Robust Audio Indexing for Dutch Spoken-word Collections

Roeland Ordelman, Franciska de Jong and Marijn Huijbregts
Department of EEMCS, Human Media Interaction
University of Twente, Enschede, The Netherlands

David van Leeuwen
TNO Defence, Security and Safety
Soesterberg, the Netherlands

***Abstract*— Whereas the growth of storage capacity is in accordance with widely acknowledged predictions, the possibilities to index and access the archives created is lagging behind. This is especially the case in the oral history domain and much of the rich content in these collections runs the risk to remain inaccessible for lack of robust search technologies. This paper addresses the history and development of robust audio indexing technology for searching Dutch spoken-word collections and compares Dutch audio indexing in the well-studied broadcast news domain with an oral-history case-study. It is concluded that despite significant advances in Dutch audio indexing technology and demonstrated applicability in several domains, further research is indispensable for successful automatic disclosure of spoken-word collections.**

I. Introduction

The number of digital spoken-word collections is growing rapidly. Due to the ever declining costs of recording audio and video, and due to improved preservation technology huge data sets are created, both by professionals at various types of organisations and non-professionals at home and underway. Partly because of initiatives for retrospective digitisation, data-growth is also a trend in historical archives. These archives deserve special attention because they represent cultural heritage: a type of content which is rich in terms of cultural value, but has a less obvious economical value. Spoken-word archives belong to the domain of what is often called oral history: recordings of spoken interviews and testimonies on diverging topics such as retrospective narratives, eye witness reports, historical site descriptions, and modern variants such as 'Podcasts' and so-called amateur (audio/video) news[1].

Where the growth of storage capacity is in accordance with widely acknowledged predictions, the possibilities to index and access the archives created is lagging behind though [4]. Individuals and many organisations, often do not have the resources to apply even some basic form of archiving. Spoken word collections may become the stepchild of an archive—minimally managed, poorly preserved, and hardly accessible. The potentially rich content in these collections risk to remain inaccessible.

For 'MyLifeBits' chronicles collected by non-professionals under uncontrolled conditions [7] the resemblance with shoe-box photo collections (i.e., little annotation and structure) may be acceptable. But for audio collections with a potential impact that is not limited to the individual who happened to do the recording, there is a serious need for disclosure technology. Tools for presenting and browsing such collections and to search for fragments could support the information need of various different types of users, including archivists, information analysts, researchers, producers of new content, general public, etc.

The observation that audio mining technology can contribute to the disclosure of spoken word archives has been made many times [8], and several initiatives have been undertaken to develop this technology for audio collections in the cultural heritage domain. Worthwhile mentioning are projects such as ECHO (European CHronicles Online), that focused on the development of speech recognition for historical film archives for a number of European languages [1], and MALACH, applying ASR and NLP for the disclosure of holocaust testimonies [3]. But the high expenses required to process historical material in combination with the expected limited financial return on investment have prohibited real successes. A break through for the application of audio mining outside standard domains (typically: English news) is still pending.

This paper addresses the history and development of robust audio indexing technology for Dutch spoken-word collections in various domains, including radio and television broadcasts, govern-

[1] 'Podcasts' are home-brew radio shows covering personal interest items and can be viewed as the audio variant of a 'blog' which is basically a journal that is made available on the web. Amateur news is news compiled by amateurs and broadcasted via the web.

mental proceedings, historical archives, lectures and meetings. The application of audio indexing technology for the oral history domain will receive special attention. Audio indexing involves topics such as audio partitioning, keyword spotting, speech recognition, speaker identification and information extraction. After introducing the technology in section II, this paper compares Dutch audio indexing in the well-studied broadcast news domain (section III) with an oral-history case-study, consisting of a collection of spoken audio material from the Dutch novelist Willem Frederik Hermans (section IV). It is concluded in section V that despite significant advances in Dutch audio indexing technology and demonstrated applicability in several domains, further research is indispensable for successful automatic disclosure of spoken-word collections.

II. Audio mining overview

Audio mining involves a number of research areas that 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. Low-level and high-level features are distinguished. Low-level features are for example acoustic features such as note duration and pitch in musical audio mining [9], or bandwidth and spectral features in speech mining. Instrument classifications, speech transcripts or even textual summaries are examples of high-level features.

