PANOPTIS: a System for Intelligent Monitoring of the Hellenic Broadcast Sector

Iason Demiros[1,3], George Carayannis[1], Vassilios Antonopoulos[1,3], Georgios Kambourakis[2], Vassilios Katsouros[1], Panayotis Kolevris[1], Marios Nottas[1], Harris Papageorgiou[1,3], Vassilios Papavasiliou[1], Spyros Raptis[1], Fotini Simistira[1], Themos Stafylakis[1]

[1]Institute for Language and Speech Processing, Artemidos 6, Marousi 151 25, Athens, Greece
{iason, gcara, vantonop, vsk, pkolevr, manottas, xaris, vpapa, spy, fotini, themosst}@ilsp.gr
[2]National Technical University of Athens, Iroon Polytechniou 9, Athens, 157 73, Greece
gcamb@cs.ntua.gr
[3]Qualia Technologies of Understanding, Voriou Ipirou 4, Kifisia 14564, Athens, Greece

Abstract

In this paper we describe a system that applies emerging technologies for speech recognition, language processing, multimedia indexing and retrieval, all integrated into a large video and audio library that covers broadcast news and current affairs in Greece. It assists the Greek National Council for Radio and Television (NCRTV) in compiling information, annotating and analyzing news and monitoring national, political, social, economic, cultural and environmental issues concerning Greece in general. It further assists supervision of the broadcast A/V sector by offering citizens an efficient way to seek and get hold of, an 'official' copy of aired programming, in order to safeguard their interests and promote NCRTV's goals.

1. Introduction

The advent of multimedia databases and the popularity of digital video as an archival medium poses many technical challenges and has profound implications for the underlying model of information access. With the recent advances on Natural Language Processing, Speech Recognition and Video Processing, and the synergy obtained by seamless integration of different technologies into a single multimedia database, the potential of a large collection and further processing of broadcast news (as well as other types of programming) can be explored. We incorporate an extensive set of multimedia processing engines in a project that we are developing for the Greek National Council for Radio and Television (NCRTV), with the hope that it will be a powerful tool in promoting the reuse of existing resources, in automatically creating metadata for indexing and retrieval from a large internal multimedia archive, in gaining full strategic value from the inherent value of media assets and in supporting users within the Council in their analysis and research tasks. Future academic needs are considered, as well. Content processing involves automatic speech recognition, speaker identification, speech synthesis and finally video text detection. Moreover, expert users manually produce metadata of different sorts, such as transcriptions, summaries, named entity and term indexes, etc. Finally, users can query manually or automatically annotated stories with the help of an ergonomic integrated environment for media annotation. Universities and private companies have worked together to achieve this sort of goal under a unified framework, to provide coordinated e-government solutions and an integrated service provision. PANOPTIS combines speech and language processing, multimedia retrieval and a fully functional video annotation environment that assist the users in organizing and characterizing news stories and current affairs.

2. Targeting e-Government activities

Electronic Government (e-Government) is a system within Public Administration composed of a mixture of political objectives, organizational procedures, information content delivery and ICT technologies [1].

It provides a modern way to serve its target group with electronic services via alternative channels in a way more efficient than traditional bureaucracy. Its target group can be any of the following: citizen, businesses, government, civil servants and the services provided are Government to Citizen (G2C), Government to Business (G2B), Government to Government (G2G) and Government to Employee (G2E) correspondingly.

NCRTV collects information from a wide variety of news sources in Greece. Journalists, analysts and political scientists process and assess national news, photos, as well as radio and television material. Among the various areas of responsibility of the NCRTV, our focus was on the following activities:

1. Formulating state policy and ensuring the adoption of the necessary legislative and prescriptive initiatives regarding the regulation of the wider sector of the mass media and helping shape and implement policies for providing all kinds of radio and TV services necessary to the Information Society.

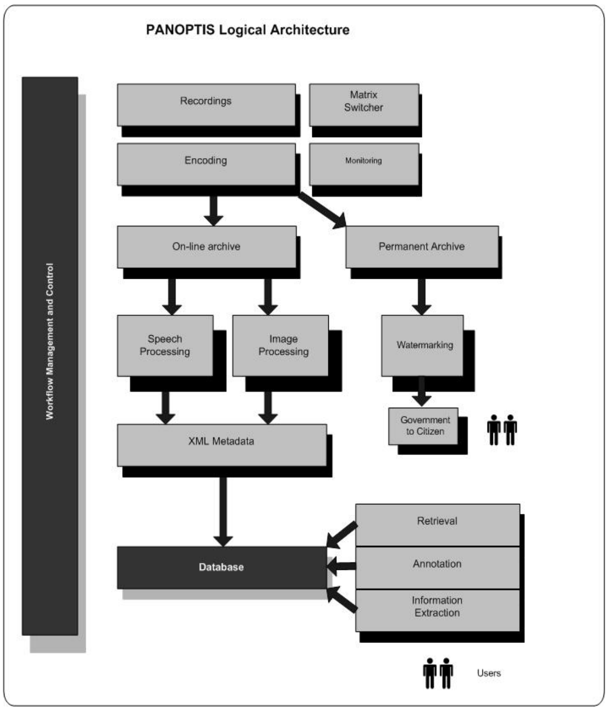
2. Collecting and making good use of data, relevant to the wider terms of reference of the NCRTV, especially in the field of national, political, social, economic, cultural and environmental issues concerning Greece, as well as international issues that are relevant to the country and/or the international bodies of which Greece is a member.

