Evaluation of Spoken Document Retrieval for Historic Speech Collections

W. Heeren, F. de Jong, L. van der Werff, M. Huijbregts and R. Ordelman

Human Media Interaction

Department of Electrical Engineering, Mathematics and Computer Science

University of Twente

The Netherlands

E-mail: {w.f.l.heeren, fdejong, l.b.vanderwerff, m.a.h.huijbregts, ordelman}@ewi.utwente.nl

Abstract

The re-use of spoken word audio collections maintained by audiovisual archives is severely hindered by their generally limited access. The CHoral project, which is part of the CATCH program funded by the Dutch Research Council, aims to provide users of speech archives with *online*, instead of on-location, access to relevant *fragments*, instead of full documents. To meet this goal, a spoken document retrieval framework is being developed. In this paper the evaluation efforts undertaken so far to assess and improve various aspects of the framework are presented. These efforts include (i) evaluation of the automatically generated textual representations of the spoken word documents that enable word-based search, (ii) the development of measures to estimate the quality of the textual representations for use in information retrieval, and (iii) studies to establish the potential user groups of the to-be-developed technology, and the first versions of the user interface supporting online access to spoken word collections.

1. Introduction

Audiovisual archives in the cultural heritage (CH) domain maintain large numbers of spoken word audio collections with great potential for e.g., research and educational purposes. A recently published report gives estimates of 9.4 million hours of audio heritage across 288 respondents, and 10.5 million hours of video for 277 respondents across Europe alone (Klijn and de Lusenet, 2008). The actual (re-)use of such collections, however, is severely hindered by their generally limited access. Firstly, whereas catalogs are in an increasing number of cases searchable online, only audiovisual item descriptions, but not the items themselves (or documents as we will call them here) can be retrieved. Secondly, once users gain access to the documents – by visiting the archive and requesting a copy – their exploration remains cumbersome as descriptions are often insufficiently specific and playing documents can be time-costly.

In the CHoral project (http://hmi.ewi.utwente.nl/choral), which is part of the CATCH program funded by the Dutch Research Council (http://www.nwo.nl/catch), we aim to provide users of speech archives with *online*, instead of on-location, access to relevant *fragments*, instead of full documents. To meet the goal of providing end users with fast access to relevant spoken word documents we start out with digitized speech collections. For index generation, techniques from automatic speech recognition (ASR) and audio processing are used. The observation that audio indexing technology can contribute to the disclosure of spoken word archives has been made many times, see (Goldman et al., 2005), and several initiatives have been undertaken to develop this technology for such audio collections, see (Byrne et al., 2004; Hansen et al., 2005). Still there are only few examples of online access to spoken audio from the CH domain. For retrieval of spoken documents we aim to employ a combination of information retrieval and word spotting techniques. Finally, to allow end users access to the audio, a user interface is being developed that supports document search and playback.

The investigation and development of a spoken document retrieval (SDR) system involves intrinsic evaluation of the system components, index generation and retrieval performance, but also usability evaluations of the search interface's functionality. In this paper we present the evaluation efforts that have been undertaken so far to assess and improve various aspects of an SDR environment for access to spoken word collections from Dutch audiovisual heritage. Section 2 deals with the quality of automatically generated metadata, section 3 presents issues in using automatic speech recognition output for information retrieval purposes, and section 4 summarizes the efforts undertaken to develop the user interface of our SDR framework.

2. Evaluation of Automatically Generated Textual Representations

The technology needed for generating a textual representation of spoken audio depends on the type and amount of metadata that are already available for a collection. If manual transcripts or elaborate summaries are available, alignment of text and speech can be employed, see e.g., (Christel et al., 2006; Munteanu et al., 2006). If transcripts are not available, a textual representation can be generated through automatic speech recognition, see e.g., (Goldman et al., 2005, Hansen et al., 2005). At the University of Twente (UT) both methods have been applied to Dutch audiovisual collections, and the results are presented in the following subsections.

2.1 Quality of Automatically Generated Transcripts

At the University of Twente, a system has been developed for automatically generating transcripts for Dutch broadcast news recordings. This system typically generates textual transcripts with an error rate of 20 to 30% (Huijbregts et al., 2007a) and these transcripts are of high enough quality to be applied in SDR applications. As a comparison, error rates reported for English broadcast news lie in the range of 10-20% (see e.g., Pallett et al., 1998).