The main focus in this paper is on the extraction of features that are relevant for generating higher-level *speech* related features. Traditionally these include the localisation of the speech fragments (audio partitioning/segmentation), the speaker (identification and clustering) and the speech itself (speech recognition, information extraction). More recently the extraction of emotional features in the audio signal (e.g., affect bursts such as laughter or words that express emotions), for example to detect so called 'hot spots' in collections, has been added to this list.

Recent years have shown large improvements in the performance of automatic speech recognition (ASR) systems, and speech transcripts can now be generated at nearly the same quality as manual transcripts, at least for well-studied domains such as the broadcast news domain [5]. As speech recognition systems label recognised words with exact time information as a standard accessory, the time-labelled speech transcripts can directly be used to search within audio documents. Parts of a document that are relevant for a query can be accessed by linking to the time-label of relevant words.

Before we give an overview of the most frequently applied speech recognition techniques in audio mining in section II-B, fundamental auxiliary techniques for a successful application of speech technology, audio segmentation and audio source labelling, are discussed in brief in section II-A. In section II-C, important issues for the presentation of search results for audio/video collections will be addressed.

A. Segmentation and audio source labelling

Although time-labelled speech transcripts can directly be used to search within audio documents, in practice longer audio documents are often pre-structured: segmented according to a particular condition such as speaker-turns, silence, or even topic, into homogeneous sub-documents that can be accessed individually. This is convenient, as scrolling through a large unstructured audio or video document to identify interesting parts can be cumbersome. Audio segmentation can be advantageous from a speech recognition point of view as well, as it allows for segment based adaptation of the recognition models as will be discussed below. A frequently applied adaptation scheme is based on speaker identity.

Using a fixed overlapping time window, or fixed number of words to segment an audio stream is a simple but in cases very effective segmentation approach that does not rely on special segmentation tools. When the window and overlap ranges are chosen well, it can provide a document structure that can already usefully be deployed for certain retrieval tasks, such as word-spotting. But a segmentation based on audio features is much more informative and helpful both from a retrieval and speech recognition point of view. With a segmentation according to speaker for example, retrieval results can be structured and presented according to speaker identity. In addition speaker dependent modelling schemes can be applied in order to improve speech recognition performance. Useful segmentation cues are in general provided by techniques that aim at the labelling of the source of audio data (e.g., acoustic environment, bandwidth, speaker, gender), often referred to as 'diarisation' or 'non-lexical information generation' [18].

From a retrieval point of view, topic-based audio segmentation would be an obvious choice. Such a

segmentation would resemble the topic structure in textual documents. In addition, topic segmentation allows for the selection of topic specific language models and re-scoring the speech recognition results with these topic-specific models. Topic segmentation can for example be based on word frequency and co-occurrence information (see e.g., [15]).

B. Speech recognition

Information retrieval research that uses the spoken audio parts of documents for retrieval is commonly referred to as spoken document retrieval (SDR) or alternatively, speech-based retrieval. 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 domains such as the broadcast news domain which is very general and makes data collection for system training relatively easy. For the broadcast news domain speech transcripts therefore approximate the quality of manual transcripts for several languages and 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 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 the oral-history domain in particular.

The most obvious approach in spoken document retrieval is the word-by-word translation of the encountered speech using a large vocabulary continuous speech recognition (LVCSR) system. Having generated a textual representation (full text transcription) of an audio or video document, the document can be searched as if it were a text document. As mentioned above, the time-labels provided by the speech recognition system and the segmentation of a large document into sub-documents, provide additional means for structuring the document.

For some applications, applying a word-spotting approach can be beneficial as an alternative or auxiliary tool. This is especially the case when the mismatch between speech recognition vocabulary and domain vocabulary is hard to model and tends to produce many out-of-vocabulary words. That is, when in a certain domain very specific words are used frequently (e.g., jargon, names), and there are no resources to rely on for the prediction of these words (e.g., related text documents). As a consequence, the speech recognition vocabulary will not 'know' the specific words and will not be able to recognise them. This implies that when these words are used as search terms the retrieval system will not find the documents that contain this term. In addition, as the 'unknown' word is replaced by the speech recognition system for a substitute word the system knows (and is acoustically similar to the unknown word), the wrong documents will be produced by the retrieval system when a query contains exactly these substitute words.

