2009 IEEE International Conference on Semantic Computing

Enhanced Multimedia Content Access and Exploitation using Semantic Speech Retrieval

Roeland Ordelman*†
†*Netherlands Institute for Sound and Vision*
Hilversum, The Netherlands
rordelman@beeldengeluid.nl

Franciska de Jong*
**Human Media Interaction*
University of Twente
Enschede, The Netherlands
fdejong,ordelman@ewi.utwente.nl

Martha Larson‡
‡*Department of Mediamatics*
Delft University of Technology
Delft, The Netherlands
m.a.larson@tudelft.nl

***Abstract*—Techniques for automatic annotation of spoken content making use of speech recognition technology have long been characterized as holding unrealized promise to provide access to archives inundated with undisclosed multimedia material. This paper provides an overview of techniques and trends in semantic speech retrieval, which is taken to encompass all approaches offering meaning-based access to spoken word collections. We present descriptions, examples and insights for current techniques, including facing real-world heterogenity, aligning parallel resources and exploiting collateral collections. We also discuss ways in which speech recognition technology can be used to create multimedia connections that make new modes of access available to users. We conclude with an overview of the challenges for semantic speech retrieval in the workflow of a real-world archive and perspectives on future tasks in which speech retrieval integrates information related to affect and appeal, dimensions that transcend topic.**

***Keywords*-speech retrieval; spoken content; speech recognition; multimedia retrieval and access; semantics**

I. Introduction

Automatic semantic annotation has an important role to play in providing conceptual semantic access, referred to here as semantic retrieval, to today's continuously accumulating audiovisual collections and to enhancing their potential for exploitation. The traditional approach of manual annotation of content falls short given both data quantity and the growing demand for rich annotations. In this context, "rich" makes reference to the detail and diversity of the information contained in content descriptions. Preferably, annotation goes beyond what is covered by the traditional so-called tombstone metadata, which treat archived items as monolithic objects. Ideally conceptual annotation also zooms in on segments of individual audiovisual items or on links between related objects. Audiovisual content that is annotated for facetted access is often considered optimal for exploitation. What is needed to turn interaction with content into an attractive and effective experience is fine-grained semantic structure and connectedness to complementary multimedia content sources.

The many hours of new content that are produced (or retrospectively digitized) on a daily basis, and the implicit request for more fine-grained annotation would put exceedingly heavy demands on content repositories choosing exclusively manually assigned annotation as a means of exploiting their rich and potentially valuable content stores. In practice, for some repositories even the most basic form of archiving is hardly feasible, whereas others need to make selective use of their human annotation capacity. As a result, the level of disclosure is often insufficient for successful exploitation. At the same time, undisclosed content continues to accumulate, and the precious annotation capacity available fails in its fight to gain ground against the waves of incoming new material.

The automation of semantic annotation has the potential to reduce the pile of undisclosed content in archives and to enhance the support for facetted search. It may also lead to a revaluation of the manual craftsmanship of enriching content via interpretation and the deployment of knowledge of a collection or culture, as opposed to basic archiving. In this paper, we investigate this potential by focusing on automatic annotation based on the processing of the speech in audiovisual content. The use of speech technology, and in particular automatic speech recognition (ASR), for the exploitation of the linguistic content that is available as spoken content in videos has proven to be helpful in bridging the semantic gap between low-level media features and conceptual information needs and its use has been advocated for many years. However, success stories of applications in real-world scenarios seem to lag behind the promised potential.

On the basis of a number of past and present use cases that we encountered in a decade of spoken document retrieval research in collaborative projects in a national and European context, and in the real-world practice of the large audiovisual archive of the Netherlands Institute for Sound and Vision[1], in which automatic annotation based on speech technology is deployed, we reconfirm the merits for audiovisual content exploitation, discuss aspects that underly the still limited possibilities for applying semantic access

[1]http://www.beeldengeluid.nl/

978-0-7695-3800-6/09 $26.00 © 2009 IEEE
DOI 10.1109/ICSC.2009.80

technologies for spoken audio content, and briefly address future directions.

