Automated speech and audio analysis for semantic access to multimedia

Franciska de Jong[1][2], Roeland Ordelman[1], and Marijn Huijbregts[1]

[1] University of Twente, Dept. of Computer Science,
P.O. Box 217, 7500 AE, Enschede, The Netherlands
[2] TNO-ICT, Delft, The Netherlands
{fdejong,ordelman,m.a.h.huijbregts}@ewi.utwente.nl
http://hmi.ewi.utwente.nl

Abstract. The deployment and integration of audio processing tools can enhance the semantic annotation of multimedia content, and as a consequence, improve the effectiveness of conceptual access tools. This paper overviews the various ways in which automatic speech and audio analysis can contribute to increased granularity of automatically extracted metadata. A number of techniques will be presented, including the alignment of speech and text resources, large vocabulary speech recognition, key word spotting and speaker classification. The applicability of techniques will be discussed from a media crossing perspective. The added value of the techniques and their potential contribution to the content value chain will be illustrated by the description of two (complementary) demonstrators for browsing broadcast news archives.

1 Introduction

The growing role expected for networked electronic media and the increasing size of content repositories require augmented attention for the automation of content-based extraction and integration of metadata for video, audio and textual content. Content-based metadata are a prerequisite for conceptual search both for professional users and the general public, and they play an important role in the exploitability of content. The adagium used to be 'Content is king', but metadata rules.

Semi-automatic metadata extraction has been given attention for a variety of monomedia content types and formats. For video and image content this has led to an interesting growth in the number of objects and events that can be detected on the basis of low-level features [13]. But in view of the huge range of concepts that users may want to search for, the field of video analysis can be argued to be still in its infancy.

A strategy that can help compensate for the limitations of image analysis is the exploitation of surface features. Surface features are those properties of (multimedia) documents that do not describe content. Examples include the length of a document, references to the document's location, and the production date. Although these features do not directly relate to the document's coverage, they have proven to be valuable additional sources of information in a retrieval setting [17]. More importantly they illustrate the importance of not being too restrictive in exploiting available secondary data streams.

As is widely acknowledged, the exploitation of linguistic content in multimedia archives can boost the accessibility of multimedia archives enormously. Already in 1995, [3] demonstrated the use of subtitling information for retrieval of broadcast news videos, and in the context of TRECVID [13] the best performing video retrieval systems always exploit speech transcripts. The added value of linguistic data is of course limited to video data containing textual and/or spoken content, or to video content with links to related textual documents, e.g. subtitles, generated transcripts etc. But when available, using linguistic content for the generation of a time-coded index can help to bridge the semantic gap between media features and user needs.

In the next two sections we first explore some methods that deal with the exploitation of already available linguistic content in, or attached to, multimedia databases. We introduce the concept of cross-media mining in section 4. Automatic audio indexing techniques are overviewed in section 5. The system architecture of the recognition environment is detailed in section 6. Finally, the added value of links that are automatically generated across media via high level annotation will be illustrated in section 7. This section will provide a description of two complementary demonstrators: one for on-line access to an archive of news broadcasts linked up to a newspaper archive, the other illustrating a crucial aspect in browsing multimedia databases, a technique known as *document clustering* applied in combination with topic detection.

2 Exploiting collateral text

The semantic gap between user needs and content features is as old as the concept of archiving itself. The traditional approach towards the creation of an index is to rely on manual annotation with controlled vocabulary index terms. With the emergence of digital archiving this approach is still widely in use and for many archiving institutes the creation of manually generated metadata is and will be an important part of the daily work. When the automation of metadata generation is considered, it is often seen as something that can enhance the existing process rather than replace it. The available metadata will therefore often be a combination of highly reliable and conceptually rich annotations, and (semi)automatically generated metadata. One of the challenges for search environments is to combine the various types of metadata and to exploit the added value of the combination. In this paper we will explain how available high level annotations for media archives can be exploited for improved automated generation of additional language-based annotations, and *vice versa*, how automatic content processing can help to generate ontological and thesaurial media annotations. For the content-based processing tasks the main focus will be on the various ways in which automatic speech and audio analysis can be deployed.