A keyword spotter searches the audio material for single keywords. This can be done 'on-line', immediately after a user has posed a query, but this requires that the archive is not too large as it takes some time to do the on-line processing. In an 'off-line' approach, the word spotter searches for the appearance of a relatively small set of keywords in the documents in advance. The words that are found are then used to label the document. An acoustic model is used to recognise phones and a, usually small, vocabulary of keywords with phonetic transcriptions provides the link to the keywords.

In specific cases, speech-based retrieval can do without speech recognition technology. In the ORL Medusa multimedia retrieval system [2] for example, teletext subtitles were exploited for purely speech-based retrieval. Subtitles contain a nearly complete transcription of the words spoken in the material and provide an excellent information source for indexing when aligned (labelled with time-codes) to the actual speech in the collection. A comparable strategy was used for disclosing meetings of the Dutch Government. Here, instead of subtitles, the official stenographic minutes ("Handelingen van de Tweede Kamer") that closely follow the discourse of the meeting, have been aligned to the speech. When available, deploying such external information sources can be a practical solution for certain task domains. If necessary, speech recognition technology may still be called upon to time-align the external transcripts with the speech in the documents.

C. Presentation of search results

The presentation of search results for a given collection in spoken document retrieval deserves special attention. People are generally very good at browsing globally through text documents, and often recognise at a single glance whether a document is relevant for their specific information need or not. For audio/visual documents, browsing glob-

ally is not an option; one usually has to play the entire document to be able to assess its relevance. Providing fast-play functionality can sometimes be helpful to speed up the process.

By structuring an audio/visual document via the deployment of segmentation techniques for the creation of subdocuments, presentation functionality can be improved considerably. A speaker based segmentation for example, makes it possible at least to skip certain speakers. In television broadcast news shows, anchor people pause longer between topics then within topics. Here, a simple pause based segmentation thus provides a helpful topic structure for browsing.

Providing the automatically generated speech transcripts with the (sub)documents for rapid relevance assessment can be an option. However, the raw speech recognition transcripts contains errors , and, as punctuation is absent, can be hard to read. In addition, distracted by the speech recognition errors in the transcripts, a user may be prejudiced concerning the relevance of results. Research aiming at the generation of readable transcriptions incorporating capitalisation, punctuation, and speaker markers is currently investigated among others in the DARPA-EARS Rich Transcription project[2]. The speech transcripts can still be used for content-based topic segmentation (e.g., in less structured domains) and automatic summary generation, provided that the speech recognition transcripts are reasonably accurate.

A relatively novel approach is the linking of collection fragments to external sources of information [10]. For example, fragments of digital meeting recordings can be linked to meetings agendas or policy documents pertaining to specific topics of a meeting. Such documents can be considered as part of the same archive. However, for a meeting archive links to external sources may also be useful, such as newspaper articles, an archive of broadcast news items and documentaries, and a 'Who-is-Who' of politicians. Given the potential size of such external sources, and the variety of perspectives that can be taken on them, automatically generated links could be very beneficial. As the external information sources can have different media formats (text, audio, video) this type of disclosure is referred to as cross-media disclosure. (For work on cross-media disclosure for meeting recordings in the IST-project AMI, cf. [14].)

Another strategy to enhance the presentation of search results for audio/video collections is making use of visualisations. For example, the entire audio document can be represented by a bar and fragments in the audio document that match a query are then represented within the bar as red markers. These markers may vary in length or colour-depth to represent the degree of relevance of the particular fragments. By pointing at the markers with the mouse, the audio fragment can be listened to.

[2]The Effective Affordable Reusable Speech-To-Text (EARS) program: `http://www.darpa.mil/ipto/programs/ears`

III. Dutch Spoken Document Retrieval

Spoken document retrieval research for Dutch first targeted at the development of a Dutch LVCSR system for the broadcast news domain. Broadcast news (BN) data has been extensively used as a benchmark domain for both international speech recognition research and SDR research. The domain is well-structured and a lot of resources exist that can be explored for the development of a broadcast news recognition system (e.g., newspaper data for language modelling, auto-cues for acoustic modelling). In the next section, the Dutch broadcast news speech recognition system is described and the results of a Dutch broadcast news retrieval experiment are discussed.

