Exploration of Audiovisual Heritage Using Audio Indexing Technology

Roeland Ordelman and **Franciska de Jong** and **Willemijn Heeren**[1]

Abstract. This paper discusses audio indexing tools that have been implemented for the disclosure of Dutch audiovisual cultural heritage collections. It explains the role of language models and their adaptation to historical settings and the adaptation of acoustic models for homogeneous audio collections. In addition to the benefits of cross-media linking, the requirements for successful tuning and improvement of available tools for indexing the heterogeneous A/V collections from the cultural heritage domain are reviewed. And finally the paper argues that research is needed to cope with the varying information needs for different types of users.

1 INTRODUCTION

A number of techniques from the AI-realm have proven to have added value for spoken document retrieval. Browsing tools for audio and/or video archives not only benefit from speech recognition, but also from techniques such as clustering, topic detection, speaker classification, segmentation, etc. Most of the proofs of concept delivered for spoken audio access technology in the past decade focused on archives in the broadcast news domain. Cf. [11].

Due to the ever declining costs of recording audio and video, and due to improved preservation technology and initiatives for retrospective digitization (cf. projects such as PrestoSpace[2]), the number of digital audio collections in the cultural heritage domain is growing rapidly. This paper will review some pilot projects that have applied and tuned audio indexing tools to Dutch heritage collections. It will analyze the specific requirements imposed by the nature and formats of the collections from a technological point of view.

Once having the technology available for indexing audiovisual content on multiple levels, new questions emerge: how can we use the technology to support the information need of various different types of users including e.g., archivists, information analysts, researchers, producers of new content, the general public, etc.? Or even better, how can the information need of users be redirected by enabling new views on the data. In this paper we will start to explore these issues for a specific audiovisual collection from the perspective of historians.

In the remainder of this paper we will describe the state-of-the-art in audio indexing and identify the particular requirements for oral history collections. In section 4 the concept of full text transcription as basis for audio search and the complications stemming from the nature of cultural heritage collections will be discussed, and in section 5 the possibilities for applying alignment techniques will be reviewed. In section 6 the variance in user perspectives will be analyzed. The concluding section will identify some issues for future research.

2 TRENDS IN ORAL HISTORY COLLECTIONS

In The Netherlands, a major part of the Dutch audiovisual heritage sits at the Dutch national audiovisual archive[3]. For this audiovisual archive, one of the largest in Europe, retrospective digitization has been a major topic for quite some time now. But also at the regional level, audiovisual heritage is becoming generously available as retrospective digitization is being taken up there as well. The Municipal Archives of Rotterdam, for example, are digitizing historical radio broadcasts from a local radio station. Apart from broadcast material that has been created by professionals, there is a substantial amount of audiovisual heritage, recorded on different occasions by semi-professionals and non-professionals. But much of such content remains invisible, as it has not yet been inspected properly, let alone digitized. However, as the interest for cultural heritage is progressing, small and medium-scale digitization and disclosure initiatives can be witnessed. An example of such an initiative in The Netherlands is the digitization and indexing of a collection of interviews and lectures (literally a bunch of tapes in a box in a closet) of the Dutch novelist Willem Frederik Hermans which have been made available via a web server[4].

These audiovisual spoken-word archives, as they are often referred to, belong to the domain of what is usually called oral history: recordings of spoken interviews and testimonies on diverging topics such as retrospective narratives, eye witness reports and historical site descriptions. It also has modern variants such as 'Podcasts' and 'Vodcasts'. Note that these modern variants impose a problem that is opposite to what we are used to in the domain: due to their fleeting nature they are in fact more perishable than analogue tapes.

Whereas the growth of storage capacity is in accordance with widely acknowledged predictions, the possibilities to index and access these archives is lagging behind [3]. On the one hand, the large quantities of data of the 'broadcast' archives do not allow for costly manual annotation on multiple levels. As a consequence, particular information may only be accessible via manual browsing of a collection of files. On the other hand, especially for smaller organizations, resources are often lacking to apply even basic forms of archiving. Hence, spoken-word collections risk to become the stepchild of an archive—minimally managed, poorly preserved, and hardly accessible. For the 'MyLifeBits' chronicles collected by non-professionals

[1] Human Media Interaction, Faculty of Electrical Engineering, Mathematics and Computer Science, University of Twente, The Netherlands

[2] `http://www.prestospace.org`

[3] The national audiovisual archive is administered by the Institute of Sound&Vision and contains, among other things, all radio and television broadcasts of the Dutch public broadcasters.

