Media Monitoring by Means of Speech and Language Indexing for Political Analysis

Iason Demiros
Harris Papageorgiou
Vassilios Antonopoulos
Andreas Pipis
Athena Skoulariki

ABSTRACT. In this article, we describe a media monitoring system that we have developed and implemented for the Secretariat General of Communication and Secretariat General of Information in Greece (SGC-SGI). The system applies emerging technologies for audiovisual recording, 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 Greek and English. It assists SGC-SGI

Iason Demiros is a researcher at the Institute for Language and Speech Processing (ILSP) in Athens and a Ph.D. student at the Division of Signals, Control, and Robotics of the National Technical University of Athens (NTUA). His main research interests include information extraction, information retrieval, and machine learning. Since 2004, he has been the CEO of Qualia Technologies of Understanding.

Dr. Harris Papageorgiou is the Deputy Head of the Department of Nature Language and Knowledge Extraction at the Institute for Language and Speech Processing (ILSP), Athens, Greece, and a Chief Scientist at Qualia Technologies of Understanding, where he works on language and speech applications. His research interests focus on mathematics of language, grammatical inference, and machine learning.

Dr. Vassilios Antonopoulos is the Technical Manager at Qualia Technologies of Understanding. He has also been a researcher at the Institute for Language and Speech Processing since 1998, working mainly in the areas of automatic translation and multimedia information retrieval. He holds a Ph.D. in Language Modeling and Natural Language Processing from the National Technical University of Athens, Faculty of Electrical and Computer Engineering.

Dr. Andreas Pipis received his Ph.D. from the Department of Computer Engineering and Informatics of the University of Patras, Greece. He currently works as a Computer Engineer at the General Secretariat for Information Systems of the Greek Ministry of Economy and Finance. Since 2004, he has been the project leader of the Information Society Project Administration Group, of the Secretariat General of Communication-Secretariat General of Information.

Athena Skoulariki has taught at the Sociology Department of the University of Crete since the fall of 2006. She graduated from the Faculty of Philosophy, University of Athens and completed her postgraduate studies in Media Studies and Communication at the University of Paris 2 (Pantheon-Assas). In March 2005, she obtained her Ph.D. for the dissertation *In the Name of the Nation: Public Discourse in Greece on the Macedonian Issue and the Role of the Media (1991–1995)* [in french], University of Paris 2. She specializes in discourse analysis, nationalism, media, and politics.

The authors wish to thank Mr. Dionysios Gardelis from the Secretariat General of Communication (SGC) for his valuable comments and suggestions and Mr. Mikes Nitis from the IT Office of SGC, who had the technical responsibility of the project. The authors also wish to thank the various users of the SGC for their contribution to the success of the project.

Address correspondence to: Iason Demiros, Qualia Technologies of Understanding (E-mail: idemiros@qualia.gr).

Journal of Information Technology & Politics, Vol. 5(1) 2008
Available online at http://jitp.haworthpress.com

doi:10.1080/19331680802149632

in compiling information; annotating and analyzing news; and monitoring national, political, social, economic, cultural, and environmental issues concerning Greece in general.

KEYWORDS. Video annotation, speech recognition, multimedia information retrieval, e-government, political analysis, media content analysis

The advent of multimedia databases and the popularity of digital video as an archival medium pose many technical challenges and have profound implications on the underlying model of information access. Imagine a large collection of broadcast news and documentaries that supports content-based retrieval over multimedia objects such as audio and video. A user seeks information on a specific person or event; with the recent advances in natural language processing, speech recognition, and image processing, and the synergy obtained by seamless integration of different technologies into a single multimedia database, the potential of such a collection can be explored. We incorporated an extensive set of language processing technologies in a project that we have developed and implemented for the Secretariat General of Communication and Secretariat General of Information in Greece (SGC-SGI),[1] with the hope that it will be a powerful tool in promoting the reuse of existing resources, in gaining full strategic value from the inherent value of media assets, in retrieving audiovisual material from a large internal multimedia archive, and in supporting political scientists within the organization in their analysis and research tasks.