A. Broadcast news system

In [11] research on Dutch spoken document retrieval in the broadcast news domain is described in detail. In [11], the *ABBOT* speech recognition system [16] was used for the generation of speech transcripts. Recently the Dutch models have been ported to a new speech recognition system (referred to as UT-BN2005 system), that is based on a recogniser developed at the University of Colorado (CSLR) which has been made available for research purposes. Its acoustic models are decision-tree state-clustered Hidden Markov Models [12] and the broadcast news specific models are trained using twenty-two hours of broadcast news recordings from the Spoken Dutch Corpus ([17]). It uses a large vocabulary of 65 K (65 thousand) words and a statistical trigram language model derived from a Dutch newspaper corpus of some 400 M (400 million) words (referred to as TWente News Corpus). On a broadcast news test set consisting of 4 hours of broadcast news material from the Spoken Dutch Corpus, the UT-BN2005 obtained a word error rate (WER) of 30 %. Although at TREC-8, BN systems for American-English produced error rates below 20 %, this figure can be considered an adequate baseline given that it concerns a relatively simple (unadapted, single-pass) system trained on

a medium size acoustic training data set and with a standard 65 K newspaper vocabulary and language model.

To investigate the applicability of the developed Dutch speech recognition system in a retrieval task, a known-item retrieval task that simulates a user seeking one particular document [19], was performed. We used a set of 18 television news broadcasts (*NOS Acht uur journaal*) that were segmented manually: 180 stories with a mean length of 257 words. Introductions and weather reports were excluded. Story topics were generated by students who were instructed to create topic "titles" that in a few words (with a maximum of ten words) give a reasonable impression of the contents of the story. These titles were further interpreted as queries aiming at the retrieval of the respective stories. The titles were used as queries for retrieval given the following evaluation modes: using document representations that are based upon perfect, human-transcribed reference, using document representations based upon a speech recognition system producing a relatively large number of errors (word error rate of 50 %) and representations based upon a relatively well performing speech recognition system (word error rate of 30 %).

Table I shows the results of the known-item retrieval task. Using the reference transcript as document representation gave the best retrieval performance in terms of found documents. Using high quality speech recognition produced comparable results: only one document less was found (10 instead of 9) and the mean rank when found was even slightly better compared to the one obtained in the reference condition. As could be expected, deploying a low quality speech recognition system significantly worsens retrieval performance. Almost a quarter of the documents could not be found, on average, the target stories were retrieved almost one rank lower, and the mean reciprocal rank decreased almost 35 % relative compared to the high quality speech recognition condition.

document representation	MRWF	not found
Reference	2.0778	9 (5 %)
ASR with low WER	1.9278	10 (5.6 %)
ASR with high WER	2.8556	43 (23.9 %)

TABLE I

Results of the known-item retrieval task: mean rank when found (MRWF) and number of documents not found.

This retrieval experiment showed that for spoken retrieval in the broadcast news domain, a speech recognition accuracy of some 70 % produces similar retrieval results as manually generated speech transcripts. When recognition accuracy decreases to 50 % however, retrieval performance drops considerably. This is in accordance with figures seen in the TREC SDR evaluations and in a Dutch spoken document retrieval simulation study as described in [20].

IV. Willem Frederik Hermans Case Study

First explorations outside the broadcast news domain made clear that for other domains there is still a lot to accomplish, especially for the accuracy level of ASR. Experiments with the ECHO historical archive collection 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, dialect speech.

In the case study that will be described below, the target is on a slightly more heterogeneous oral history collection with (almost) only one speaker: lectures and interviews of the well-known Dutch novelist Willem Frederik Hermans (1921–1994). The collection can be searched via the Willem Frederik Hermans web-portal[3].

Although the performance of a BN system in the oral history domain was expected to be poor, we used the tools and resources collected and developed for a broadcast news (BN) system as a starting point. As similar systems are available in many labs, the conversion of the BN system and tuning to a collection from the oral history domain might be a case of a more general interest for research groups that want to pursue applications for their ASR tools for similar purposes.

