poster

agma

automatic generation of meta data for audio-visual content in the context of mpeg-7

Joachim Köhler, Konstantin Biatov, Martha Larson, Christian Eckes, Stefan Eickeler
Fraunhofer – Institut Medienkommunikation (IMK)
{Joachim.Koehler, Konstantin.Biatov, larson, christian.eckes, stefan.eickeler }@imk.fraunhofer.de

Abstract

This paper describes our MPEG-7 project AGMA. AGMA stands for Automatic Generation of Metadata for Audio-visual content in the context of MPEG-7. The goal of the 3-year project funded by the German Federal Ministry of Education and Research is to develop new methods for automatic content analysis and to integrate these methods in a MPEG-7 compliant retrieval system. AGMA is divided into four components that deal with speech recognition, music recognition, face recognition and system design. The content analysis methods are developed and optimized on a multimedia database of material recorded in the German parliament. The content analysis is based on statistical methods including Hidden-Markov-Models (HMM) and Neural Networks (NN), which extract features that characterize the data and can be encoded in the MPEG-7 framework. MPEG-7 is the new standard for multimedia description interfaces. One of its main components is the representation of audio-visual metadata in XML. The paper presents background information on media archiving, the MPEG-7 standard, speech and image recognition, describes the German parliamentary database and relates preliminary recognition results.

Keywords: MPEG-7, Spoken Document Retrieval, Face Recognition, XML Schema, Content Analysis, multi-modal

Project URLs:
http://www.3sat.de/nano/cstuecke/19078/index.html

Introduction

The enormous increase in the number of multimedia sources and the availability of documents requires automatic methods to segment and analyse staggering amounts of audio-visual data. Technologies like speech and image recognition make this goal attainable. Projects like Informedia [1] and the DARPA Spoken Document Retrieval tasks [2] show that it is still a challenge to transcribe broadcast data totally automatically. Methods for face recognition and object segmentation are still under development and huge improvements are required before this technology can be used in commercial applications in the field of media archiving. The media industry is demonstrating an acute interest in reusing broadcast content that is in storage in databanks. Automatic analysis methods would make it possible to archive, retrieve and filter audio-visual segments (for example, a special interview), objects (a special person) and events (a special goal in a football match). In addition to improvements in the underlying algorithms needed for automatic content analysis, it is important to find a common description interface to describe multimedia data. This is the focus of the new MPEG-7 standard, which supports interoperability between different archiving and retrieval systems. Although MPEG-7 specifies a long list of multimedia descriptors and description schemes, the project AGMA uses only a select few of them. The fact that we can implement our system with a subset of the total functionality reflects the modular nature of the MPEG-7 standard.

The paper is divided in 5 sections. First, we describe the multimedia database of the German parliament used in the project. Second, we summarise some principles of the MPEG-7 standard that are important for our project. Then we introduce our speech recognition system, which is used to transcribe speech of the German parliament. The fourth section contains some preliminary work in the area of object segmentation and face recognition. In the final section we present our MPEG-7 system architecture followed by an outlook detailing the direction the ongoing work.

1 The Database of the German Parliament

For the AGMA project we have collected speech and video data from the German parliament. Although other domains may have higher entertainment potential, German parliamentary sessions offer three important advantages. First, the German parliament is a naturally delimited domain. Second, the high volume of speech and video with the corresponding transcriptions and annotations necessary for system training is available. Third, the data is publicly available, and its use does not give rise to intellectual property issues.

Previous projects in the area of speech recognition and spoken document retrieval have shown that designing a retrieval system for unlimited domains is not a tractable undertaking. A system conceptualized to recognise the content of sport events like a soccer game as well as other events like a TV talk show will show a low performance because the content is to diffuse for the system to learn the necessary generalizatons. Domain limitation is an important aspect to build a system that can be used for real world applications.