The role of the SGC-SGI is to inform state and public sector agencies on important events as well as views and reactions of Greek and foreign public opinion, including those of mass media, of issues affecting the country. SGC-SGI formulates state policy and ensures the adoption of the necessary legislative and prescriptive initiatives regarding the regulation of the wider sector of the mass media. Moreover, it collects data in the fields 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. SGC-SGI collects information from a wide variety of news sources in Greece and abroad. Journalists, analysts, and political scientists then process the news sources to assess national and international news and photos, as well as radio and television material.

Not surprisingly, media analysis for both scholarly research and political communication professionals requires sophisticated media monitoring systems in order to effectively collect, manage, and classify media content. In the case of a public communication administration, like the SGC-SGI, the challenge is even greater, given the amount and the variety of media data that need to be collected and analyzed. As a result, and as part of a larger electronic government (e-government) initiative, SGC-SCI works with mass-media authorities, universities, and private companies, within a unified framework for information access, to provide innovative electronic services (e-services) for both the government and the public. The media monitoring system that we describe in our article is a heavily used component of the framework, due, in part, to its ability to combine speech and language processing, multimedia retrieval, and a fully functional video annotation environment that assists users in organizing and characterizing news stories and current affairs.

QUALIARC: AN ENVIRONMENT FOR MEDIA ANNOTATION

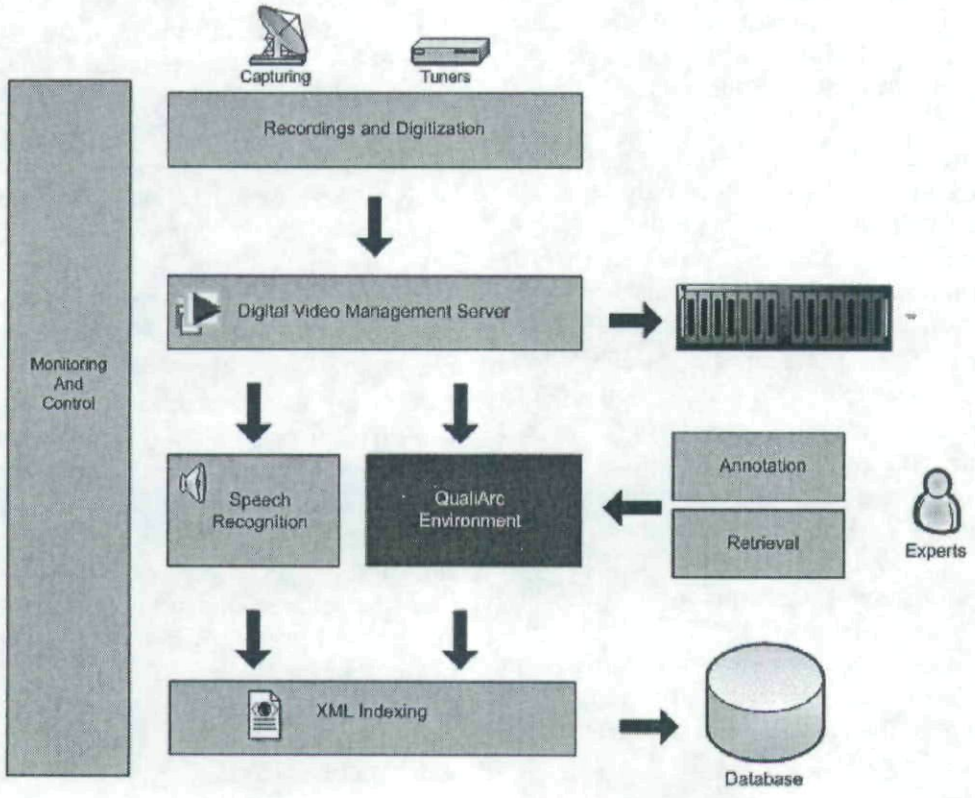
QualiArc is the integrated environment that we have developed for accessing and annotating multimedia digital content, providing advanced searching and browsing capabilities. We have successfully applied emerging technologies for data storage, indexing, and retrieval, and we have integrated them into a large video and audio library system that covers broadcast news and current affairs in Greek and in English. The SGC-SGI digital video archive uses intelligent

automatic mechanisms that provide full-content search and retrieval from a large online digital video library containing around-the-clock recordings from eight Greek TV channels, four Greek radio channels, and four international English speaking TV channels. There exist two types of metadata generation tools used in our setting, for manual and automatic annotation, respectively.