II. Speech-based indexing

The potential of speech-based indexing has been demonstrated most successfully in the broadcast news domain. Broadcast news (BN) involves relatively clean, planned and well-structured speech and the domain was studied in-depth along benchmarks focussing on both automatic speech recognition (HUB-4, [1]) and spoken document retrieval (TREC SDR, [2]). As a result of the attention directed at this domain, through the years large amounts of BN (related) training and test data became available via international channels[2] or more locally oriented initiatives for several languages, allowing the development of robust broadcast news systems with high performance levels. Improving ASR performance, e.g., by reducing out-of-vocabulary (OOV) words remains an important topic of research in the news domain given its named entity dynamics, but by using large vocabularies and updating these by deploying information from available textual resources (e.g., the Web) OOVs can to a large extent be reduced. Moreover, successful attempts to alleviate the OOV problem by exploiting n-best lists or lattices, or by combining word-based systems with phone-based systems have been reported [3], [4]. More recently research has been conducted into the possibility to use BN transcripts as a basis for conceptual indexing. This research theme will be outlined in section III). Outside the broadcast news domains, attempts to improve ASR or SDR performance have been focused on the processing of audio produced in unknown conditions (section III-A), on exploiting collateral content for e.g., language models improvement (section III-B), and on alignment of ASR transcripts and other textual resources (section III-C).

III. Concept detection in speech transcripts

In case content to be indexed is from a known domain, typically domain knowledge is available that can guide the automatic analysis of the content. Knowledge-assisted content analysis has been given an enormous boost since the emergence of the attention for ontologies: formal hierarchical representations of the set of concepts that together constitute the knowledge structure for a specific domain. Techniques for the detection of content elements corresponding to concepts and their relations as well as concept instantiations (e.g., named entities) have gained popularity for all content modalities. For the detection of conceptual elements a variety of techniques can be deployed, among which are machine learning approaches that can help to capture the statistical relation between low-level content features and the higher level conceptual structure.

For non-linguistic content, concept detection tools are a key contributor to conceptual search based on automatically generated annotation [5], [6]. For spoken word content, semantic search is of course feasible insofar as the transcripts contains words that are suited as search terms, such as proper names. As is the case with textual content, speech transcripts can also be the basis for named entity extraction, automatic classification or even summarization [7], [8]. In addition, conceptual annotation layers can be generated based on the labeling of segments with speaker identities, and affect labeling. Coupling of these types of labeling to the labeling based on the detection of concepts in the content layers for other modalities (e.g., image, text) has only recently become a topic of research. As shown by initial experiments conducted in the context of IST project MESH[3] the number of the parameters involved (length of segment, density of speech track, voice-over *versus* dialogue, user preference for results in image or speech content) requires a very careful design of the experiments in order to draw any conclusion on how to set up the fusion of retrieved multimodal search results.

A. Surprise data

As is widely cited, based on experimentation on a corpus of broadcast news, spoken document retrieval was declared a solved problem in 2000 [2]. However, this "mission accomplished" proclamation can be considered to have over-generalized limited domain results. Outside the broadcast news domain, the exact features of the audiovisual data are often unclear and the data may be far more heterogeneous in nature than those usually seen in the laboratory. Moreover, annotated sample data resembling the audio conditions in these data are typically not available via the usual channels. As a result, the accuracy of automatic transcription will often be lower then the accuracy obtained in the BN domain. We therefore refer to these as '*surprise data*'.

A content set that can be regarded as an instance of surprise data that has been studied the past years in the TRECVID community (2007-2009) is the Academia[4] collection, provided by the Netherlands Institute for Sound and Vision. The Academia collection represents video from a real-life archive and consists of some 400 hours of Dutch news magazine, science news, news reports, documentaries, educational programmes and archival video. The audio and speech conditions vary enormously and range from read speech in a studio environment to spontaneous speech under degraded acoustic conditions. Furthermore, a large variety of topics are addresses and the material dates from a broad time period, and the collection contains historical items as well as contemporary video. (The former with poorly preserved audio; the latter with varying audio characteristics, some even without 'intended' sound, just noise.)