Obtaining high quality, automatically generated transcripts for types of data other than broadcast news is not easy. For example, automatically generated transcripts were created for a number of lectures and interviews of the Dutch novelist Willem Frederik Hermans (1921-1994). These recordings form a relatively homogeneous collection, since most speech was produced by one speaker. Simply deploying the broadcast news ASR system for transcribing this, however, resulted in error rates over 80% (Huijbregts et al., 2005). This is because this type of data contains more spontaneous speech, partly from non-professional speakers, recorded under less optimal circumstances (e.g., background noise, reverberation). Since one speaker dominated the collection, acoustic speaker adaptation was applied (so-called 'supervised adaptation'), and the word error rate was reduced from 81.6% to 66.7% for the dominant speaker.

Heterogeneous collections with varying audio conditions (e.g., including music and bandwidth changes) and multiple speakers are even harder to process automatically. These kinds of collections typically require more pre-processing, such as Speech Activity Detection and Speaker Clustering. Especially when the collection contains audible non-speech such as background noise or musical fragments instead of just speech and silence, Speech Activity Detection (SAD) is a challenging step. Before performing automatic speech recognition, the system should filter out all non-speech fragments. Performing ASR on these fragments would introduce high amounts of noise in the resulting transcripts. On the other hand, the system should be careful not to remove speech segments instead of non-speech segments as this mistake will lead to missing fragments in the transcripts.

UT measured the quality of SAD on the TRECVID 2007 collection, a subset of the Academia collection of the Netherlands Institute for Sound and Vision, one of Europe's largest audiovisual archives. This data set contains speech and audible non-speech from various unknown sources. Without optimization, the error rate of speech activity detection was 18.3%, but it reduced to 11.4% after optimization (Huijbregts et al., 2007b).

Another important pre-processing step is speaker clustering or speaker diarization, see e.g., (Tranter and Reynolds, 2006). Speaker diarization is the process of automatically clustering all speech fragments of each individual speaker in a recording. Once it is known which speech fragments are pronounced by a particular speaker, it is possible to automatically optimize the ASR system for this speaker. For example, the system can determine the pitch of the speaker's voice and adapt its statistical models in order to increase performance, e.g., Vocal Tract Length Normalization. On the TRECVID 2007 collection this technique improved the transcripts with 4.4% word error rate, resulting in an optimized word error rate of 64.0% (Huijbregts et al., 2007a). Note, however, that the word error rate can easily exceed 100% due to its definition, i.e. the sum of the numbers of substitutions, insertions and deletions divided by the number of words in the reference transcript.

Although it is possible to apply SDR using the automatically generated transcripts of both the Willem Frederik Hermans collection as the TRECVID 2007 collection, the relatively high word error rates of the transcripts will introduce errors for each query result. Ongoing research at the University of Twente is focussed on increasing ASR performance for heterogeneous collections, see also section 3.

2.2 Alignment Quality

For the collection of radio speeches that Queen Wilhelmina (1880--1962) addressed to the Dutch people during World War II, the so-called 'Radio Oranje' collection, both the recordings and their 1940s transcripts are available in digitized form. For generation of the word-level index UT used alignment, a technique that matches a speech sound representation of the text against the sequence of speech sounds detected in the audio, see (van der Werff et al., 2007) for more detail.

Optimal performance on this collection was obtained using speaker-dependent, monophone acoustic models. The performance evaluation showed that over 90% of all word boundaries were found within 100 ms of the reference, which is more than sufficient given the application of indexing word positions within a spoken document. Moreover, our results show that the method performs adequately, even under the adverse acoustic conditions of historic recordings. Alignment therefore proved a relatively straightforward method to process those collections from the cultural heritage domain for which manual transcripts are available, such as interview collections and speeches.

3. Evaluating ASR for IR

In 2000 automatic speech recognition for spoken document retrieval was declared 'solved' for the broadcast news domain. Many spoken word collections from the cultural heritage domain, however, are not in this domain and give word error rates (WERs) that are relatively high compared to the WERs of 10-20% obtained on broadcast news. Our results on Dutch audiovisual collections from the cultural heritage domain

are somewhat lower than results reported for English collections that lie in the range of 30-60% on e.g., a collection of historic speeches, news broadcasts and debates (Hansen et al., 2005), and the MALACH collection of interviews with Shoah survivors (Byrne et al. 2004).

In a retrieval context, however, word error rate is a flawed optimization criterion because (i) it is only defined on a (literal) transcription and cannot be calculated on other types of speech recognition output such as n-best lists or lattices, and (ii) information retrieval performance depends not only on the *amount* but also on the *type* of errors: recognition errors for function words have less impact than for content words.