Depending on the resources available within an organization that administers a media collection, the amount of detail of the metadata and their characteristics may vary. Large national audiovisual institutions such as Beeld&Geluid in The Netherlands[3], annotate at least titles, dates and short content descriptions (descriptive metadata). Many organizations with multimedia collections however, often do not have the resources to apply even some basic form of archiving.

To still allow the conceptual querying of video content, collateral textual resources that are closely related with the collection items can be exploited. A well known example of such a textual resource is subtitling information for the hearing-impaired (e.g., CEEFAX pages 888 in the UK) that is available for the majority of contemporary broadcast items, in any case for news programs. Subtitles contain a nearly complete transcription of the words spoken in the video items and provide an excellent information source for indexing. Usually, they can easily be linked to the video by using the time-codes that come with the subtitles. The Dutch news subtitles even provide topic boundaries that can be used for segmenting the news show into subdocuments. Textual sources that can play a similar role are teleprompter files: the texts read from screen by an anchor person (also referred to as auto-cues).

The time labels in these sources are crucial for the creation of a textual index into video. As in full text retrieval, where all words in a document can function as index terms and thus as a link to a document, the exploitation of collateral transcriptions for speech in video will allow that all words spoken offer a link to the fragment in which they occur. And though full text retrieval is certainly not the ultimate solution to the semantic gap, natural language is inherently closer to the level of concepts than low-level image features.

3 Time alignment

In the collateral text sources mentioned above, the available time-labels are not always fully reliable and can even be absent. In such cases the text files will have to be synchronized. Examples of such text sources are minutes of meetings, or written versions of lectures and speeches. This section will describe methods for the automatic generation of time-stamps for minutes in two pilot projects in the domain of e-Government. These minutes pertain to the so-called *Handelingen*, i.e. the meetings of the Dutch Parliament, and to city council meetings. Due to the difference in accuracy of the minutes, two different approaches had to be developed.

The minutes of the meetings of the Dutch Parliament are stenographic minutes that closely follow the discourse of the meeting, only correcting slips of the tongue and ungrammatical sentences. Given the close match with the actual speech, a relatively straightforward so-called forced alignment procedure could be used. Forced alignment is a technique commonly used in acoustic model training in automatic speech recognition (ASR). In order to be able to train phone

[3] Beeld&Geluid:`http://www.beeldengeluid.nl/`

models, words and phones in pre-segmented sentences are aligned to their exact location in the speech segments using an acoustic model[4]. Given a set of words from a sentence the acoustic model tries to find the most optimal distributions of these words given the audio signal on the basis of the sounds the words are composed of. When using alignment for indexing, pre-segmented sentences are evidently not available but as long as the text follows the speech well enough, the word alignment can be found by using relatively large windows of text.

The alignment procedure works well even if some words in the minutes are actually not in the speech signal. However, if the text to be aligned does not match the speech too well, as was the case with city council meetings, and if the text segments are too large, the alignment procedure will fail to find a proper alignment. In order to produce suitable segments, we used a two-pass strategy, similar as proposed in [9], incorporating the following steps:

1. a baseline large vocabulary speech recognition system[5] is used to generate a relatively inaccurate transcript of the speech with word-timing;
2. the transcript is aligned on the word level to the minutes using a dynamic programming algorithm;
3. where the transcript and the minutes match (three words in a row correctly aligned), so-called ‘anchors’ are inserted
4. using the word-timing labels provided by the speech recognition system, the anchors are used to generate suitable segments;
5. individual segments of audio and text are accurately synchronized using forced alignment;

The described methods allow for the synchronization of audiovisual data to available linguistic content that approximates to a certain extent the speech in the source data and they enable the processing of conceptual queries of the audiovisual content with readily available tools.

4 Cross-media mining

Ideally one would not only synchronize audiovisual material with content that approximates the speech in the data, but take even one step further and exploit *any* collateral textual resource, or even better: any kind of textual resource that is accessible, including open source titles and proprietary data (e.g., trusted webpages and newspaper articles). Another way of putting it: we propose to shift the focus from indexing individual multimedia documents to video mining in truly multimedia distributed databases. In the context of meetings for example, usually an agenda, documents on agenda topics and cv’s of meeting participants can be obtained and added to the repository. Mining these resources can support

[4] In the first iteration usually an ’averaged’ bootstrap model is used. The alignment and the model should improve iteratively

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

information search because it yields annotations that offer the user not just access to a specific media type, but also different perspectives on the available data. An agenda could help to add structure that can for example be presented in a network representation, whereas cv's can be linked to annotations resulting from automatic speaker segmentation. In addition, both documents and cv's would allow for multi-source information extraction.