Below we will first describe the Willem frederik Hermans audio collection (IV-A) and the data that was used to train the speech recognition system (IV-B). Next, we will compare the speech recognition performance on the collection of the broadcast news system and a system that is tuned to the domain (IV-C).

A. The collection

The collection of audio recordings to be disclosed consists of some 10 to 15 hours of lectures and interviews featuring Willem Frederik Hermans (WFH). More data will become available at a

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

later stage. Although WFH is not the only speaker present in the material, his voice dominates the larger part of the collection. The lectures were recorded at different locations with different acoustics. The lectures which were studied in more detail have applause, laughter, coughing and questions from the audience that—even for a human listener—sometimes are hard to recognise. Parts of the interviews are quite informal and recorded in a home environment on celluloid tape.

B. Training data

One of the lectures and a television documentary with a number of interviews were manually annotated at word level (130 minutes of speech), with WFH speaking approximately 85 % of the time. A training set (78 minutes) was used for training the acoustic models. A test set was used to evaluate both the acoustic models and the language models during development. An evaluation set was used for the final evaluation of the system.

In the annotated speech material, WFH is speaking 110 out of the 130 minutes. It is therefore reasonable to expect that the recognition rate will improve when a speaker (WFH) dependent acoustic model is used instead of the broadcast news acoustic model. Two new acoustic models were trained. The first model was trained solely on the part of the training set in which WFH is speaking. The second model was created by adapting the broadcast news acoustic model (UT-BN2005) to the training data using an acoustic adaptation algorithm referred to as Structured Maximum a Posterior Linear Regression (SMAPLR) adaptation [12].

For training BN language models we used the Dutch newspaper corpus. Two other text collections were available for domain adaptation. A number of written interviews with WFH and one of his short novels made up the first collection (further referred to as WFH-text) containing one and a half million words. Word-level transcripts of general conversational speech from the Spoken Dutch Corpus [17] formed the other text collection. This collection consists of 1.65 M words. Both text collections were used to apply language model adaptation schemes in order to create a language model that better fits the WFH domain.

Two domain specific trigram language models were trained. These models both use a 30 K vocabulary containing domain specific words. The most occurring words from the WFH-text described above were complemented with the most occurring words from the newspaper corpus.

From each of the two domain specific text collections a language model was created. A third model was created using the newspaper corpus and the $30K$ vocabulary. From these three models, a mixture language model that combines the statistics of each separate language model into a single language model using a weight factor was created. Mixture weights were computed using the transcripts of the acoustic training set.

C. Experimental results

The word error rates of the broadcast news acoustic model and the adapted acoustic models on the WFH-collection are shown in Table II. In order to make a fair comparison with the broadcast news system, the 65 K broadcast news language model was used during these recognition runs. Table II shows three word error rates for each model. The first WER is of the part of the audio in which WFH is speaking, the second one is based on speech from other people and the third is the overall word error rate.

Both adapted models perform better than the broadcast news model. Although the broadcast news model performs best on the small subset with various speakers (15 % of the total amount of speech), the adapted models show improved WERs on the part of the data in which WFH is speaking. The SMAPLR adapted model (66.9 % WER) outperforms the speaker dependent model (76.6 % WER). Apparently, the 78 minutes of speech used for training the speaker dependent model does not contain enough data for building a robust acoustic model.

AM	%WER WFH	%WER other	%WER total
UT-BN2005	81.6	67.2	80.4
WFH	76.0	83.2	76.6
BN-SMAPLR	66.7	77.1	67.5

TABLE II

WERs of three acoustic models: the 2005 broadcast model, the WFH model and the SMAPLR adapted model. The second column shows the WER of the part in which WFH is speaking, the third the WER on the other speech parts of the evaluation set. The last column shows the total WER.