The second advantage of the German Bundestag database is the generous amount of data which is available. Every year there are about 100 debates containing in average 6 hours of speech and video. Even though the last 3 years are only partially available, we are able to exploit roughly 1000 hours of speech and video material. The speech and the video are provided in RealNetworks format. An additional advantage is the existence of annotations for the parliament speeches. The debates are tagged and annotated by humans and the speech is transcribed using stenography. For each debate there is a html-page and a electronic text file which can be downloaded from the Internet. The existing annotations help to train and evaluate a statistical recognition and classification system.

2 The MPEG-7 Standard

This section presents a brief introduction to the MPEG-7 standard and explains its import for the future of multimedia. More information can be found at the MPEG and MPEG-7 homepage [3] [4]. The major goal of MPEG-7 is to provide a framework for the description of multimedia content. Previous MPEG standards have focussed on audio and video coding to reduce storage requirements (e.g. MP3) and transmission issues (e.g. MPEG-2 for DVB). The huge amount of available media content requires metadata to access, reuse and to personalize this content. Like other MPEG standards, MPEG-7 aims to provide a basis for interoperability among different applications and vendors. The MPEG-7 initiative started in 1998 with the call for proposals. Within 3 years the MPEG group will have produced the ISO standard 15938 which is called MPEG-7 [5]. The components of the standards are:

System: This component specifies the architecture of the standard. It contains tools for transport, storage and synchronization between content and descriptors. It also contains tools for protecting intellectual property.

Description Definition Language: This component describes the language for defining descriptors and descriptor schemes. It is mainly based on XML schema defined by W3C [6].

Visual: This component contains descriptors related to video annotations, such as color histogram, texture information, shape information.

Audio: This component contains descriptors related to audio, like waveform and spectral envelope, melody description and spoken content description.

Multimedia Description Schemes: This component contains information about high level metadata, such information about media format, date of creation, IPR, financial, structure of the content, etc.

Reference Software: This component contains the reference software, which is mainly the eXperimental Model. Here, all descriptors are integrated in a full system used to test the MPEG-7 components.

Figure 1: Focus of MPEG-7

Conformance: This component contains guidelines and procedures for testing conformance of MPEG-7 implementations.

MPEG-7 is not primarily focussed on automatic methods to annotate and recognise audio-visual input. Rather, it is meant to provide a common description of the syntax and semantics of audio-visual scenes. This principle is shown in figure 1.

The scope of MPEG-7 covers the definition of a standard for the description. Feature extraction, which is either an automatic procedure, or the work of human annotators, and search and retrieval are not part of the standard. These processes can be implemented and realized in different way by different research groups and vendors.

An important role of MPEG-7 is related to XML schema, which is used for describing multimedia content. Although schema does not cover all functionality required by MPEG-7 (i.e. matrix and vector definitions), it is a promising standard of W3C, which is used in many database systems and Internet applications. The decision to use XML will help to integrate MPEG-7 into existing database applications.

3 Speech Recognition and SpokenContentDS

For automatic segmentation and transcription of speech data we use the speech recognition system from Mississipi State University, called the ISIP system [7]. It is a state-of-the-art speech recognition system based on Hidden-Markov–Models (HMM) and is distributed under an Open Source license agreement. The whole system is implemented in C++ and contains tools for HMM training and large vocabulary decoding. The acoustic models are trained on 33.000 German sentences from the Phondat and Siemens100 databases yielding in 48 monophone models. Each model consists of 3 states with 8 mixture components. The lexicon we use for recognition contains 50.000 word entries. Higher level constraint on the recognizer is provided by a bigram language model trained with the SRI language modeling toolkit (http://www.speech.sri.com/projects/srilm/). Currently we use the ISIP system to transcribe the existing German parliament speeches to produce the training material needed

to adapt the acoustic models to the domain. The main issue is to process the imperfect transcription of the stenographs. We generate a word grammar which allows the insertion and deletion of words given in the stenography. In a second processing step, we apply confidence scoring to identify segments of speech which have been transcribed with a high degree of reliability and which can be used for retraining the acoustic models. In a first evaluation we achieved a recognition rate of 65% for the German parliament domain tested on 400 sentences[8].