3. PANOPTIS: a content-based multimedia retrieval system

PANOPTIS treats continuous media as consisting of physical media accompanied by meta-information. Several classes of meta-information are identified in order to support flexible access and efficient reuse of continuous media. Meta-information is produced by a variety of tools working independently, making the integration of video and audio documents with their meta-information indispensable, in order to index the material and give the user the ability to browse, navigate and search the media database. All processing modules in the corresponding three modalities (Audio, Image and Text) converge to a textual XML metadata annotation scheme. These XML metadata annotations are merged and loaded into the PANOPTIS Multimedia DataBase. There exist two types of metadata generation tools used in our setting, for manual and automatic annotation respectively.

PANOPTIS integrates a user-friendly graphical environment for assisting the manual annotation of broadcast news recordings. Journalists, editors and political analysts can insert various levels of meta-information to the multimedia material. They can insert title, summary, full or partial transcription either in narrative or in dialog form, speakers and speaker turns, journalists and anchormen/anchorwomen, media information (channel, date, show, type of broadcast), sentiment classification (degrees from negative to positive), duration and a flag signing breaking news. They have direct access to any position in the video by several types of playback modes. Reports can be created containing the metadata that accompany the story, for immediate forwarding to the Council. Recordings can be managed from any workstation connected to the network. They can be scheduled in advance or programmed on-demand, as it is usually the case.

Regarding automatic annotation, the tools that we have developed during the project automatically populate the library and support access to it. The approach that we have followed uses large vocabulary automatic speech recognition, speaker identification, video text detection, speech synthesis and language processing technologies, in order to automatically transcribe, segment and index the video. Retrieval is text-based, performed on the material that results from intelligent processing, together with metadata created by the user. We present a logical diagram of the system in Figure 1.



4. Metadata extraction for intelligent video and audio indexing

4.1. Large Vocabulary Automatic Speech Recognition

Broadcast news exhibit a wide variety of audio characteristics, including clean speech, telephone speech, conference speech, music, and speech corrupted by music or noise. Transcribing the audio, i.e. producing a (raw) transcript of what is being said, determining who is speaking when, what topic a segment is about or which organizations are mentioned, are all challenging problems. We apply a highly accurate, speaker-independent speech recognizer in Greek, in order to automatically transcribe audio recorded from broadcast news which is then stored in a full-text information retrieval system [3].

The system is based on the Hidden Markov Model (HMM) technology [4]. It comprises three basic components, namely the audio signal processor, the audio segmentation component and the core speech recognition engine. The first component is responsible for the proper extraction of certain features from the audio signal which are then exploited by the following components. The audio segmentation module automatically identifies speech regions, rejects non-speech ones and clusters homogeneous regions of speech, i.e. same speaker, same background conditions, which carry the information to be decoded by the speech recognition engine. In this last component, speech content is automatically transcribed with a certain level of confidence, depending mainly on how well the models match the data being processed. Multiple passes with increasing complexity and speaker adaptation techniques are applied to upgrade performance. The overall speech recognition WER is 25% for decoding speed of 1,2xRT.

4.2. Speaker Recognition

We use the audio component of the video signal in order to identify each speech utterance according to a predefined dataset of target-speakers. The outcome is stored in both plain text and XML files, with the labeling of the regions that matched with a target speaker, as well as the total time of each speaker's utterances. In addition, speakers that appear in the broadcast news, who show but fail to match a speaker profile from the predefined dataset are also labeled, using symbolic names that define their gender, the recording conditions and speaker's serial number (e.g. UNKNOWN-F-W-03). The architecture of the unit is fully modular, while the algorithmic implementation is at the state-of-the-art.