[4] `http://www.willemfrederikhermans.nl/`

under uncontrolled conditions [6] the resemblance with shoe-box photo collections (i.e. little annotation and structure) may be acceptable. However, for potentially rich audio collections with a latent impact that is not limited to the individual who happened to do the recording, there is a serious need for disclosure technology.

The observation that audio indexing technology can contribute to the disclosure of spoken word archives has been made many times [7] and several initiatives have been undertaken to develop this technology for audio collections in the cultural heritage domain. A well-known example is the MALACH project that applies ASR and NLP for the disclosure of holocaust testimonies [2]. In this paper we give examples of similar audio indexing approaches applied to a number of Dutch audiovisual heritage collections.

3 AUDIO INDEXING

Audio indexing is a topic in a number of research areas. The various programs have in common that they aim at the automatic extraction of information from audio documents that can directly or indirectly be used for searching. The extracted information can be regarded as document features; each feature adds to the overall representation of a document. Speech related features play a major role. Traditionally these include the localization of the speech fragments (audio partitioning/segmentation), the speaker (identification and clustering) and the speech itself (speech recognition). More recently the extraction of emotional features in the audio signal (e.g., affect bursts such as laughter, cheering or words that express emotions) has been added to this list. Until now, emotional features have been primarily used to detect so called 'hot spots' in collections, but using emotional features for indexing a collection, represents a nice example of how historical data may be looked at in new ways.

3.1 Variance in performance

Recent years have shown that automatic speech recognition can successfully be deployed for equipping spoken-word collections with search functionality. This is especially the case in the broadcast news domain which is very general and makes data collection for system training relatively easy. For the broadcast news domain speech transcripts approximate the quality of manual transcripts for several languages. Spoken document retrieval in the American-English broadcast news (BN) domain was even declared "solved" with the NIST-sponsored TREC SDR track in 2000 [5]. In other domains than broadcast news, a similar recognition performance is usually harder to obtain due to a lack of domain-specific training data, in addition to a large variability in audio quality, speech characteristics and topics that are addressed. This applies to historical data in particular.

3.2 Cross-media browsing

Having produced a textual representation of audiovisual content it is a small step to relate it to textual representations of other audiovisual documents or text documents, enabling cross-media linking and browsing [9]. An application that demonstrates the added value of this option takes the textual representation of a broadcast news topic as a query to search for related newspaper articles in a newspaper database. Especially for scenarios involving professional users, such as historians, cross-media browsing has a great potential as it allows for relating material across multiple multimedia databases.

4 FULL TEXT TRANSCRIPTION

First explorations of spoken document retrieval outside the broadcast news domain made clear that for other domains there still is a lot to accomplish, especially with respect to the accuracy level of ASR. In the project ECHO[5], experiments with historical data from a number of national audiovisual archives learned that search technology based on ASR might easily collapse due to shockingly high word error rates caused by the typical characteristics of historical material, for example a wide variety in audio quality, background noise, overlapping speech, spontaneous speech, topics that are unknown beforehand, old-fashioned speech and dialect speech.

4.1 Language modeling

In ECHO, one of the focus points was to solve the mismatch of statistical language models based on contemporary text, with the old-fashioned language and unknown words in the task domain. However, in order to do this, we needed historical in-domain text data –preferably in large amounts, and at least some information on the topics of the particular documents. As the only, fairly historical text data available were copies of carbon copies of commentaries on news items, the only option was to use OCR for the available paper material. Due to the quality of the copies this approach did not yield any useful results. In recent years however, we saw that more and more historical text data is becoming digitally available, especially newspaper data (see e.g., the National Newspaper Archive in The Netherlands[6]), that could be of help for language model purposes for speech recognition in the cultural heritage domain.