A typical example of what the cross-media perspective can yield in the broadcast news domain is the linking of newspaper articles with broadcast items and *vice versa* (Cf. Fig 1). Links can be established between two news objects which count as similar on the basis of the language models assigned to them via statistical analysis. Typically such language models are determined by the frequency of the linguistic units such as written or spoken words and their co-occurrences. The similarity between two documents can be decided for each pair of documents, but a more common approach is to pre-structure a document collection into clusters of documents with similar language models. Similarity of language models predicts similarity of topic, and therefore this technique is know as topic clustering[6].

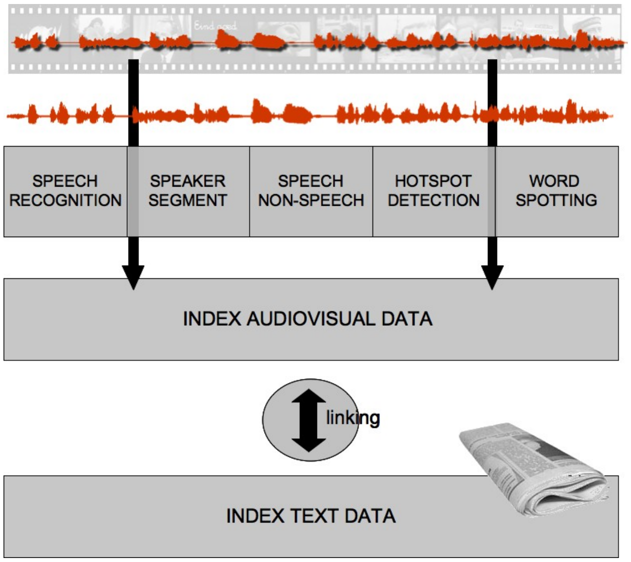

Fig. 1. Linking audio and textual sources

In addition to linking documents with a similar topic profile, which can be supportive in a browser environment, also the available semantic annotation for

[6] The functionality commonly known as *topic detection and tracking* (TDT) for dynamic news streams has been built upon it and plays a central role in the evaluation series for TDT organized by DARPA.

documents with similar profiles can be exchanged and exploited for conceptual search. If a newspaper article has been manually classified as belonging to e.g., economy or foreign politics, a broadcast item with a similar language model can be classified with these conceptual labels as well. In section 7 below, we describe a cross-media news browser demonstrator that incorporates this functionality.

For the linking of audiovisual data with textual resources that are not directly related to the speech, such as documents on agenda topics in the context of meetings, or newspaper data in the broadcast news domain, we have to step up the use of speech recognition compared to the speech recognition deployed in the alignment procedures described in section 3. In the more elaborate alignment procedure, an initial hypothesis is generated by a large vocabulary speech recognition system. As this hypothesis is only needed for finding useful segments, we do not really care about the performance of the system as long as it is able to provide us with 'anchors'. However, the relevance of speech recognition performance increases when textual resources suitable for alignment with audiovisual data are *not* available. In the next section, the application of speech recognition technology as the *primary* source for generating a textual representation of audiovisual documents that can be linked to other linguistic content, is described.

5 Audio indexing

Recent investigations have shown the feasibility of deploying large vocabulary speech recognition for the generation of multimedia annotations that allow the conceptual querying of video content and the synchronization to any kind of textual resource that is accessible, including other full-text annotation for audiovisual material. The potential of ASR-based indexing has been demonstrated most successfully in the broadcast news domain. Data collection for training a speech recognition system in this general domain is relatively easy. Word-error-rates below 10% are no longer exceptional. For the broadcast news domain, ASR transcripts approximate the quality of manual transcripts, at least for several languages. Spoken document retrieval in the American-English broadcast news (BN) domain has even been declared 'a solved problem' with the NIST-sponsored TREC SDR track in 1999 [6]. It should be noted however that in other domains than broadcast news and for many less favored languages, a similar recognition performance is usually harder to obtain due to (i) lack of domain-specific training data, and (ii) large variability in audio quality, speech characteristics and topics being addressed. However, when recognition performance remains within certain boundaries (an ASR performance of 50% WER is typically regarded as a lower bound for successful retrieval) the damage in terms of retrieval performance may be acceptable, especially when no other means (metadata) are available for searching.