The speaker adaptation we employed here, is a so-called 'supervised adaptation.' The 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.' For speaker segmentation an acoustical segmentation/clustering algorithm can parti-

tion the audio stream in segments belonging to the same speaker [6]. Using a first speech recognition pass, an automatic transcript can be generated. For each cluster of audio-segments spoken by the same speaker, this automatic transcript can be used to adapt the speaker-independent acoustic models to cluster-dependent models. These models can then be employed for a second speech recognition pass for the speech segments in that cluster, in order to obtain more accurate transcripts. Our supervised SMAPLR experiments give an impression of the maximum achievable performance increase, reducing the WER from 81.6 % to 66.7 %, for speaker WFH.

In Table III the word error rates are shown of a system that uses both the adapted acoustic model and the domain specific $30K$ mixture language model. The combination of the 30K mixture language model and the SMAPLR adapted acoustic model results in the best system performance: 66.9 % WER.

AM	%WER WFH	%WER other	%WER mixLM	%WER BN-LM
WFH	73.8	83.6	74.6	76.6
ADP	66.4	72.5	66.9	67.5

TABLE III

The word error rates (WER) of the two adapted acoustic models combined with the 30K mixture language model. The first row contains the AM trained on WFH solely. The second row contains the SMAPLR adapted acoustic model. For comparison, the results from the baseline BN language models are given as well.

By creating a mixture LM and a speaker dependent AM the word error rate was reduced with 13.5% (16.8% relative). To determine possible further improvements, a brief error analysis was conducted. To investigate to what extent different audio conditions influence the word error rate, speech segments were classified into five classes: clean speech (F_0), speech with audible reverberation (F_1), speech containing background music (F_2), speech with background noise (F_3) and overlapping speech (speech interrupted by other speakers, F_4).

Table III shows the word error rates in each of the conditions. Two third of the segments in the WFH evaluation set are classified as 'clean.' One third contains reverberation, music, noise or overlapping speech. Although all speech in the music class is clearly understandable for human listeners music increases WER substantially, in the WFH task by more than 10%, which is comparable with the statistics reported in [13].

Class	%WER WFH	%WER other	%WER total
F_0	63.9	61.8	63.8
F_1	75.4	100	76.5
F_2	76.1	86.1	78.6
F_3	82.4	83.3	82.4
F_4	100	100	100

TABLE IV

The word error rates of the five manually classified parts of the WFH evaluation set.

V. Discussion and conclusion

The Willem Frederik Hermans case study revealed that simply deploying the broadcast news system for transcribing oral history data resulted in high error rates of around 80% WER. In order to improve on the BN system, several adaptation schemes were applied on the acoustic level and on the language model level that indicated that a maximum achievable performance increase, reducing the WER from 81.6 % to 66.7 % for speaker WFH, is achievable. The overall word error rate was 66.9%, a 13.5% absolute improvement on the baseline BN system. Although near perfect transcripts are not required for a successful application of spoken document retrieval, error rates in this range are clearly below threshold.

It has already been noted that the large variability in audio quality, speech characteristics and topics are typical for the oral history domain and make the successful application of speech recognition technology difficult. A descriptive study on the characteristics of a certain collection is an important minimal prerequisite for identifying useful developments strategies.

What makes a successful application even more difficult is the fact that for the oral history 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. This can be due to the fact that oral-history archives have limited resources so that links to useful metadata are simply missing, or to the historical nature of the collections. For example, in the ECHO collection we could only use contemporary newspaper texts to model the ancient, out-dated speech of the Dutch Queen Wilhelmina (1880-1962), as there were no example text data digitally available that could be used to model this type of speech. An attempt to apply OCR techniques on related historical text data failed because of the low quality of the paper copies. Next to text (language model) related problems, ancient or dialectic speech that does not or only minimally occur in contemporary

speech training databases, impose additional constraints to the effort to obtain an adequate speech recognition performance. A first step towards the successful application of the automatic disclosure of oral-history collections, should therefore be to collect (from a speech recognition developer point of view) or make available (from a content provider point of view) as much related data sources as possible for fine-tuning the system. A strategy to deal with the lack of acoustic training data is deploying (partly) unsupervised training strategies.

Other topics that need to be addressed are related to the retrieval functionality proper. Dependent on the users that are expected to search the collections, this functionality may need to be adapted. For example, users may use highly selective, domain specific-words in their queries that are infrequent in the material. Because of their low frequency, such words have a low chance of being selected for the speech recognition vocabulary and thus become so called 'query out-of-vocabulary'. Applying a word spotting approach to search for such words would then be an option. Another example concerns the presentation of the retrieval results. Presenting a user with short excerpts from the collection that contain query words may not be very informative. Instead, providing coherent fragments that are structured according to speaker or topic may be preferred.