Manual annotation is time consuming and highly subjective, yet it provides broadcast news descriptions at the right level of abstraction. Specialized journalists that monitor the broadcasting channels locate and export TV video segments and radio audio segments in a predefined compressed format. They can monitor the broadcast in real-time or search in the audiovisual archive by means of automatically produced or manually inserted metadata. Segments correspond to news, stories, and events that are of interest to SGC-SGI, and they are stored in a large multiterabyte storage system in order to be annotated. Another team of journalists and political analysts reviews the stored segments and produces annotations while interacting with the QualiArc environment that we have developed within the project. Expert users produce metadata of different sorts, such as transcriptions, summaries, named entity and term indexes, etc. Annotated stories, accompanied by the corresponding video or audio, are stored in the system database.

Regarding machine 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 speech recognition in Greek and English and language processing technologies to automatically transcribe, segment, and index the video. Retrieval is text-based, performed on the material that results from speech recognition or on the metadata manually created by the user. In Figure 1 we present the logical diagram of the system.

FIGURE 1. Diagram and modules of the media monitoring system.



Semantic Annotation of News Stories

QualiArc is 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 a title, summary, or 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 signaling breaking news. They have direct access to any position in the video using several types of playback modes. A report can be created containing the metadata that accompany the story for immediate forwarding to the Minister of State or the Secretary General. Recordings can be managed from any workstation connected to the network. They can be scheduled in advance or programmed on-demand, as is usually the case. An example is the coverage of parliamentary elections in Greece that were held on September 16, 2007. It is noteworthy that the system was extensively used for monitoring both the elections and the electoral campaigns, since the Greek Prime Minister announced early general elections.

According to their user privileges, journalists and annotators can produce individual stories referring to the media that broadcast them. The same story can appear in different media with various degrees of sentiment expression, usually reflecting the beliefs of the different segments of the political spectrum. Political analysts aggregate individual stories in order to produce consolidated stories, where in fact they link source reference (and corresponding videos/audio recordings) that belong to the same entity (the task is technically called *source co-reference resolution*). They extract and report judgments, opinions, sentiments, and evaluations from stories. Fine-grain opinion extraction concentrates on words, sentences, clauses, polarity, and strength. A screenshot of the QualiArc environment is presented in Figure 2.

Metadata Scheme

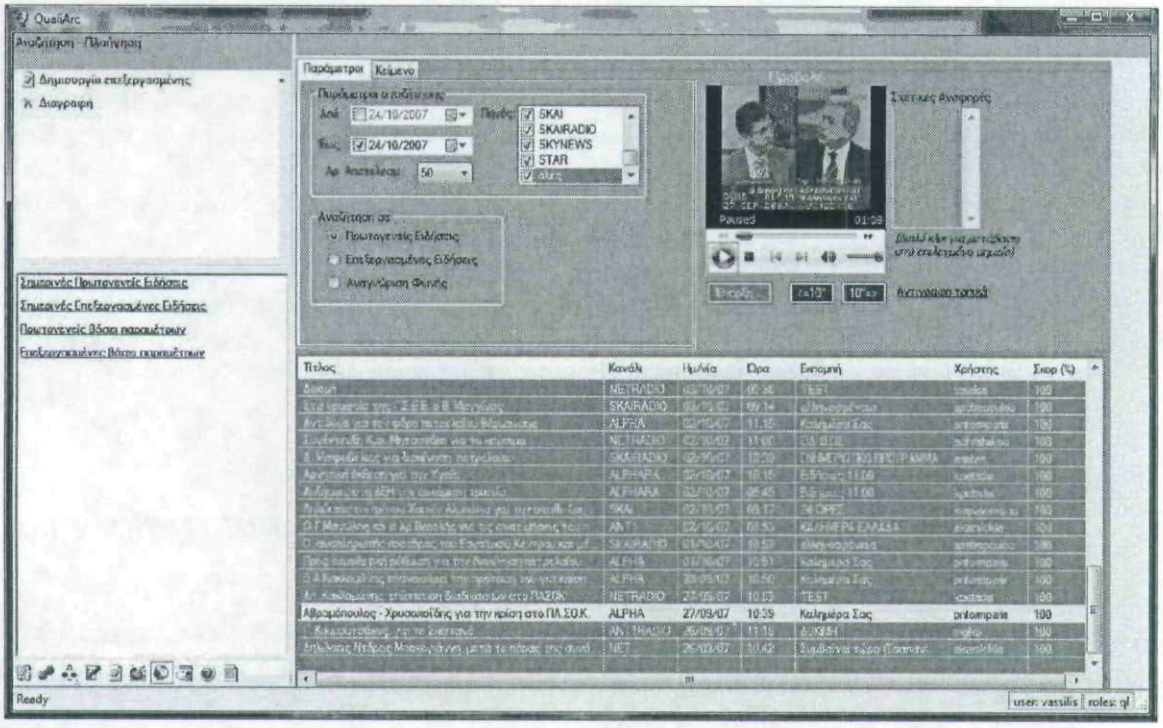
The core feature of the information system we have developed for the access and exchange of the news lies on the metadata scheme that we have devised for news documentation. The main requirements the scheme had to meet include the following:

(a) Encoding of different types of information pertaining to the form and content of news, as well as to administrative and technical information regarding its handling
(b) Encoding of information regarding the original media type or format
(c) Common representation for similar types of information (packaging of news information), regardless of the medium format of the item to which these are attached

In our case, the most appropriate scheme for news management is NewsML, a media-independent news exchange format,[2] catering not only for the core news content but also for data that describe the content in an abstract way, for information about how to handle news in an appropriate way, for information about the packaging of news information, and finally for information about the technical transfer itself. NewsML 1.2 acted as the source from which we derive our simpler scheme; to this end, we adapted and occasionally modified/expanded the NewsML metadata set (see example in Appendix B).

The metadata scheme is hierarchically organized; the *NewsML* element is situated at the top level while middle nodes are occupied by specific groupings: the *NewsEnvelope* element representing exchange information, the *NewsItem* element and its children nodes represent management information, the *NewsComponent* element and its children nodes represent structural information, and finally the *ContentItem* element represents content information. The metadata scheme has been implemented and incorporated in the QualiArc environment, and it offers an interface to all SGC-SGI political analysts enabling efficient metadata collaborative editing of digitized news resources and search thereof.

FIGURE 2. Screenshot of the QualiArc environment.



The purpose of NewsML is to make possible the exchange of news accurately and quickly in a way that allows automatic processing. The core technology is XML. The ContentItem element is a wrapper that makes available to the NewsML processor both the content itself and metadata describing the physical characteristics of the content. The different types of content metadata provided by NewsML are media type, format, mime type, notation, and a set of characteristics. The DataContent element represents the textual annotation of the body of the story, which can range from a short description to the full transcription, word-by-word. We have included textual data in the DataContent element by direct inclusion without namespace declaration. This decision leads to a simpler representation; however, it has the drawback that it cannot be validated by an XML schema. We plan to refine this issue during the continuation of the project. We always include a media type in the ContentItem that can be a text body, a news story, an article content, a video segment, or an audio segment. The characteristics of video (video coder, width, height, total duration, frame rate, etc.) and audio (audio coder, total duration, average bit rate, etc.) are automatically added by the digital video management server that controls the recording and digitization procedure.