To reach speech recognition accuracy that is acceptable for retrieval given the difficult conditions, the various com-

[2] e.g., The Linguistic Data Consortium (LDC)

[3] http://www.mesh-project.eu/

[4] http://www.academia.nl

ponents of an audio processing system—including both several kinds of pre-processing steps, such as speech/non-speech detection, speaker segmentation, and language detection, and speech-to-text transcription—must be made robust against problems that typically emerge when technology is transferred from the lab and applied in a real life context, and against mismatches between training and testing conditions.

At the University of Twente, work has been carried out to develop robust, open-source ASR technology that can be deployed in unknown domains without the need for expensive manual annotation work for system training and without extensive manual tuning [9]. However, as will be discussed in the next section, it can be worthwhile to explore how existing textual information sources that encompass collections can be exploited for generating time-labeled semantic descriptions.

B. Exploitation of collateral data

The 'surprise data' problem discussed above constitutes a major obstacle towards forms of access that require time-labeled annotations. Next to the development of a robust speech recognition system, one of the strategies to overcome the low accuracy of speech transcripts in certain domains is to make smart use of available resources, such as descriptive metadata or collateral data, to provide useful annotations based on the speech in collections.

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

Despite the fact that metadata and collateral text data can be formally differentiated, collateral text data may show great overlap with content descriptions that are part of the metadata. Collateral text may also be used to generate metadata descriptions, but once these have been created, the multimedia documents and those collateral data sources become separate objects again. Take, for instance, subtitling information for the hearing-impaired (e.g., CEEFAX pages 888 in the UK) that is available for the majority of contemporary broadcast items, at least for news programs. Subtitles contain a nearly complete transcription of the words spoken in the video items, and provide an excellent information source for automatic indexing. Textual sources that can play a similar role are teleprompter texts—also referred to as auto-cues—read from a screen by an anchor person. Although teleprompter texts are usually an accurate representation of the anchor person's speech (with an accuracy measured on a Dutch collection of around 90% they often do not include transcripts of interviews and dialogs with on-site reporters. At the Netherlands Institute for Sound and Vision, subtitles for the hearing impaired are currently physically linked to incoming broadcasts when available.

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

In sum, for many spoken word documents there is some kind of collateral text data available. Two remarks have to be made however. First, availability is often a relative concept. Making several metadata and/or collateral data streams available for a proof-of-concept demonstrator as in the broadcast news cross-media browser mentioned in section IV-A, figure 1 below, restructuring a complex real-life workflow for a running application, possibly involving commitment from multiple branches or even companies, is something else. Second, the level of similarity between the collateral data and the speech, may differ. At one end of the spectrum are the full transcripts of the spoken content, and via extensive summaries or documents laying out the linear structure, such as slides or agendas, at the other end of the spectrum there we find the textual documents that relate only generally to the semantic themes of the spoken content.

C. Alignment

Alignment is the process of using an ASR system to recognize an utterance, where the words occurring in the utterance, but not their timing, are known beforehand. The result is a set of time-aligned word labels. Alignment is a well-known procedure used frequently in ASR, for example when training acoustic models. It applies best when available transcripts closely follow the speech as it was found in the data, such as can be the case with accurate minutes from a meeting, although it holds for surprisingly low text-speech correlation levels as well, especially when some additional trickery is applied. When the available data allows for the successful application of the alignment strategy, alignment has a number of benefits: it saves the development and tuning of collection-specific automatic speech recognition,

it is accurate (e.g., with respect to collection-specific words) and fast.