MPEG-7 provides the SpokenContentDS to describe the content of speech material using an existing speech recognition engine, like Viavoice, HTK or ISIP. The SpokenContentDS contains a description of a word and a phoneme lattice to allow retrieval on a lattice. This descriptor scheme was proposed and developed by [9] and is now part of

[10]. A lattice is a graph containing recognition alternatives weighted by probabilities. Figure 2 shows a typical lattice with similar words:

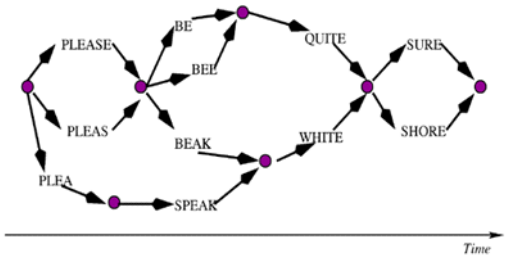

Figure 2: An example of a lattice for the SpokenContentDS (from [9])

Previous investigations on information retrieval have shown that the use of a lattice improves the performance of the system. Further, the phoneme lattice allows identification of words that are unknown to the recognition system, but that occur in retrieval queries. Proper names and new technical terms are particularly likely to be missing from the recognizer's vocabulary.

The SpokenContentDS is fully integrated in the MPEG-7 XM-system. Currently we are working on converting the ISIP recognition output to the format of the SpokenContentDS and combining this with the Transcriber toolkit [11] to annotate speech material. This system uses also XML and DTD to describe the segmentation and transcription of a spoken document.

4 Video Proccessing and FaceRecognitionDS

Although this work is in preliminary stage, we present our approach for video segmentation and face recognition. The video analysis is composed from three successive processing steps: temporal segmentation, face detection and tracking (figure 3), and face recognition. The temporal segmentation uses the difference image and the difference of the histograms of two consecutive frames to find the scene boundaries (cuts and dissolves) in the video sequence. A frontal face detector is used to find the faces in the scenes. The gaps between the detected frontal views of the face, where the person looks to the left or right, are filled using a simple histogram based tracking algorithm. The recognition of the detected faces is based on pseudo 2D Hidden Markov Models. One model trained on face images of each member of the German Bundestag is used to recognize the speaker at the parliament.

Figure 3: Face tracking result for a single speaker

5 AGMA System Design

The system design in AGMA is closely related to the design of the MPEG-7 XM (eXperimental Model). Because there are more than 100 different DS in MPEG-7 we have focused on those that are relevant to our goal of transcribing and annotating speeches of the German parliament. For example, the MelodyDS is not required for the German parliament domain. The most important DSs are the SpokenContentDS and the FaceRecognitionDSs. Both DSs are high-level descriptor schemes that contain semantic information already high level. Information on this level is important to do meaningful retrieval and have an effective application.

The components of the systems are shown in figure 4.

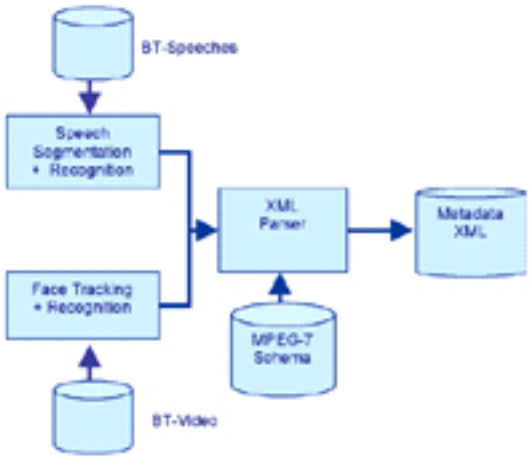

Figure 4: content extraction architecture in AGMA

The feature extraction process is divided in two stages. The speech recognition system generates a segmentation and transcription of the German parliament speeches. The output of the recognizer is a word and a phoneme lattice. In a simple version only the first best hypothesis is generated. The recognition engine is the ISIP speech recognition toolkit described in section 3. The output of the recognition process is passed to the XML parser. Here we use the Xerces parser from Apache, which is also used in the MPEG-7 XM system. The parser creates a DOM tree and is able to check the MPEG-7 schema. This step is an important stage in the process because here the recognition output format is validated. The validation guarantees conformance with the MPEG-7 standard and the interoperability with future applications.