Performance of information retrieval systems is typically measured using the mean average precision, a score that is calculated from the amount of relevant documents found for some set of queries, the amount of non-relevant documents that is produced, and their ranking. Calculating such a score can only be done using an evaluation platform that contains ground-truth (i.e. human) relevance judgments for a set of queries and documents. In practice, information retrieval evaluation platforms are only readily available for a limited amount of collections. When optimizing the speech recognition component of an SDR system for a collection for which no matching evaluation platform can be found, one needs to develop a new evaluation set, which requires a prohibitive amount of work.

Instead, UT proposed three new performance measures for speech recognition in an SDR context. Instead of simply counting the number of errors, each error is weighted using the *same weight* that is also used for calculating the rank of documents, *tf.idf* (Sparck Jones et al., 2000). Since users are expected to enter query terms that are discriminative towards the documents they are seeking and term weights are optimized towards ranking documents for expected relevance, the use of this weight to determine the relative importance of a term in practical information retrieval settings seems plausible.

These measurements were calculated on a test set with noisy spontaneous Dutch speech (van der Werff & Heeren, 2007). Results were encouraging and in line with expectations. Removing stopwords from the transcription and performing stemming, showed that these new measurements are relatively robust towards these standard information retrieval techniques, whereas traditional quantitative measures such as word error rate and out-of-vocabulary rate were not. Moreover, a direct comparison between one of the new measures and the traditional word error rate on a set of broadcast news data from the TREC SDR benchmarks showed that the new measure was better at predicting changes in retrieval performance.

4. Evaluation of the User Interface

From a user perspective, the goal is to get fast access to those spoken word documents one is interested in. In the CHoral project, we have so far addressed two research issues concerning users. We first performed a general requirements analysis with keepers of Dutch, spoken word collections from the cultural heritage domain to determine who the users of these collections actually are. This revealed that users are mostly researchers/students (e.g., oral historians, sociologists) and content producers. Comparable results were recently reported in a Europe-wide study on audiovisual heritage collections (Klijn & de Lusenet, 2008, p. 15). We therefore target further research into user interfaces for our SDR framework at these user groups.

Second, in our pursuit of supporting the interaction between users and the audio, we developed tools aimed at improved navigation and listening support during audio playback for the 'Radio Oranje' collection. These tools were built using the word-level index for visualization of the spoken content: karaoke-style subtitling and word position information in timelines for navigation control (see Figure 1). Subtitling highlights the word being spoken by changing its color, and shows query terms in bold. The interactive timeline gives an overview of the entire speech (upper bar) as well as a zoomed-in view of the interval being played (lower bar), and shows the exact locations of query terms and sentence boundaries through colored indicators. The complete demonstrator system – which is mostly in Dutch – can be found at http://hmi.ewi.utwente.nl/choral/demo.

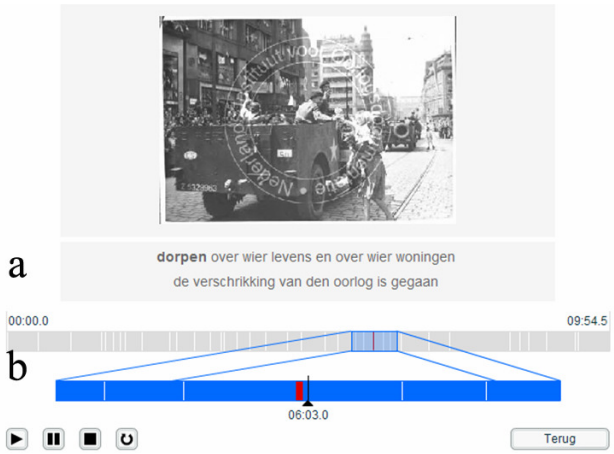

Figure 1: Playback page for the Radio Oranje collection including subtitling (a) and annotated timelines (b). While the audio plays, a slide show of semantically related photos is presented.

Earlier user evaluations of speech browsers that exploit time-stamped metadata have shown that users only benefit from on-screen transcripts when these are of high-quality, i.e. contain only few recognition errors (e.g., Munteanu et al., 2006), and that visual feedback on the positions of relevant sections in the audio is helpful (e.g.,

Whittaker et al., 1999).