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

An example of an archive for which full-text alignment has been applied is the historical collection of radio speeches that Queen Wilhelmina of the Netherlands addressed to the Dutch people during World War II, the so-called 'Radio Oranje' collection. The 'Radio Oranje' project[5] aimed at the transformation of a set of World War II related mono-media documents – audio, images, and text – into an on-line multimedia presentation with keyword search functionality [10], [11]. The mono-media documents consisted of the audio of Queen's speeches, the original textual transcripts of the speeches, and a tagged database of WWII related photographs. Within the NWO-CATCH project CHoral a demonstrator search system was developed for this collection. The demonstrator is discussed further in section IV-B, here we discuss the alignment between the spoken audio recordings and the the original textual transcripts. The collection consists of 37 speeches with lengths varying between 5 and 19 minutes. Their style is very formal and language use is complex. The recordings as well as their 1940s transcripts have been digitized by the Netherlands Institute for War Documentation (NIOD) and the Netherlands Institute for Sound and Vision. The audio quality be considered poor; the recordings are noisy and contain artifacts (e.g., hiss, pops).

The alignment tool from an off-the-shelf multi-mixture, Gaussian HMM-based speech recognition engine was used, [12]. This produces Viterbi-optimized, word-based alignments. Optimal alignment performance was obtained using speaker-dependent, monophone acoustic models, trained from gender- and speaker-independent models optimized for broadcast news (see also [13]). Performance was adequate for this task: $>$90% of all word boundaries were found within 100 ms of the reference, i.e., within the correct syllable. On the basis of alignment an index was built which turned this historical collection into on online, searchable asset. This combination of functionalities offered is received with enthusiasm by users and since the launch of the demonstration website at the beginning of 2007 it has been visited over 1500 times.

In the case that the match between the collateral text and the spoken content is incomplete, the transcripts, such as meeting minutes, can still be automatically enhanced via the generation of time-stamps. As long as the text follows the spoken content well enough, the word-level alignment can be found by using relatively large windows of text. This alignment procedure works well even if some words in the minutes are not actually present in the speech signal. In case the speech-transcript correlation is low, applying a two-pass strategy, similar to the one proposed in [14] could be used. A baseline large vocabulary ASR system[6] is used to generate a relatively inaccurate transcript of the speech with word-timing labels. This transcript is referred to as 'hypothesis'. Next, the hypothesis is aligned to the minutes at the word level using a dynamic programming algorithm. At the positions where the hypothesis and the minutes match so called 'anchors' are placed. A match is defined as three correctly aligned words in a row. Using the word-timing labels provided by the speech recognition system, the anchors are used to generate segments. Individual segments of audio and text are accurately synchronized using forced alignment.

IV. Connectedness to multimedia content sources

In order to boost the interaction with audiovisual content, the connectedness to complementary multimedia content sources needs special attention. Currently there is limited connectivity between audiovisual content and correlated information. At the Netherlands Institute for Sound and Visionconnecting the archive to external information sources—referred to as *contextualisation* of the audiovisal archive—is being taken up via several threads as it is regarded as an important instrument to improve access and the potential for exploitation of the archive.

Sound and Vision curates and makes accessible a large audio/visual archive whose contents are collected from Dutch public broadcasters and other national media. The archive includes documentaries, films, commercials, political programs, music, and video from cultural and scientific organizations. Currently, the archive contains 700,000 hours of content and each year an additional 22,000 hours of radio and 8,000 hours of television, which represents the complete production of the Dutch public broadcasters, are added.

To link the archival content to relevant information sources (web data, program guides, content from trusted parties such as libraries) Sound and Vision is building a context data platform that consists of a content aggregator that collects information sources, an analysis layer that creates the links between various types of content sources, and a distribution layer that delivers the connected information to different types of consumers in the archive workflow, such as the professional archivist generating a description for an asset, a broadcast professional who is searching for material for his documentary, and the public user or researcher that is interested in the background of a video item. Speech

[5]Radio Oranje:http://hmi.ewi.utwente.nl/choral/radiooranje.html

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

recognition technology is another possible client of the context platform that requires context information to include video specific terms in the language model vocabulary.

A. Connecting speech to text

As an example of how speech recognition could be deployed for connecting audiovisual content with external information sources, we refer to an application that was developed at University of Twente to demonstrate on-line access to an archive of Dutch news broadcasts ('NOS 8 uur Journaal'). See also fig 1.