The similar approach is carried out for the face recognition process. Before the speakers are recognized, a face finder locates the position of the face. Then the recognition process is initiated and the results are also passed to the XML parser. The results of the speech recognizer and the face recognition system are written to a XML file that contains the metadata of the audio-visual input file. This

metadata can be used for retrieval and information filtering applications. Although most work should be done by automatic methods like speech and face recognition a human annotation and verificatoin is necessary. We use for the speech domain the Transcriber Toolkit to check the metadata and to improve the transcriptions. In the step, confidence values are essential to identifying regions which are not reliably transcribed using automatic methods. Exploration of the potential of semi-automatic methods is an critical research direction in the area of metadata generation.

Finally, the combination of visual and spoken annotation are important to improve recognition and transcription rate. The synergies that can be exploited under a multi-modal approach will be a key research issue in the project.

7 Conclusion and future work

In this paper we have presented our MPEG-7 reference system designed for the German parliamentary domain. The combination of speech and face recognition yields in a powerful system for automatic annotation of multimedia data from a specific domain. Our AGMA project is one the first projects covering all parts of the MPEG-7 standards except for binary presentation and streaming of metadata.

Our future work will focus on a combination of video and speech recognition in a multi-modal system. We will also concentrate on object and face segmentation in order to be able to extract and reuse the images of different politicians. We have committed ourselves to improving the overall performance of our system to the level necessary to fulfill the demands of a real-world application.

8 Acknowledgement

We thank the ISIP team for providing the speech recognition system and Phil Garner for his work on the SpokenContentDS. This research is supported by the German Federal Ministry of Education and Research (BMBF) and has the official project number: VFG0011B.

9 Literature

[1] Hauptmann, A.G. and Witbrock, M.J.: Informedia: News-on-demand multimedia information acquisition and retrieval. In M. Maybury, Ed., Intelligent Multimedia Information Retrieval, AAAI Press, 1997.

[2] Garofolo et al.: 1998 TREC-7 Spoken Document Retrieval Track Overview and Results, Proc. Spoken Language Systems Technology Workshop, 1999

[3] MPEG-URL: http://www.cselt.it/mpeg/index.htm

[4] MPEG-7 URL: http://www.mpeg-7.com/

[5] Jose M. Martinez: Overview of the MPEG-7 Standard, ISO/IEC JTC1/SC29/WG11 N4031, 2001

[6] W3C Recommandation: XML-Schema Part1 + Part2, http://www.w3.org/ ,May 2001

[7] Picone et al.: A Public Domain Decoder for Large Vocabulary Conversational Speech Recognition, http://www.isip.msstate.edu/resources/technology/projects/1998/speech_recognition, Mississippi State University, 1998.

[8] M. Larson, D. Willett, J. Koehler, G. and Rigoll, Compound splitting and lexical unit recombination for improved performance of a speech recognition system for German parliamentary speeches, ICSLP 2000.

[9] Charlesworth, J. P. A. and Garner, P. N.: Proposal for MPEG-7 Descriptor: Output of an automatic speech recognition system, ISO/IEC JTC1/SC29/WG11 – MPEG Proposal M4458, Seoul, 1999

[10] MPEG-7 Audio – ISO/IES CD 15934-4 Multimedia Content Description Interface – Part 4: Audio, 2001

[11] C. Barras, E. Geoffrois, Z. Wu, and M. Liberman : Transcriber: development and use of a tool for assisting speech corpora production Speech Communication special issue on Speech Annotation and Corpus Tools, Vol 33, No 1-2, January 2000.