The functionality of showing query term locations for easier navigation, and of subtitling for better perception of the low-quality, old-fashioned speech was evaluated with a group of ten potential users, i.e. students from the departments of history and linguistics. They may be expected to search speech fragments for different types of analyses. The evaluation showed that searchers were supported by high-quality subtitling in the correct understanding of old-fashioned speech, but it did not help their short term memory of the content. This was probably due to the fact that cognitive load was increased by presenting the spoken words both visually and auditorily at the same time, see e.g., (Kalyuga et al., 1999). In line with expectations we found that content navigation was supported by visual information on the exact positions of relevant words. If available, this information should therefore always be presented to the searcher.

5. Conclusion

In this paper we presented the various types of evaluation taken up by UT, and the CHoral project in particular, to develop an SDR environment for access to spoken word collections. We focused on access functionality for speech collections from the Dutch cultural heritage domain.

Firstly, automatic speech recognition performance on Dutch audiovisual collections from the cultural heritage domain was reported. Performance was somewhat lower than for comparable English collections. For use in information retrieval, however, word error rates in the range of 30-40% are deemed sufficient (Garofolo et al., 2000). Since recognition performance on spoken documents from the heritage domain is generally lower, improvement is needed. On the other hand, an automatic index may in some cases – despite its errors – be more helpful than no index at all.

Secondly, we have concluded that the traditional word error rate measure for evaluating speech recognition quality seems unfit for evaluating speech recognition performance in the context of information retrieval. Therefore, a number of alternative measures were proposed that have shown promising behaviour. These measures will be studied further in future work.

Finally, we have narrowed down our potential user groups, and are taking steps to further develop user interfaces that facilitate access to content from historic spoken word collections. We are currently applying the lessons learnt from the 'Radio Oranje' project to user interfaces for access to other types of spoken word collections from the heritage domain, such as interview collections.

6. Acknowledgements

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

7. References

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

Christel, M.G., Richardson J., and Wactlar, H.D. (2006). Facilitating access to large digital oral history archives through Informedia technologies. In *Proceedings of JCDL* '06, pp. 194--195.

Garofolo, J.S., Auzanne, C.G.P., and Voorhees, E.M. (2000). The TREC Spoken Document Retrieval Track: A success story. In *Proceedings of RIAO 2000*, pp. 1--20.

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

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

Huijbregts, M., Ordelman, R., and de Jong, F. A. (2005). Spoken Document Retrieval Application in the Oral History Domain, in *Proceedings of 10th SPECOM,* pp. 699--702.

Huijbregts, M., Ordelman, R., and de Jong, F. (2007a). Annotation of Heterogeneous multimedia content using automatic speech recognition. In *Proceedings of SAMT 2007*, pp. 78--90.

Huijbregts, M., Wooters, C. and Ordelman, R. (2007b) Filtering the Unknown: Speech Activity Detection in Heterogeneous Video Collections. In *Proceedings of Interspeech 2007*, pp. 2925--2928.

Kalyuga, S, Chandler, P., and Sweller, J. (1999) Managing split-attention and redundancy in multimedia instruction, *Applied Cognitive Psychology*, 13, pp. 351--371.

Klijn, E. and de Lusenet, Y. (2008). *Tracking the reel world. A survey of audiovisual collections in Europe.* Amsterdam: European Commission on Preservation and Access.

Munteanu, C., Baecker, R., Penn, G., Toms, E., and James, D. (2006). The effect of speech recognition accuracy rates on the usefulness and usability of webcast archives. In *Proceedings of CHI 2006*, pp. 493--502.

Pallett, D., Fiscus, J., Garofolo, J., Martin, A., and Mark Przybocki. (1998). 1998 Broadcast News Benchmark Test Results: English and Non-English Word Error Rate Performance Measures. In *Proceedings of 1999 DARPA Broadcast News Workshop*.

Sparck Jones, K., Walker, S. and Robertson, S. (2000). A probabilistic model of information retrieval: development and comparative experiments (part 1 and 2). *Information Processing & Management*, 36(6), pp. 779--840.

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

Van der Werff, L., Heeren, W., Ordelman, R, and De Jong, F. (2007). Radio Oranje: Enhanced access to a historical spoken word collection. In *Proceedings of. CLIN 2007*, pp. 207--218.

Van der Werff, L., and Heeren, W. (2007). Evaluating ASR Output for Information Retrieval. In *Proceedings of ACM SIGIR SSCS Workshop*, pp. 7--14.

Whittaker, S., Hirschberg, J., Choi, J., Hindle, D., Pereira, F.C.N., and Singhal, A. (1999). SCAN: Designing and Evaluating User Interfaces to Support Retrieval From Speech Archives. In *Proceedings of ACM SIGIR 99*, pp. 26--33.