A preprocessing module is responsible for transforming the waveform (in 16-bit/16-kHz sampling frequency format) into a domain that is adequate for speech/speaker processing tasks. We employ the Mel-Frequency Cepstral Coefficients (MFCCs), while the differentials of first and second orders of the MFCC are computed on-the-fly, when necessary. We then segment the signal in order to detect the time instances that a transition from one to another sound class takes place. The transitions include all possibilities, such as speaker-to-speaker, music-to-speaker, silence-to-speaker, etc. Those chunks that include speech are further classified according to gender and recording conditions (studio, outdoor, telephone), using Gaussian Mixture Model (GMM) classifiers. Instead of applying speaker identification directly to the chunks, the unit includes a blind clustering module. The role of this module is to group all chunks labeled as speech into clusters, so that a one-to-one mapping between the ground truth speakers and the chunk clusters is achieved. Having collected the maximum possible portion of each speaker utterances, we are ready to apply speaker identification to each participant. Our method is based on the state-of-the-art speaker identification/verification algorithms found in literature. We first built two Universal Background Models (UBMs), one for each gender, using 512-component GMMs [5]. Then, for each target-speaker in the dataset, we collect speech data under different recording conditions. We apply Cepstral Mean Subtraction to each distinct utterance and adapt the mean vectors of the corresponding UBM in a MAP-like fashion. Finally, the evaluation phase is based on maximum likelihood probability and hypothesis testing.

4.3. Video Text Detection

Recognized text extracted from newscast video sequences can be used to search for a specific topic. The Video Text Detection Unit (VTDU) of PANOPTIS focuses on localization, segmentation and recognition of artificial, horizontal and static text in video sequences. Such text occurrences mostly appear in newscasts, commercials, sports, etc. and are often the important carriers of information and herewith suitable for indexing and retrieval.

The VTDU consists of six sub-modules, each one of which exploits some characteristics of the text areas. The input of the VTDU can be a video document of MPEG-1, MPEG-2 or MPEG-4 format with resolution of 720x576 pixels. The output is a metadata file of the transcribed text in an XML format. The first sub-module deals with the sampling and aims to the construction of a frames sequence from the video document. Treating this frames sequence as a set of independent images, the next sub-module deals with the localization of the candidate text areas in each frame (Fig.2a). The algorithm involves operations like conversion to gray-scale, filtering by using a horizontal Sobel mask, binarization, iterations of dilation and erosion and merging of the candidate regions (Fig.2b-f). At the final step, a bounding box for every text-like area is extracted and a number of geometrical constrains are imposed, in order to eliminate small candidate text regions (Fig.2g-h). The algorithm locates successfully text with height ranging from 8 to 100 pixels. Next, a text verification sub-module aims at the elimination of the non-text regions (Fig 2i-j). The algorithm [6] explores the spectral properties of the horizontal projection of candidate text regions and by using a GMM classifier, discards the majority of the false alarms (non-text regions).

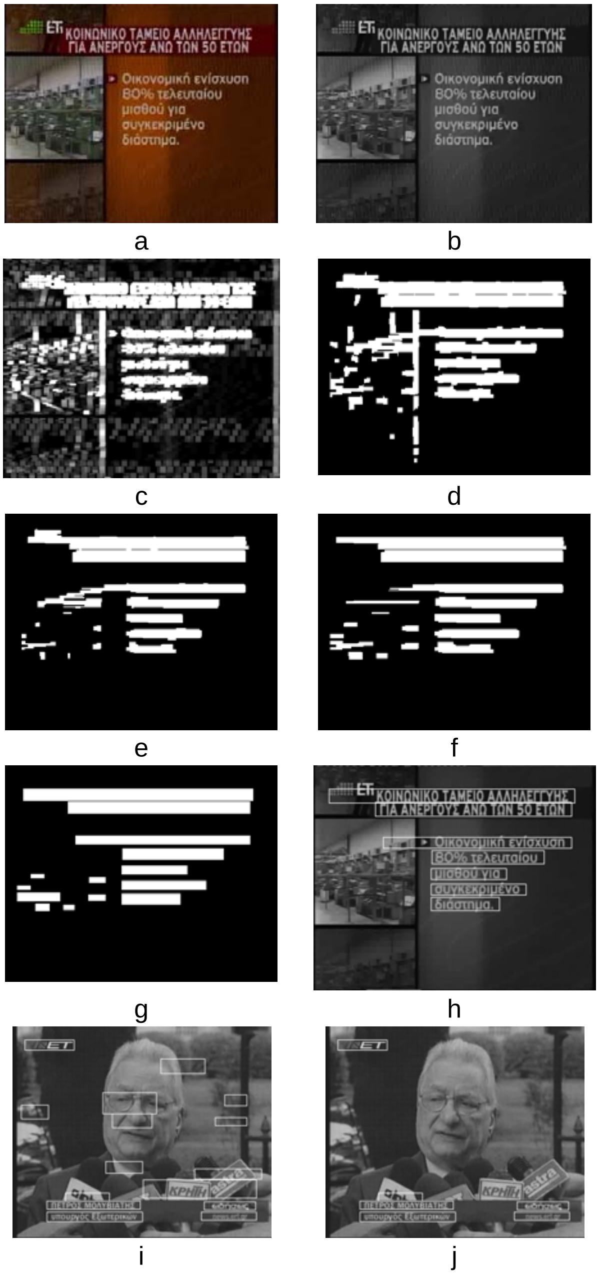


Figure 2. The intermediate results of the text localization process: (a) original, (b) gray scale, (c) filtered, , (d-g)