The NewsComponent element groups together news objects that share metadata or that are conceptually associated into collections—for instance, a photograph and its caption. In our model, the constituents of the NewsComponent are complementary to each other. NewsML assigns four classes of metadata to the NewsComponent element: administrative, rights, miscellaneous, and descriptive. Into the latter category falls information about the content: title, summary, topic, polarity. The HeadLine element provides a displayable title. The KeywordLine element provides the set of keywords relevant to the news item. The SummaryLine element provides the summary of the news item, the Topic element provides its topic, and the Polarity element provides the annotator's judgment on the sentiment of the story. Administrative metadata provide information about the creation of the content: FileName, Provider,

Creator, etc. Above the structural level, the NewsItem is the primary unit for news management. This is the news object that we create, store, manage, reuse, and link to from other NewsItems. The NewsItem has a unique identifier, the NewsIdentifier, since it needs to move through the news workflow. We ensure that no two NewsItems carry the same ProviderId, DateId, NewsItemId, or RevisionId. The NewsManagement element provides information about a NewsItem's type, history, and status as well as its relationship to other NewsItems. NewsItems are exchanged between editorial systems in NewsEnvelopes.

In the NewsML framework, values of two basic types can be assigned to XML elements and attributes: controlled and uncontrolled. We use controlled vocabularies in order to ensure that metadata take well defined values that are constant over time and can be used for automated processing. The majority of the elements in our scheme are governed by controlled languages; among them are Format, Language, Provider, MediaType, Status, Priority, Topic (selected among a subset of the IPTC Subject Codes[3]), etc. On the other hand, the HeadLine, Summary, and DataContent elements are uncontrolled. We could not restrict either the syntax or the vocabulary of these elements, especially of the DataContent, which may contain the full transcription of what actually has been said. Informal guidelines have attempted to standardize terminology and to reduce ambiguity in headlines and summaries, albeit with limited success (i.e., sentences with less than a fixed number of words, simple syntax, correct spelling, etc.). However we were able to enforce user agreement on a single form for named entities, especially politicians, journalists, locations, and organizations that appear frequently on the news.

The polarity element reflects a story classification by overall sentiment. Sentiment-based classification is manual and controlled, and it can take one value out of the following set: strongly positive, positive, neutral, negative, strongly negative. The users try to detect the expressive subjective elements in the story and to recognize beliefs, emotions, evaluations, and judgments, in the given context (Wilson, Wiebe, & Hwa, 2006). When the story expresses a mixture of feelings, some positive and some negative, we asked the users to annotate such stories as neutral. We did not compute at this phase an interannotator agreement in order to understand whether the meaning of the values is uncontroversial. We are currently monitoring the way they perceive the meaning of positive and negative polarity, and we plan to refine our annotation model in a future phase of the project, where we also envisage the implementation of automatic polarity classification that subsequently will be confirmed by the users.

SEARCHING MULTIMEDIA CONTENT

Users can locate the news of interest either by watching and listening or by searching. In the first case, channels are assigned to users, and as many as 16 channels can be displayed simultaneously on the view screen. For most of the day, a single channel corresponds to a single user, especially when news production is intensive.[4] Regarding radio, since the human brain cannot process multiple acoustic channels simultaneously, it is always the case that one user is assigned to one channel. In the second case, the system functions as a full content multimedia information retrieval of current and archived TV and radio news and broadcasts. A well known example of such a large-scale, online digital video library that is developing base technology for machine understanding of video and film media is the CMU Informedia project (Hauptmann, Jin, & Ng, 2003).

SGC-SGI analysts can search and retrieve multimedia news items concerning the following:

(a). Foreign or domestic affairs
(b). Interviews with politicians
(c). Answers to journalists' questions on specific topics
(d). Opinions and different views on a subject
(e). The "who spoke, when, and where" list
(f). Packaging of critical events in different media

(g). Topic ranking following various criteria
(h). Weekly/monthly news bulletins
(d). Consolidated reports tracking news in their life cycle

The user generates a query in order to obtain information. The system then presents to the user a ranked list of video and audio segments (stories), ideally with those most relevant to the query topic first. Typical queries reflecting the user interests are listed below:

List all the broadcast news in a defined time range from the *Greek National TV channel (NET)*.
List the videos where the name of the anchorwoman is *Maria Houkli*.
List audios on terms *bioinformatics* and *genetics*. Account for possible morphological variations.
List videos about *wildfires* where the name *Costas Karamanlis* [Greek Prime Minister] appears.
List news released between *Dec. 25, 2006* and *Dec. 31, 2006*.
List news from the audiovisual archive about *Olympic Games 2004 in Athens*.
List news on *Minister Theodoros Roussopoulos*, on Oct. 10, 2007 where the semantic orientation of the story was positive.