5.1 Vocabulary selection via collateral text

One of the main research topics in large vocabulary speech recognition is vocabulary selection. Given the huge quantities of available training data for the broadcast news domain, the acoustic models and language models can usually be trained adequately and in addition, various acoustic adaptation procedures (using e.g., bandwidth/gender-dependent models, speaker-adaptive training, etc.) can be applied to boost ASR performance. However, language models and recognition vocabularies are usually created using fixed and, with respect to broadcast news, often outdated training corpora. Vocabularies are based on word frequencies within corpora, while the linguistic properties in broadcast news are continuously changing: previously infrequent names of places and people can start occurring frequently without prior indication, people dominating the news during a period of time may disappear from the headlines after a while, jargon may suddenly 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 highly improbable in another period (e.g., *hurricane*, or *Christmas tree*.

To limit the number of out-of-vocabulary (OOV) words, the ASR engine of an SDR environment should be based on models that adapt to linguistic variation. OOVs damage retrieval performance in two ways: firstly, a query consisting of an OOV word, a so-called QOV (query-out-of-vocabulary), will never match a transcript, even though the QOV occurs in the speech. Secondly, the word occurring in the transcript at the position of the OOV may induces a false hit. Although document expansion and query expansion techniques may be deployed to compensate for QOV words [18, 7], tackling OOVs in an earlier stage is favorable. As named entities play an important role in the mining of broadcast news, regular updates of a recognition vocabulary with 'new' proper names is crucial. To keep an ASR engine 'tuned', up-to-date training material is required. Ideally, this is dealt with via a daily feed of newspaper content. Alternatively contemporary data can be collected via the Web [1] or by capturing subtitling information from news programs. A number of vocabulary selection methods have been proposed, based on parallel corpora. They are based on the use of a narrow look-back time window to select new words [2], the use of word history information [11], or the use of vectors combining word frequency data from multiple corpora [1].

5.2 Word spotting

Another way of reducing the effect of OOV words is to use word spotting. Word spotting is the audio search functionality that matches query terms for audio content either directly or via a phone (or phone-lattice) representation of query and content (cf. [8]) and can be very effective, especially when the vocabulary for the domain is hard to predict, resulting in high OOV-rates. Word spotting can be combined with ASR based on a full text transcription to 'correct' OOVs or misrecognised names. In this approach the following steps are taken: (i) the initial ASR transcript is used to identify related collateral textual data; (ii) named entity detection applied to the collateral textual data sources provides a list of named entities that are relevant given the audio document topic; (iii) the occurrence (and timestamps) of these named entities in the audio are recovered using a word spotting approach.

5.3 Speaker classification

There is more information in the speech than the words alone. Speaker characteristics can be extracted from the speech (speaker's voice, word usage, syntax) as well and may serve as an additional information layer, for example to add structure for browsing (speaker segmentation and identification) or to extract features that could not be accessed using traditional views on the data. Automatic speaker classification can especially be beneficial for spoken documents in cultural heritage collections. Historical spoken word archives receive attention from professional information analysts in various fields. Historians for example, may be interested both in the exact words that were spoken, but also in the speaker's profile. The latter may partly be reconstructed using the identification of speaker characteristics such as accent, age, gender, speaking behavior and even emotion and cognitive state.

Instead of aiming at the extraction of speech features from a single speaker's voice, research has been directed to extract features from multiple voices, for example emotional features in order to detect so called 'hot spots' in collections. Typical examples of such 'hot spot' detections are the cheering of a crowd in a sports game, or laughter in the context of meetings (cf. [15]).

6 System description

In this section the ASR system architecture will be discussed in more detail. As indicated above, ASR can be deployed in different scenarios and for content in several domains, both within and outside the broadcast domain: documentaries, interviews, historical archives, recordings of lectures and recordings of meetings (corporate, scientific, parliamentary, etc.). As the characteristics of the speech encountered in these domains vary, robustness of recognition is taken as a requirement. A robust speech recognition system can be defined as a system that is capable of maintaining good recognition performance even when the quality of the speech input is low (environment, background noises, cross-talk, low audio quality), or when the acoustical, articulatory, or phonetic characteristics of the speech encountered in the training data differ from the speech in the task data. Even systems that are designed to be speaker independent cannot cover all the speaker variations that may occur in the different task domains.