The linguistic annotations of news items (based on either subtitles or ASR) were linked to an up-to-date database of Dutch newspaper articles. We can use that database for demonstration purposes by courtesy of PCM publishers, one of the largest publishers in the Dutch language region. For copyright reasons, the public version of the demonstrator does not contain the links to these articles. The links from broadcast news fragments to related, i.e., collateral, newspaper articles are generated by (i) using a stopped version of the textual video annotation to query the newspaper archive, (ii) matching the query with the content in the newspaper archive using Okapi term weighting, [15], and (iii) presenting the top-n results in a clickable list, ordered by date or by relevance.

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

B. Connecting speech to images

A screenshot of the 'Radio Oranje' system is depicted in Figure 2. As already described in section III-C the system uses a word-level alignment of the original historical transcripts with the original audio recordings generated using speech recognition technology. The screenshot in Figure 2 illustrates the way in which the alignment allows the system to be able to (i) respond to queries by offering users relevant listen-in points, (ii) show the spoken words in the speeches

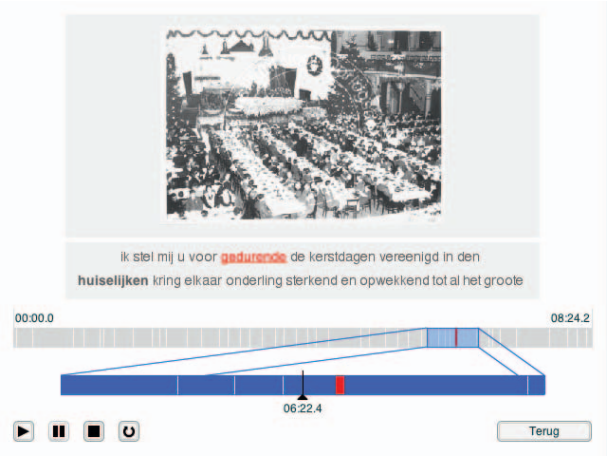

Figure 2. Screen shot of the 'Radio Oranje' application showing bars visualizing the audio (entire speech and current segment), subtitles (keyword in bold-face, current word underlined) and an image that relates to the content of the speech (in this case: Christmas).

as 'subtitling' during audio playback, and (iii) present changing images related to the content of the speech by matching the index words with the image tags from the database. Here we wish to use the system as an illustration of an outstanding real-world example of how speech recognition technology can be used in order to create a new setting in which a photo collection can be exploited for its historical value.

C. Documentalist support

Instead of regarding manual and automatic annotation as two separate threads, these can be interlinked. By automatically pre-selecting keywords for documentalists to add to content descriptions, the manual annotation process can be boosted substantially with respect to both quantity and quality of the descriptions. Keyword recommendation is part of a Documentalist Support System under development at Sound and Vision. On the basis of automatically aggregated context information and semantic analysis, this system suggests relevant keywords from the thesaurus (Common Thesaurus Audiovisual Archives, GTAA[7]) so that a documentalist can quickly pick terms from a relatively small list instead of going through a list of more than 150.000 terms.

There are two ways to find keywords for suggestion: (i) via the textual context of archived documents, and (ii) via the automatic analysis of the documents themselves by means of speech recognition technology. The latter approach has the benefit that it can also be used for audiovisual documents that cannot easily be connected to textual context on the web. A possible caveat when using speech recognition technology could be that transcription accuracy varies across collections. For certain collections, transcript error rates might simply be too high to be useful for keyword suggestion.

[7]http://informatieprofessional.googlepages.com/support

In a preliminary experiment using the ACADEMIA collection, lists of keywords were generated either from context documents or from ASR output. The unstructured lists coming from the individual processes were subsequently ranked with a ranking component, in this case the TF.IDF algorithm. Both sets of ranked lists were compared with a reference of manually assigned keywords. Although the context documents attained similar performance with much less information, the total amount of good suggestions contained by the ASR lists was larger. However, the number of wrong suggestions was also larger [16].