Vector Space Model for Transcribed Text

Information retrieval proceeds in two modes. Under the first mode, users detect new stories of interest by querying the automatically transcribed text of shows usually broadcasted within the last 24 hours. To make the retrieval and viewing of information faster, the digital video library supports partitioning video into small-sized clips. This partitioning process is accomplished by the audio segmentation module, turning audio into a series of consecutive speech segments that correspond to speaker turns. Benefits of segment retrieval include the retrieval of the most relevant portions of longer documents, the avoidance of document-length normalization problems, and the possibility of more user-friendly interfaces that return the most relevant portion of a multimedia document. They are the likely boundaries of speaker turns, music, or noise segments.

The basic retrieval unit is the audio segment that plays the role of a document of a standard retrieval system. Each segment corresponds to one video (on TV) or audio chunk (on radio) and is indexed on textual metadata produced by automatic transcription. Segments and queries are represented as vectors in a common vector space. Their similarity is quantified as the cosine similarity of their vectors weighted by a *tf*idf* weighting scheme.[5] Prior to similarity calculation, we perform tokenization and stop word filtering of both the query (online) and the transcribed text (off-line). In order to reduce inflected words to a single morphological form, and taking into account the fact that morphological variants of words have similar semantic interpretations, we have applied stemming algorithms (or stemmers) in English and in Greek. Both stemmers are variants of the classic Porter stemmer[6] that implements a set of suffix stripping rules in order to find the stem of a word (Porter, 1980). We use inverted files and inverted lists to index the XML metadata into relational tables of a commercial relational database, and the system transforms user queries into SQL queries over relational data, allowing full text search.

Weighted Boolean Model for Annotated Text

Under the second mode of retrieval, users query manually annotated stories on a set of metadata that correspond to QualiArc fields. Metadata are associated with (a) free text fields (sometimes called *zones*) such as title, summary, speakers, and full-text transcription, and (b) parametric fields that can take a finite set of values such as date, channel, name of show, polarity, and breaking news flag. The retrieval engine merges free text and parametric indexes into a single weighted Boolean model. Parametric indexes allow the retrieval of stories matching only the specified fields. For instance, the date of creation index allows us to select only the stories matching a date specified in the query. For weighted Boolean scoring, the process may be viewed as computing a linear function of the

Boolean match scores contributed by the various free text fields. Given the query q and an annotated story d, the weighted Boolean scoring assigns to the pair (q, d) a score in [0, 1] by a linear combination of field scores, where each field contributes a boolean value: $\sum_{i=1}^{l} w_i s_i$, where l is the number of fields, w_i the weight of field i, and s_i a Boolean score denoting a match between (or absence thereof) q and the *ith* field. The score of a field is 1 if any of the query terms occur in that field (Boolean function OR). The weights are specified by expert users after experimentation with training examples that have been manually evaluated. Summary, speakers, and full-text transcription fields have the same weight, while the title contributes a little more to the total match.

SPEECH RECOGNITION IN BROADCAST NEWS

We have implemented a digital video management system (DVMS) for maximum quality digital video and synchronized audio recording. The system supports multiple compression modes in real time for each channel. It offers broadcast quality video and allows for optimized management of video and audio streams to the client computers of SGC-SGI.

A simple and effective GUI enables video and audio live view, playback, cut, and export of video and audio segments via LAN or the Internet. Audio streams that conform to speech recognition requirements for broadcast are extracted from video, that is 16 KHz, 16-bit signed linear PCM, and are presented to the input of speech recognition engines in Greek and in English.

Broadcast news exhibit a wide variety of audio characteristics, including clean speech, telephone speech, conference speech, music, and speech corrupted by music or noise (Ajmera, McCowan, & Bourlard, 2002). Transcribing the audio—that is, 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) is a challenging problem. Adverse background conditions can lead to significant degradation in performance. Consequently, adaptation to the varied acoustic properties of the signal or to a particular speaker and enhancements to the segmentation process are generally acknowledged to be key areas for research and improvement in order to render indexing systems usable. We apply a highly accurate, speaker-independent speech recognizer in English and in Greek in order to automatically transcribe audio recorded from broadcast news, which is then stored in a full-text information retrieval system.