Another approach towards more useful speech recognition transcripts in the ECHO project aimed at improving the language model for individual documents. Here we used either available metadata (short content description) or the results of an initial speech recognition run as an information retrieval query. This query was then fed into a search engine covering the web [1] or a database [12] with contemporary newspaper data. In a subsequent step, the returned ranked list of documents was used to create a topic specific language model for the document of focus. This procedure is visualized in Figure 1. Although this approach had a positive effect on Word Error Rates, it did not improve results enough to enable the successful application of search technology for this type of data. Clearly, the acoustic modeling part deserved special attention but given the wide variety in audio characteristics in the ECHO collection it was not feasible to make a start here.

4.2 Acoustic modeling

We focused on acoustic modeling with a more homogeneous oral history collection with (mostly) one speaker: lectures and interviews of the well-known Dutch novelist Willem Frederik Hermans (1921–1994). Although the performance of a broadcast news (BN) system in the oral history domain was expected to be poor, we used the tools and resources collected and developed for a BN system as a starting point.

The Willem Frederik Hermans (WFH) study revealed that simply deploying a broadcast news system for transcribing oral history data results in very high error rates. With (mostly) only one speaker in

[5] European Chronicles Online; IST-1999-11994

[6] Dutch National Newspaper Archive website:`http://hennekam.archieven.org/`

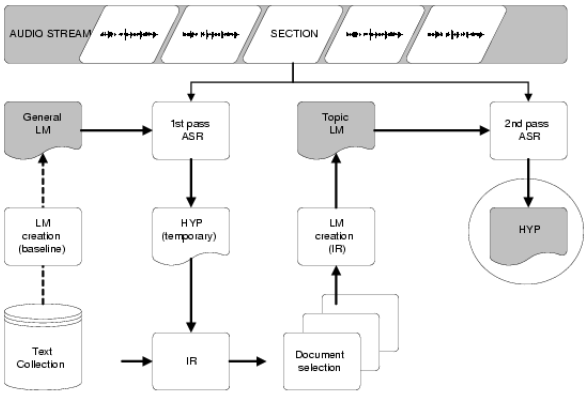

Figure 1. Results of an initial speech recognition run are used as a query to find on-topic documents. These documents are used to generate a new, topic specific, language model for a second speech recognition run.

the collection, performing acoustic speaker adaptation was an obvious step. The speaker adaptation employed, was a so-called 'supervised adaptation.' A segmentation into speakers (WFH and non-WFH) has been performed manually, and the acoustic models have been adapted to the speaker WFH using transcribed text. Both the segmentation and the speaker adaptation can in principle be performed automatically, or 'unsupervised.' The supervised experiments we performed, reducing the WER from 81.6 % to 66.7 %, for speaker WFH, give an impression of the maximum achievable performance increase. For more details, cf. [8].

4.3 Example data

For the domains we have seen until now, related audio and text sources that could be used for adapting the speech recognition components to these domain characteristics are usually only minimally available. As there are hardly any speech training databases with ancient or dialectic speech, successful application of speech recognition technology for audio indexing in the oral history domain is very difficult and requires additional measures. An important step is to collect as much related data sources as possible for tuning the system. A strategy to deal with the lack of acoustic training data is deploying (partly) unsupervised training strategies. This will at least be one of the topics in the recently started CHoral (Access to oral History) project which aims at disclosing the archives dating back to the early 1980's of the regional radio station Radio Rijnmond from the city of Rotterdam.

5 TEXT ALIGNMENT

In case collateral text data is available for an audio collection, it is worthwhile to investigate whether synchronization of text data with the audiovisual data using alignment techniques is an option [10]. Typical examples of useful collateral text are: subtitling files or teleprompter texts (e.g., in news shows), minutes of a meeting and written versions of lectures or speeches. The higher the resemblance of these speech transcripts with the actual speech the better, but a perfect match is not required. When the quality of the transcripts is low (frequent gaps, re-phrasing or compression of phrases) the best strategy is to do the synchronization in subsequent steps as depicted in Figure 2: a speech recognition system is used to generate a initial transcript of the speech, referred to as hypothesis. Next, the hypothesis is aligned on the word level to the speech transcript using dynamic programming. At positions where the hypothesis and the transcript match, so called 'anchors' are placed. As the speech recognition system also provided word timing information, these anchors can be used to segment the audio file. The audio segments and the words in the text segments can then be synchronized using Viterbi forced alignment, a procedure that is commonly applied during the acoustic training phase of a speech recognition system. In the Radio Oranje project the goal is to synchronize a collection of speeches addressed to the Dutch people by Queen Wilhelmina (1880–1962) during World War II with the text versions that have been written out. The aligned text version can be used either for subtitling or for searching within the collection.

