pp. 1763–1766.

[4] IASA: `http://www.iasa-web.org/`

[5] IST coordinated action, `http://www.ist-chorus.org/`

[18] C. Auzanne, J.S. Garofolo, J.G. Fiscus, and W.M Fisher, "Automatic Language Model Adaptation for Spoken Document Retrieval," in *Proc. RIAO 2000*, 2000, pp. 132–141.

[19] A. Allauzen and J.-L. Gauvain, "Open vocabulary ASR for audiovisual document indexation," in *Proc. ICASSP*, April 2005, pp. 1013–1016.

[20] K. Ng, *Subword-based Approaches for Spoken Document Retrieval*, Ph.D. thesis, Massachusetts Institute of Technology, February 2000.

[21] A.F. Smeaton, M. Morony, G. Quinn, and R. Scaife, "Taiscéalaí: Information Retrieval from an Archive of Spoken Radio News," in *Proceedings of ECDL2*, Crete, 1998, pp. 429–442.

[22] J. Morang, F.M.G. de Jong, R.J.F. Ordelman, and A.J. van Hessen, "Infolink: analysis of dutch broadcast news and cross-media browsing," in *Proc. CBMI 2005*, Amsterdam, 2005.

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

[24] S.E. Robertson, S. Walker, M.M. Beaulieu, M. Gatford, and A. Payne, "Okapi at TREC-4," 1996, Text REtrieval Conference.

[25] W. Byrne, D.Doermann, M. Franz, S. Gustman, J. Hajic, D. Oard, M. Picheny, J. Psutka, B. Ramabhadran, D. Soergel, T. Ward, and W-J. Zhu, "Automatic recognition of spontaneous speech for access to multilingual oral history archives," *IEEE Trans. Speech Audio Proc.*, vol. 12, no. 4, 2004.

[26] F.M.G. de Jong and W. Kraaij, "Content Reduction for Cross-media Browsing," in *RANLP workshop 'Crossing Barriers in Text Summarization Reserach*, H. Saggion and J.-L. Minel, Eds., Borovets, Bulgaria, 2005, pp. 64–69.