Sound and Vision makes it possible for the wider multimedia community to work on tasks related to the important issue of keyword recommendation through its support of the VideoCLEF video analysis and retrieval benchmark evaluation [17].[8] VideoCLEF is a track within the Cross Language Evaluation Forum (CLEF),[9] which promotes research in multilingual information access. Benchmark campaigns are important since they help to focus investigative effort on specific tasks, allow research teams to pool resources to carry out evaluation, and provide additional resources that lower the entry threshold for groups lacking specific infrastructure, and facilitating licensing of archive data for research purposes. The VideoCLEF task related to the keyword recommendation problem is called "semantic theme classification" and its goal is to group videos into classes each treating a specific semantic theme. Within the archiving workflow, each theme would receive a different keyword from the archivist. The semantic theme classification task in VideoCLEF is a manifestation of a larger interested in video classification, as reflected, for example, by the genre classification task set out by Google as an ACM Multimedia Grand Challenge for 2009.[10] Here, participants are encouraged to build systems that classify internet video into genre classes, for example, those used by the Open Directory Project.[11] This list includes many categories in which genre includes a strong topical component, including *Arts, Health, News, Travel, Science*. In all its guises, classifying video with respect to the semantic topic is a classic task with real-world relevance that is still in need of operationalizable solutions.

The VideoCLEF semantic theme classification task is still relatively young and is currently in the process of laying the groundwork for future growth. It was introduced in 2008, when it ran on 40 hours of video content (development set 11 videos/test set 40 videos) of Dutch-language documentaries and talk shows from Sound and Vision. The University of Twente provided speech recognition transcripts, using their open source recognition system [9], [18]. The teams that participated in the task were asked to tag videos with 10 thematic subject labels: Archeology, Architecture, Chemistry, Dance, Film, History, Music, Paintings, Scientific research and Visual arts. Participants were required to collect their own training data and most teams chose to use either general Web content or else articles from Wikipedia as a data source. Participants were given the opportunity to explore the potential of the multilingual nature of the data. The main language of the content in the Sound and Vision archive is Dutch, however, guests on talk shows and experts interviewed in documentaries often speak other languages and are not dubbed over. VideoCLEF participants in 2008 used English language transcripts to attempt to exploit the content spoken in the English language sections of the video content. Spoken-content-based subject classification stands to benefit from the focus and momentum provided by a benchmark task such as VideoCLEF. Each year, VideoCLEF increases the amount of video data and the number of subject keywords used and attempts to encourage participants to use collateral data and innovative features, including features drawn from the visual channel, or characteristics of the audio not related to its spoken content, such as a presence of music or applause.

[8]http://www.cdvp.dcu.ie/VideoCLEF/
[9]http://www.clef-campaign.org/
[10]http://www.acmmm09.org/MMGC.aspx
[11]http://www.dmoz.org/

V. SEMANTIC SPEECH RETRIEVAL IN THE ARCHIVING WORKFLOW

There are a number of important respects in which automatic annotations fall short of providing fully effective support for multimedia access tasks. The first problem relates to the so-called semantic gap. Because automatically generated annotations are not associated with fully explicit human interpretations, certain abstractions cannot easily be made. In the context of semantic speech retrieval, the semantic gap manifests itself in the mismatch between the actual words that are being spoken and the more abstract semantic concepts that are being talked about. For an example, a discussion can treat the topic of civic responsibility, but mention only examples of specific commitments to the community, such as to a church or school, and the activities through which these commitments are fulfilled, such as baking and attending meetings.

To reduce the semantic gap between user queries and indexes, there is a well-established range of generally applicable methods to turn text-based indexing into something that can be considered to support automatic semantic annotation, e.g., topic clustering and automatic classification. The application of these techniques in multimedia archives is also expected to improve accessibility, cf. [7], [8], [19] for an overview of techniques.

The second problem concerns the fact that a large number of user queries to multimedia collections involve named entities, such as personal names and locations. Especially named entities run the risk of being out-of-vocabulary, i.e., not being present in the current vocabulary of the speech recognition system, which means that they cannot

be transcribed and do not appear in the speech recognizer generated transcripts. If textual sources for tuning vocabularies to specific collections are not available, these out-of-vocabulary terms may be irretrievable. One way of reducing this problem is by carefully annotating at least the names of places and persons that are associated with a certain multimedia document during the description process. Usually, metadata models encompass standard fields to enter such information. Alternatively, collateral data such as scripts and notes from producers of multimedia documents may provide this information.