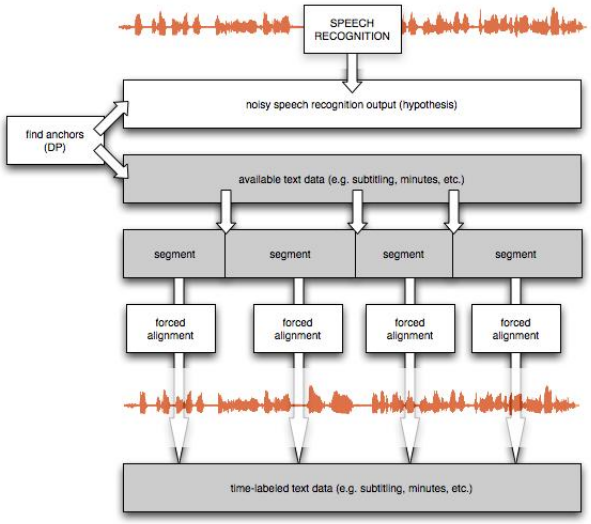

Figure 2. Alignment procedure: synchronization of text and audio in a number of steps.

6 VARIANCE IN INFORMATION NEEDS

By applying speech recognition technology, or one of the other audio analysis techniques mentioned in section 3, multiple levels of annotation will eventually become available. In addition, collection fragments can be linked to internal or external multimedia sources via cross-media linking (e.g., names that occur in an interview can be linked to a 'Who-is-Who' page). Interesting questions are (i) whether these automatically generated metadata streams and cross-links can indeed support the information need of different user types, (ii) whether they allow for new views on the data, and (iii) whether they require new ways of representing the data. Especially in the cultural heritage domain, the information need of the various user types may differ a lot. Archivists, information analysts, researchers from different fields, producers of new content and the general public all have their own questions and expectations regarding the collections and its metadata.

A survey [14] that investigated the attitude towards audio indexing technology of archivists administering a huge audiovisual archive,

showed that archivist are ambivalent towards the technology. On the one hand they acknowledge the potential added value and are interested in including all possible levels of metadata in a structured way. On the other hand, confronted for example with imperfect speech transcripts, archivist may become sceptical about the usefulness of the automatically generated metadata. Such contradictions should be dealt with carefully in order to have the technology accepted in the first place.

6.1 Researchers' point of view

In the CHoral project (disclosure of radio archives), the information needs of researchers from different fields using spoken-word archives will receive special attention. In research, questions can be asked that apply to any of the levels of information in the metadata. For example, through audio indexing historians can gain easy access to primary spoken sources such as eye witness reports. They may be interested both in the exact words that were spoken and in the speaker's profile (e.g., social background, emotion). Linguists, on the other hand, could study spontaneous speech and language use over the course of several decades. Their interest lies at the level of the speech signal and the linguistic structures it expresses, which can even be irrespective of meaning. Other types of researchers, such as information analysts, are more likely to require high-level and quantitative information on the collections. In general, researchers can be expected to be particularly interested in new insights emerging from combining a multitude of views on the data. Given the greatly varying information needs from different types of users, in CHoral the development of methodologies for the use of historical multimedia collections is a prerequisite for eventual successful deployment of the tools to be built.

On the basis of the various user groups' information demands, users are expected to differ not only in their information needs, but also in their needs for search options and data presentation. In general functionalities such as clustering, classification, extraction of headlines and proper names, time-lines, indexing and summarization can support efficient navigation and selection [4]. However, some presentation forms may either be more suitable for particular information requests, or may in itself provide a new view on the data. An example could be linking collection metadata to geographic information (e.g., by marking topics related to places on a map).

Finally, it is important to acknowledge that users of audiovisual archives could also contribute knowledge. Unknown parameters or features in the metadata of the audio archives may be added by (specialist) users, for example in a personalized peer-to-peer set-up that stimulates the exchange of content [13].