As manually adjusting a system for each different situation is time consuming and error prone, we designed and developed an ASR system that is self-adjustable as much as possible, and that is robust for changing conditions. In Figure 2 the system is graphically represented. The gray area at the left represents the sub-system responsible for adapting our baseline, broadcast news (BN), knowledge models to a specific domain. This sub-system will be described in section 6.1. The second gray area represents the main audio processing stream. This sub-system consists of four modules. In the first module, the audio is cut up in smaller segments and each segment is assigned to a unique speaker. This process of determining 'who spoke when', called speaker diarization, will be discussed in section 6.2. After diarization, the second module performs a first speech recognition pass for each speaker using the domain specific models. This recognition result is then used in the third module to create a new acoustic model (AM) for each speaker. These speaker specific models are eventually used in the fourth module doing the final decoding pass. The decoding procedure is discussed in section 6.3. Some first attempts were made to make the system more robust by monitoring and analyzing the system's performance. In section 6.4 some of our monitoring tools are discussed.

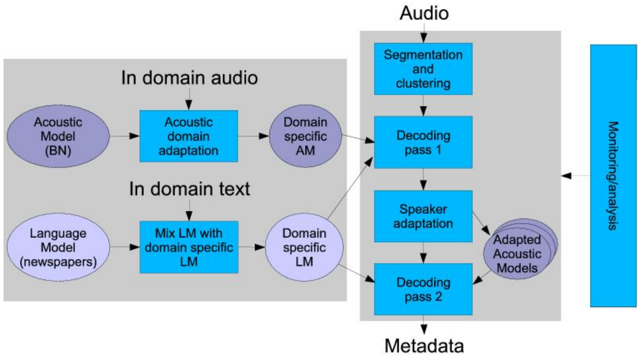

Fig. 2. The ASR system consisting of four modules: first the audio is divided in segments and each segment is assigned to a speaker. The speech of each speaker is decoded and used for adapting the acoustic model. The adapted models are used in the second decoding pass.

6.1 Acoustic and language model creation

The system uses two kinds of knowledge models. The acoustic model (AM) is used for calculating which sequence of phones has the highest probability of being pronounced at a certain time in the audio stream. The language model (LM) is used for calculating which sequence of words is most likely pronounced. Large amounts of training data are required to extract the statistics that are needed to create these models. Unfortunately, for most application domains outside the news domain, data is not sufficiently available for model training.

For the broadcast news domain relatively large text corpora are available. The Twente News Corpus contains over 450M words of newspaper text data that are used to train our BN language model. The BN acoustic model is trained on approximately 150 hours of speech from a variety of sources including the Spoken Dutch Corpus [10]. (Partly) unsupervised methods to augment the basis for training could offer compensation. In order to use the system in other domains than broadcast news, we typically use small amounts of in-domain audio and text data to adapt our BN models to the new task domain. The goal is to adjust the acoustic model parameters so that the model better fits the task domain using a model-space transformation method (such as SMAPLR [12]).

6.2 Speaker diarization

The task of the first module in the system is to cut the audio in smaller *segments* suitable for input to our decoder. Speech and non-speech segments are distinguished (Speech Activity Detection) and speaker changes are detected. Simultaneously the module determines 'who spoke when' and clusters the segments of unique speakers. This procedure is often called *speaker diarization*.

The speaker diarization module was evaluated in the speaker diarization task of the NIST Rich Transcription 2006 Spring Evaluation (RT06s). After all non-speech, such as silence and background noise, has been removed from the audio, the remaining segments are passed to a modified Hidden Markov Model (HMM) speech decoder. The HMM topology consists of a number of parallel single state HMMs connected to each other by a single (non-emitting) start- and end-state. Each state represents a single speaker from the audio. Because the exact number of speakers is unknown, the system will initially contain more states than the maximum expected number of speakers. After training the HMM with the input audio, the number of states is reduced until an optimum system likelihood is reached. At RT06s we have successfully tested two methods for reducing the number of states [16]. In the ideal situation, after this procedure there will be exactly as many states as there are speakers in the audio and each state will be trained on a single unique speaker. Performing a Viterbi alignment using the final HMM topology will result in a set of speech segments for each speaker.