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

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

Finally, automatically generated annotations often fail to capture aspects of multimedia documents that are important for retrieval, but are not directly related to the document's declarative informational content and therefore can be considered as "orthogonal to topic." In the case of spoken word collections, we distinguish two different dimensions along which automatic techniques can generate annotations with potential to provide support for multimedia content access are being developed. First, methods can be developed that annotate documents with respect to their affective content, the level of emotion experienced or evoked by the speaker. Second, methods can be developed that attempt to predict the appeal that the document will have for users by automatically analyzing the quality and style of the delivery and the production of the spoken content. These two dimensions are discussed in greater detail in the next section.

VI. Affect and Appeal

The Sound and Vision archive has a well-developed understanding of the archive needs that should be satisfied if the state of the art of spoken-content-based multimedia access is to be pushed to the next level. One of the new directions that Sound and Vision is interested in moving in the future is moving beyond the thematic content of the video and analyzing video with respect to characteristics that are important for viewers, but not related to the video topic. Information orthogonal to topic makes it possible to recommend or rank videos or video segments that are equivalent with respect to their declarative content, but are not equally valuable in the information searchers.

Sound and Vision promotes research in the area of analyzing helpful non-topical dimensions of video content by supporting the VideoCLEF task "Affect and Appeal." This two part task is currently in its pilot year. This year the tasks are carried out on data from *Beeldenstorm* (Eng. Iconoclasm), a Dutch-language documentary series on the visual arts. *Beeldenstorm* is hosted by Prof. Henk van Os, known for his expertise and insight, but also his narrative ability, which creates both "humorous" and "moving" moments according to descriptions of viewer's opinions of the series.

The "Affect" task is defined in a straightforward manner: design a narrative peak detector, an algorithm that tries to identify these moments in the same way that a human viewer would perceive them. The task requires participants to first think about what constitutes a narrative peak and which possible indicators could be used to detect one. In order to simplify the task, we constrain the task data to *Beeldenstorm* data, with the assumption that a single speaker produces narrative peaks with more easily generalizable characteristics. Then, participants build systems which automatically detect narrative peak indicators in the speech stream and predict the positions of peaks. System output is evaluated by comparing the detected peaks with human opinion on peak position. The narrative peak detector is intended to be one building block within a larger system that exploits affect detection for the support of multimedia access. The frequency of peaks is expected to be an indicator of where in the episode the most insightful most exciting information is presented. Whether or not this expectation will be fulfilled can only be determined by appropriate experimentation.

The "Appeal" task involves designing a system that makes a prediction about the level of preference a particular *Beeldenstorm* episode enjoys among viewers. This task provides participants with an opportunity to apply methods such as those that have been used for preference predictions with podcasts [20], [21] to the area of video. Again, task evaluation is carried out by comparing system output to human judgments. The task is challenging due to the difficulty of pinpointing the factors contributing to user preference and also accommodating the fact that viewer likes and dislikes

are subject to the vagaries of style and sometimes appear to be entirely quirky. In order to integrate information derived from automatic annotation involving appeal into a system, it is not, however, necessary for prediction to be perfect. The prediction can be used to prioritize items for viewer review rather than to exclude items completely from consideration.

VII. Conclusion

This paper has overviewed techniques and technology that apply automatic speech recognition to the improvement of meaning-based access to multimedia collections containing spoken content. Through the descriptions, examples and observations presented here a reserved, but positive picture emerges. We do not predict that semantic speech retrieval solutions will become ubiquitous in the immediate future, but rather we identify places in which their sophistication and influence is extending. In particular, speech recognition technology is helpful when it is used to support existing workflow steps, such as archivist keyword assignment, or enhance existing collections – such as enabling integrated access to otherwise isolated mono-media resources.