7 CONCLUSION

The aim of this paper was twofold. First, we discussed audio indexing methods and tools that have been implemented for the disclosure of Dutch audiovisual cultural heritage collections. It was explained that due to the increasing accessibility of historical text data, language models can be adapted to historical settings. Moreover, it was shown that the disclosure of homogeneous audio collections improves through adaptation of acoustic models. The challenge now is to facilitate access to heterogeneous collections from the cultural heritage domain. For such collections, extraction of speech transcripts and metadata calls for robust audio indexing technology that performs well irrespective of speaker, bandwidth or audio quality.

The second aim of this paper was to discuss the greatly varying information needs for different types of users, evoked by the diverse nature of historical audio collections. It was argued that the automatically generated stream of metadata and multimedia cross-links on the one hand, and the different types of users on the other hand, requires the development of methodologies for the use of historical multimedia collections in this context. Moreover, we stressed the importance of adequate navigation and selection tools that in the ideal case may unfold new information.

ACKNOWLEDGEMENTS

The research reported here is partly funded by the research program MultimediaN (http://www.multimedian.nl) and the NWO-CATCH project CHoral (http://hmi.ewi.utwente.nl/projects/choral). MultimediaN is sponsored by the Dutch government under contract BSIK 03031.

References

[1] I. Bulyko, M. Ostendorf, and A. Stolcke, 'Getting More Mileage from Web Text Sources for Conversational Speech Language Modeling using Class-Dependent Mixtures', in *Proceedings of the joint conference of Human Language Technology and North American Chapter of Association for Computational Linguistics (HLT-NAACL)*, (2003).

[2] W. Byrne, D. Doermann, and M. Franz, 'Automatic Recognition of Spontaneous Speech for Access to Multilingual Oral History Archives', *IEEE Transactions on Speech and Audio Processing, Special Issue on Spontaneous Speech Processing*, (July 2004).

[3] K. W. Church, 'Speech and Language Processing: Where Have We Been and Where Are We Going?', in *Eurospeech-2003*, Genève, Switzerland, (September 2003).

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

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

[6] J. Gemmell, G. Bell, R. Lueder, S. Drucker, and C.Wong, 'Mylifebits: fulfilling the memex vision', in *ACM Multimedia*, pp. 235–238, (2002).

[7] Jerry Goldman, Steve Renals, Steven Bird, Franciska de Jong, Marcello Federico, Carl Fleischhauer, Mark Kornbluh, Lori Lamel, Douglas Oard, Claire Stewart, and Richard Wright, 'Transforming Access to the Spoken Word', *International Journal on Digital Libraries*, **5**(4), 287–298, (2005).

[8] M.A.H. Huijbregts, R.J.F. Ordelman, and F.M.G. de Jong, 'A Spoken Document Retrieval Application in the Oral History Domain', in *Proceedings of 10th international conference Speech and Computer (SPECOM 2005)*, 2, pp. 699–702, (2005).

[9] Jeroen Morang, Roeland Ordelman, Franciska de Jong, and Arjan van Hessen, 'InfoLink: analysis of Dutch broadcast news and cross-media browsing', in *Proceedings of ICME 2005*, Amsterdam, (September 2005).

[10] Pedro J. Moreno, Chris Joerg, Jean-Manuel Van Thong, and Oren Glickman, 'A Recursive Algorithm for the Forced Alignment of Very Long Audio Segments', in *Proceedings of the 5th International Conference on Spoken Language Processing (ICSLP'98)*, Sydney, Australia, (1998).

[11] R.J.F. Ordelman, A.J. van Hessen, F.M.G. de Jong, and D.A. van Leeuwen, 'Speech Recognition for Dutch Spoken Document Retrieval', in *Content-Based Multimedia Indexing (CBMI)*, Brescia, (2001).

[12] Roeland Ordelman, *Dutch Speech Recognition in Multimedia Information Retrieval*, Ph.D. dissertation, University of Twente, The Netherlands, October 2003.

[13] Jun Wang, Johan Pouwelse, Jenneke Fokker, and Marcel J.T. Reinders, 'Personalization of a Peer-to-Peer Television System', in *Proc. of European Conference on Interactive Television, EuroITV 2006*, (2006).

[14] E. Zuurbier, 'Onderzoek naar de haalbaarheid van Spoken Document Retrieval', Master's thesis, University of Twente, (2004).