The major advantage of the used diarization method is that hardly any parameters need to be tuned for different audio or domain conditions. As a consequence the resulting system is robust to changing conditions.

The diarization module can cluster all speech from a single speaker within a single audio document. This in-document speaker information can already be valuable metadata in itself (see section 5.3), but in order to track speakers across documents we would like to cluster speakers over document boundaries. This should be possible by extending this technology so that speaker models of different documents can be compared.

6.3 Two-pass decoding

Once the audio has been segmented and clustered, the speech is recognized in either one or two decoding passes. Our most recent systems use the recognition of the first pass to adapt the acoustic model to the speakers encountered in the data. This so called speaker adaptation is performed using audio with high confidence scores using the SMAPLR adaptation method [12]. The new speaker specific acoustic models are then used in the second decoding pass to produce the final recognition.

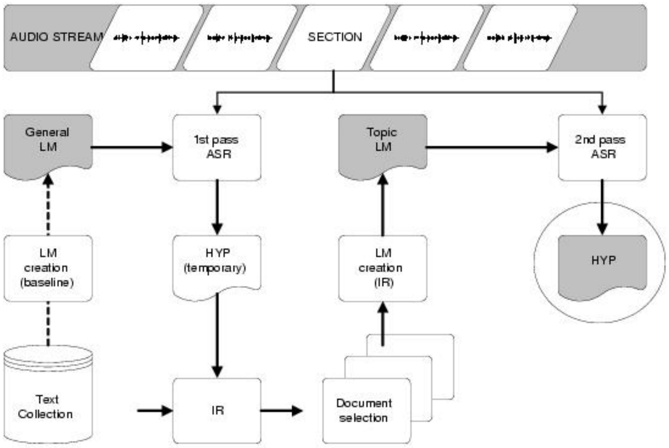

Fig. 3. Topic based LM: segmentation of the audio file, initial speech recognition on the audio segments, defining the 'topic' on the basis of the speech transcripts (here using an IR system), creating a topic specific language model, and finally the final speech recognition run using the topic-based language model.

A two-stage recognition run allows for domain adaptation on the word level as well. Here, the strategy is to assign a topic to segments of the input file on the basis of the speech transcript of the first run. The topic assignment is then used to select a topic-specific language model. As evidently real topic-based segmentations are not known a-priori, readily available segmentations, such as on the change of speaker, on longer silence intervals or even fixed time-windows, are chosen to divide the audio document in smaller parts. These parts are then further regarded as representing single topics.

The topics can be assigned either implicitly or explicitly. An explicit topic assignment refers to using specific topic labels, for example generated on the basis of a topic-classification system that assigns thesaurus terms. From a collateral text corpus (e.g., a newspaper corpus) that is labeled with the same thesaurus terms, documents that are similar to the topic in the segment can be harvested for creating a topic-specific language model. For implicit topic assignment, an Information Retrieval system is used for the selection of documents from an unstructured collateral text source (e.g., internet sources) that have a similar topic: on the basis of the stopped speech transcript that serves as a query, a ranked list of similar documents is generated; the top N documents of the list in turn serve as input for language modeling. Having created a topic specific language model a second speech recognition run is performed on the same segment with the new language model to generate the final transcript. The procedure is visually depicted in Figure 3.

Both recognition passes are performed with the University of Twente 2006 (UT06) decoder. The UT06 decoder is a Large Vocabulary Continuous Speech

Recognition decoder that uses Hidden Markov Models (HMM). Its state emission probabilities are calculated by Gaussian Mixture Models (GMM). We have trained models based on Perceptual Minimum Variance Distortionless Response (PMVDR) cepstral coëfficient features (created using the Sonic LVCSR toolkit [19]) and with Mel Frequency Cepstral Coëfficient (MFCC) features.

6.4 System monitoring and output analysis

Given that (i) a large amount of system parameters need to be fine-tuned for every application domain, and (ii) the system's behavior often needs to be monitored over a longer period of time (e.g., in longitudinal tasks), various methods to monitor and analyze the system are being developed.