It can be concluded that applying audio mining techniques for the disclosure of oral-history collections is a promising approach. Proof-of-concept has already been provided in other domains. However, due to the typical characteristics of the oral history domain, substantially more effort must be directed towards obtaining speech transcripts that can be used adequately for indexing. Research aiming at the optimisation of presentation strategies, of interest for spoken-word collections in general, could also boost the usability of audio mining considerably.

ACKNOWLEDGEMENT

This paper is based on research funded in part by the Dutch projects MultimediaN and Waterland. We like to thank the Willem Frederik Hermans Institute for providing text and audio material.

REFERENCES

[1] ECHO Project Homepage. `http://pc-erato2.iei.pi.cnr.it/echo/`.

[2] 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 *Proceedings of the third ACM international conference on Multimedia*, pages 35–43, San Francisco, November 1995. ACM Press.

[3] 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.

[4] K. W. Church. Speech and Language Processing: Where Have We Been and Where Are We Going? In *Eurospeech-2003*, Genève, Switzerland, September 2003.

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

[6] Jean-Luc Gauvain, Lori Lamel, and Gilles Adda. The LIMSI broadcast news transcription system. *SPeech Communication*, pages 89–108, 2002.

[7] J. Gemmell, G. Bell, R. Lueder, S. Drucker, and C.Wong. Mylifebits: fulfilling the memex vision. In *ACM Multimedia*, pages 235–238, 2002.

[8] Jerry Goldman et al. Report of the EU/NSF working group on Spoken Word Audio Archives. `http://www.ercim.org/publication/ws-proceedings/Delos-NSF/SpokenWord.pdf`, 2003.

[9] A. Goodrum and E. Rasmussen. Sound and speech in information retrieval: an introduction. Bulletin of the American Society for Information Science, 2000. (URL: `http://www.asis.org/Bulletin/June-00/godrumrasmussen.html`).

[10] 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 (to appear)*, Amsterdam, September 2005.

[11] Roeland Ordelman. *Dutch Speech Recognition in Multimedia Information Retrieval*. PhD thesis, University of Twente, The Netherlands, October 2003.

[12] B. Pellom and K. Hacioglu. Recent Improvements in the CU Sonic ASR system for Noisy Speech: The SPINE Task. In *Proc. ICASSP*, 2003.

[13] B. Raj, V. N. Parikh, and R. M. Stern. The Effects Of Background Music On Speech Recognition Accuracy. In *Proc. of the ICASSP, Munich, Germany*, 1997.

[14] S. Renals. Ami: Augmented multiparty interaction. In *Proc. NIST Meeting Transcription Workshop*, Montreal, 2004. AMI-10.

[15] P. Rennert. Streamsage unsupervised asr-based topic segmentation. In *TRECVID 2003 - Text REtrieval Conference Video Track*, Gaithersburg, Maryland, November 2003.

[16] Tony Robinson, Mike Hochberg, and Steve Renals. *The use of recurrent networks in continuous speech recognition*, chapter 7, pages 233–258. Kluwer Academic Publishers, 1996.

[17] I. Schuurman, M. Schouppe, H. Hoekstra, and T. van der Wouden. CGN, an Annotated Corpus of Spoken Dutch. In *In Proceedings of the 4th International Workshop on Linguistically Interpreted Corpora (LINC-03)*, 2003.

[18] S.E. Tranter, K. Yu, D.A. Reynolds, G. Evermann, D.Y. Kim, and P.C.Woodland. An investigation into the interactions between speaker diarisation systems and automatic speech transcription. Technical report, Cambridge University Engineering Department, October 2003.

[19] E. Voorhees, J. Garofolo, and K. Spärck Jones. The TREC-6 Spoken Document Retrieval Track. In *Proceedings DARPA Speech Recognition Workshop*, 1997.

[20] E. Zuurbier. Onderzoek naar de haalbaarheid van spoken document retrieval. Master's thesis, University of Twente, 2004.