morphological processed, (h) final image, (i-j) false alarm reduction

Looking for similar text regions in successive frames, we define a measure of similarity that corresponds to an overlap of the text regions of at least 70% and a 2-D correlation coefficient that is greater than 0.8. Similar text images are enhanced using an averaging method.

The enhanced images feed the ABBYY FineReader OCR engine. The result of the character recognition and the corresponding log files are combined in order to create a metadata file in XML format. The VTDU has been tested explicitly in video sequences of newscasts monitored mostly from the two Greek public TV channels (ET1, NET). It scored a recall rate about 95% and a precision rate about 80%, which are at par with the results presented in [7].

5. An Every Citizen Interface to the System

A significant dimension of the PANOPTIS system is its interaction with its intended users. Aiming to provide services for a wider range of e-Government activities, PANOPTIS interacts with different users baring different roles. These can range from journalists and editors that insert, edit and update metadata, to analysts that request statistics, researchers seeking to interpret social traits and trends or even citizens that may access the system through the web to locate and retrieve information that may concern them. Whichever the mode of use and the medium employed to gain access, accessibility and usability are a valid concern.

Accessibility is not only related to the way that the interaction is driven and the information is presented to the user, but also to the modalities employed during the interaction and the ergonomies offered therein [8]. Often, speech can provide a useful way for conveying information and can complement or, in certain cases, substitute information in visual form. Speech is the preferred alternative for people with disability (for instance, blind people and people with visual impairments, people with dyslexia and so on), but also for contexts of use where visual information is harder to consume (for instance, when accessing the system's website through the tiny display of a web-enabled personal device). The integration of a high-quality, near-natural text-to-speech component in the user interface of PANOPTIS, can not only significantly enhance its accessibility by physically challenged people but can also make it possible for the system to offer advanced features such as spoken summary reports available through the telephone on a 24x7 basis, automatic telephone notifications on significant events identified by the system and so on. Advanced interface technology by ILSP and its spin-offs has been incorporated. Some samples of the high-quality speech engines that is used can be found at http://www.innoetics.com/tts.

6. Future Work

In this paper, we have described a media monitoring system that we have developed and implemented for the Greek National Council for Radio and Television (NCRTV). This project is a valuable experience that has provided us with a rich body of knowledge about the advances that are made possible by emerging speech, image and language technology. The success of the project lead NCRTV to make the decision to further develop it by applying information extraction technologies to media monitoring public services. NCRTV will continue to implement information technologies in order to improve its operations and services and to enhance the role that stems from its mandate.

7. References

[1] Aldrich D., Bertot J. C. and McClure C. R. 2002. “E-Government: Initiatives, developments, and issues”, In Government Information Quarterly, 19 4 (2002), 349--355.

[2] Jones, C.D., A.B. Smith, and E.F. Roberts, Book Title, Publisher, Location, Date.

[3] Nguyen L., Abdou S., Afify M., Makhoul J., Matsoukas S., Schwartz R., Xiang B., Lamel L., Gauvain J.L., Adda G., Schwenk H., and Lefevre F. 2004. “The 2004 BBN/LIMSI 10xRT English Broadcast News Transcription System”, In Proc. DARPA RT04, Palisades NY, November 2004.

[4] Papageorgiou H., Antonopoulos V., Demiros I., Gkiokas A. 2006. “Thematic Classification and Intelligent Indexing of Broadcast News Using Speech Recognition and Image Analysis”, EuroITV 2006, Athens, Greece.

[5] D. Reynolds, T. Quatieri, and R. Dunn, “Speaker verification using adapted Gaussian mixture models,” Digital Signal Processing, vol. 10, pp. 19–41, 2000.

[6] V. Papavassiliou, T. Stafylakis, V. Katsouros, G. Carayannis: A Parametric Spectral-Based Method for Verification of Text in Videos. ICDAR 2007: 879-883

[7] R. Lienhart, “Video OCR: A Survey and Practitioner’s Guide” in Video Mining, Kluwer Academic Publisher, pp. 155-184, Oct. 2003.

[8] W3C, Essential Components of Web Accessibility, http://www.w3.org/WAI/intro/components.php.