One of these methods is blame assignment [4]. This method uses a small evaluation set (audio that has been manually annotated on the word level) to evaluate recognition accuracy. Incorrectly decoded words are grouped in *error regions*. For each region it is calculated in which error class the region belongs. There are five error classes: (i) the region contains out of vocabulary words, (ii) the error is caused by a LM mismatch, (iii) an AM mismatch or (iv) a mismatch in both the LM and AM, and (v) the error is caused by pruning away the correct path during decoding (search error).

The blame assignment method can only be used when evaluation data is available. When monitoring a BN system that needs to decode broadcasts on a daily basis, transcribed data may not be available. We are currently investigating system behavior over a longer period of time using collateral data sources such as subtitling information that comes with broadcast news programmes or minutes of meetings as evaluation data. The difficulty here is how the mismatch between the noisy speech transcripts containing errors and the incomplete and/or reformatted collateral text data should be interpreted.

7 The power of transcripts demonstrated

A number of techniques described above have been implemented in two demonstrators described below. They illustrate how, on the basis of textual transcripts, the concept of cross-media news browsing for a multifaceted or layered multimedia archive can be realized.

7.1 Cross-media news browser

The so-called cross-media news browser demonstrates on-line access tools to an archive of Dutch news broadcasts (NOS 8 uur Journaal). It shows how either available collateral data sources (subtitling information for the hearing-impaired) or full-text speech recognition transcripts can be used as linguistic content for the generation of time-coded indexes for searching within news shows. Although the subtitling information in itself would already be enough to enable access, speech recognition transcripts are generated as well for demonstration purposes. The subtitling information is captured using a teletext capturing card and synchronized with the video using a manually determined off-set value. The speech recognition system consists of decision-tree state-clustered acoustic models trained on approximately 20 hours of speech from the Spoken Dutch Corpus [10], 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 the 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, we can use real topic boundaries for the segmentation of the news show into 'subdocuments'. In case we have to rely on ASR transcripts only, the segmentation news can be based on acoustic information such as speaker changes, speech/non-speech transitions and silences.

In order to demonstrate the added value of links that are automatically generated across media types via high level annotation, the linguistic annotations of the news items (either based on subtitles or ASR) are linked to an up-to-date database of Dutch newspaper articles made available for demonstration purposes by PCM publishers[7]. The links from broadcast news fragments to related newspaper articles are 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.

7.2 Novalist news browser

The broadcast news browser described above primarily demonstrates the added value of automatic linguistic annotation of audiovisual content. The functionality of the so-called Novalist browser (developed at TNO) can be regarded as complementary: it aims to facilitate the work of information analysts in the following way: (i) related news stories are clustered to create dossiers or 'threads', (ii) dossiers resulting from clustering are analyzed and automatically annotated with several types of metadata, and (iii) a browsing screen provides multiple views on the dossiers and their metadata. All analysis steps can be performed data-driven.

The corpus covered by the Novalist demonstrator consists of a collection of news items published by a number of major Dutch newspapers and magazines, web crawls, a video corpus of several news magazines and a video archive with all 2001 news broadcasts of *NOS 8 uur Journaal*. Here, the teleprompter files for the video archive function as collateral text. Transcripts of broadcast audio generated with automatic speech recognition (ASR) can also be incorporated.

[7] PCM publishers is one of the largest publishers in the Dutch language region: `http://www.pcmuitgevers.nl/`

The entire demonstrator collection consists of some 160,000 individual news items from 21 different sources, and recent new releases even more.

The system has to deal with dynamic information, about which no full prior knowledge is available. There is no fixed number of target topics and events types. The system must both discover new events as the incoming stories are processed, and associate incoming stories with the event-based story clusters already created. Document clustering is done incrementally: for a new incoming story, the system has to decide instantaneously to which topic cluster the story belongs. Since the clustering algorithms are unsupervised, no training data is needed.

Via document clustering, structure is generated in news streams, while the annotations can be applied as filters: search for relevant items can be limited to relevant subsets of the collection. Novalist dossiers are visualized in a compact overview window with links to a time axis. Additional functionality consists of the automatic generation of links to related sources, both internal and external. For a detailed explanation of the concept of topic detection and the similarity concept applied in the language modeling approach that is underlying Novalist, and for an overview of the performance evaluation of some components, cf. [14], [5].