Acknowledgment

This paper is based on research that was partly funded by IST project MESH (http://www.mesh-ip.eu) and by bsik program MultimediaN (http://www.multimedian.nl). The third author acknowledges the EU-FP7 PetaMedia Network of Excellence (http://www.petamedia.eu).

References

[1] D. P. Jonathan, J. G. Fiscus, J. S. Garofolo, A. Martin, and M. Przybocki, "Broadcast news benchmark test results: English and non-english word error rate performance measures," in *Proc. DARPA Broadcast News Workshop*, 1998, pp. 5–12.

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

[3] M. Siegler, "Integration of continuous speech recognition and information retrieval for mutually optimal performance." Ph.D. dissertation, CMU, 1999.

[4] R. P. Yu, K. Thambiratnam, and F. Seide, "Word-lattice based spoken-document indexing with standard text indexers," in *2008 SIGIR Workshop on Searching Spontaneous Conversational Speech*, Singapore, 24 July 2008, pp. 1–5.

[5] M. S. Lew, N. Sebe, C. Djeraba, and R. Jain, "Content-based multimedia information retrieval: State of the art and challenges," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 2, no. 1, pp. 1–19, 2006.

[6] J. Yang and A. G. Hauptmann, "(un)reliability of video concept detection," in *CIVR '08: Proceedings of the 2008 international conference on Content-based image and video retrieval*. New York, NY, USA: ACM, 2008, pp. 85–94.

[7] L.-S. Lee and B. Chen, "Spoken document understanding and organization," *Signal Processing Magazine, IEEE*, vol. 22, no. 5, pp. 42–60, Sept. 2005.

[8] K. Koumpis and S. Renals, "Content-based access to spoken audio," *Signal Processing Magazine, IEEE*, vol. 22, no. 5, 2005.

[9] M. Huijbregts, "Segmentation, diarization and speech transcription: Surprise data unraveled," Ph.D. dissertation, University of Twente, November 2008.

[10] W. Heeren, L. van der Werff, R. Ordelman, A. van Hessen, and F. de Jong, "Radio Oranje: Searching the Queen's speech(es)," in *Proceedings of the 30th ACM SIGIR*, 2007.

[11] R. Ordelman, F. de Jong, and W. Heeren, "Exploration of audiovisual heritage using audio indexing technology," in *Proceedings of the first Workshop on Intelligent Technologies for Cultural Heritage Exploitation*, 2006.

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

[13] L. van der Werff, W. Heeren, R. Ordelman, and F. de Jong, "Radio Oranje: Enhanced access to a historical spoken word collection," in *Proceedings of CLIN 17*, 2007.

[14] P. Moreno, C. Joerg, J.-M. V. Thong, and O. 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.

[15] S. Robertson, S. Walker, M. Beaulieu, M. Gatford, and A. Payne, "Okapi at trec-4," 1996, text REtrieval Conference.

[16] V. Malaisé, L. Gazendam, W. Heeren, R. Ordelman, and H. Brugman, "Relevance of ASR for the automatic generation of keywords suggestions for TV programs," in *TALN 2009*, Senlis, june 2009 2009.

[17] M. Larson, E. Newman, and G. Jones, "Overview of VideoCLEF 2008: Automatic generation of topic-based feeds for dual language audio-visual content," in *Working Notes for the CLEF 2008 Workshop*, 2008.

[18] M. Huijbregts, R. Ordelman, and F. de Jong, "Annotation of heterogeneous multimedia content using automatic speech recognition," in *Proceedings of SAMT 2007*, vol. 4816, 2007.

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

[20] E. Tsagkias, M. Larson, W. Weerkamp, and M. de Rijke, "Podcred: A framework for analyzing podcast preference," in *Second Workshop on Information Credibility on the Web (WICOW 2008)*, ACM. Napa Valley: ACM, October 2008.

[21] E. Tsagkias, M. Larson, and M. de Rijke, "Exploiting surface features for the prediction of podcast preference," in *31st European Conference on Information Retrieval Conference (ECIR 2009)*, April 2009.