8 Conclusion

The two demonstration browsers described here show how automatically extracted annotation based on non-image features can successfully support the exploitation of multimedia content. The possibility to link textual content from diverse sources to media files and vice versa, strengthens the impact of audio and speech analysis. The transcript processing techniques deployed can be linked to query functionality at several levels of conceptual abstraction: from the words spoken to higher level semantic concepts that have automatically detected in the textual content via clustering and classification. Future research will be directed towards the integration with metadata generation based on visual analysis.

Acknowledgment

This work was partly supported by the Dutch bsik-programme MultimediaN (`http://www.multimedian.nl`) and the EU projects AMI (IST-FP6-506811), MESH (IST-FP6-027685), and MediaCampaign (IST-PF6-027413).

References

1. A. Allauzen and J.L. Gauvain. Diachronic vocabulary adaptation for broadcast news transcription. In *InterSpeech*, Lisbon, September 2005.
2. C. Auzanne, J.S. Garofolo, J.G. Fiscus, and W.M Fisher. Automatic Language Model Adaptation for Spoken Document Retrieval. In *Proceedings of RIAO 2000, Content-Based Multimedia Information Access*, pages 132–141, 2000.

3. 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.
4. Lin Chase. Blame assignment for errors made by large vocabulary speech recognizers. In *proceedings Eurospeech '97*, pages 1563–1566, Rhodes, Greece, 1997.
5. F.M.G. de Jong and W. Kraaij. Content reduction for cross-media browsing. In H. Saggion and J.-L. Minel, editors, *RANLP workshop 'Crossing Barriers in Text Summarization Reserach*, pages 64–69, Borovets, Bulgaria, 2005.
6. J.S. Garofolo, C.G.P. Auzanne, and E.M Voorhees. The TREC SDR Track: A Success Story. In *Eighth Text Retrieval Conference*, pages 107–129, Washington, 2000.
7. 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*, pages 8–13, Cambridge, UK, 1999.
8. W. Kraaij, J. van Gent, R. Ekkelenkamp, and D. van Leeuwen. Phoneme based spoken document retrieval. In *Proceedings of the fourteenth Twente Workshop on Language Technology TWLT-14*, pages 141–153, University of Twente, 1998.
9. 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.
10. N. Oostdijk. The Spoken Dutch Corpus. Overview and first evaluation. In M. Gravilidou, G. Carayannis, S. Markantonatou, S. Piperidis, and G. Stainhaouer, editors, *Second International Conference on Language Resources and Evaluation*, volume II, pages 887–894, 2000.
11. R.J.F. Ordelman. *Dutch Speech Recognition in Multimedia Information Retrieval*. Phd thesis, University of Twente, Enschede, October 2003. publisher: Taaluitgeverij Neslia Paniculata publisherlocation: Enschede, ISSN: 1381-3617; No 03-56, ISBN: 90-75296-08-8, Numberofpages: 268.
12. O. Siohan, T. Myrvol, and C. Lee. Structural maximum a posteriori linear regression for fast hmm adaptation, 2000.
13. A.F Smeaton, W. Kraaij, and P. Over. Trecvid - an overview. In *Proceedings of TRECVID 2003*, USA, 2003. NIST.
14. Martijn Spitters and Wessel Kraaij. Unsupervised clustering in multilingual news streams. In *Proceedings of the LREC 2002 workshop: Event Modelling for Multilingual Document Linking*, pages 42–46, 2002.
15. Khiet P. Truong and David A. van Leeuwen. Automatic detection of laughter. In *InterSpeech*, pages 485–488, Lisbon, September 2005.
16. David van Leeuwen and Marijn Huijbregts. The ami speaker diarization system for nist rt06s meeting data. In *in NIST 2006 Spring Rich Transcrition Evaluation Workshop*, Washington DC, USA, 2006.
17. Thijs Westerveld, Arjen P. de Vries, and Georgina Ramírez. Surface features in video retrieval. In *3rd International Workshop on Adaptive Multimedia Retrieval, AMR'05*, 2005.
18. 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*, pages 372–374, Athens Greece, 2000.
19. U. Yapanel and J. H. L. Hansen. A new perspective on feature extraction for robust in-vehicle speech recognition. In *Proceedings of Eurospeech*, pages 1281–1284, 2003.