A large vocabulary, speaker-independent, automatic speech recognizer is generally based on the hidden Markov model (HMM) technology. It comprises three basic components: 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 two other components. The audio segmentation module automatically identifies speech regions, rejects nonspeech ones, and clusters homogeneous regions of speech, that is, same speaker and 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 basic building blocks of a state-of-the-art speech recognition engine are the acoustic model (AM) and the language model (LM). The AM represents the acoustic parameters of the problem (acoustics, phonetics, environment, audio source, speaker and channel characteristics, recording equipment) while the LM contains a representation of the neighborhood of the words contained in the vocabulary and their statistical properties.

A large vocabulary, speaker-independent, gender-independent, continuous automatic speech recognizer transcribes Greek (Papageorgiou, Antonopoulos, Demiros, and Gkiokas, 2006).

In order to train our acoustic models, we have used 90 hours of carefully transcribed audio wave files of recent news and current affairs shows that were recorded from the Greek TV. We gathered a large corpus of 210 million for constructing the language model in order to predict the likelihood of sequences of words.[7]

We use BBN's very large vocabulary speech recognition system for transcribing English. The BBN Byblos Engine forms the core of the application (Nguyen et al., 2004). The system uses HMMs and Gaussian mixture models for the acoustic model (AM) and n-gram models for the language model (LM) of the recognizer (Colthurst et al., 2000). The engine was trained on 137 hours of acoustic data and 140 million words of CNN text.[8] Both recognizers generate time stamped transcriptions in XML format (see example in Appendix A).

Speech Recognition and Retrieval Evaluation

We evaluate the speech recognition engines per se for the purpose of multimedia retrieval. We did not evaluate the retrieval of manually annotated stories, only of machine annotated stories. Users detect new stories by searching the transcript and promote the ones that are subsequently annotated. This way, retrieval based on automatic transcription is crucial for the detection of stories of interest.

Beginning in the early 1980s, evaluation of automatic speech recognition (ASR) stabilized on the current performance measure of word error rate (WER). This measure scores ASR performance using a caseless, lexicalized from of ASR output known as the standard normalized orthographic representation (SNOR) format. The WER is defined as the sum of all ASR output token errors divided by the number of scoreable tokens in a reference transcription of the test data. There are three types of errors; namely, tokens that are missed (*deletion errors*), inserted (*insertion errors*), and incorrectly recognized (*substitution errors*). WER is considered to be a better measure of recognizer accuracy than the number of words correct alone. In our system, the overall speech recognition WERs are 25% for decoding speed of 1.2xReal-Time in Greek and 14.9% for a decoding speed of 1xReal-Time in U.S. English. In English, the reference points are the standard speech evaluation data that are used to benchmark broadcast news speech recognition systems. The evaluation deals with anchored news shows and news magazines (examples for English ASR: *ABC Prime Time, ABC World Nightly News, CNN Headline News, CNN World View,* etc.). This program material includes a combination of read speech and spontaneous speech, as well as a combination of recording environments in broadcast studios and in the field. The BBN English ASR was evaluated on the EARS RT04f evaluation set.[9]

No benchmark exists for the Greek language. We have constructed an accurate verbatim transcript of 10 hours of news and current affairs shows, recorded from Greek television. We have followed the EARS transcription guidelines, and the Greek speech recognition evaluation has been performed on this particular dataset. Although the above figures are the highest ever reported for Greek, spontaneous speech-related issues and background noise compensation techniques are currently investigated in order to further improve the robustness and effectiveness of the final system and to narrow the gap between the success of the Greek and the English ASR. We are enriching our database of recordings and transcriptions (dated 2007), and we are also building a new domain-specific language model from available textual data. It is also noteworthy that Greek TV news and current affairs are characterized by the presence of multiple speakers and analysts in constant interaction under significant variations of the speech environment, thus making the task of speech recognition more difficult than in English news broadcasts. Moreover, multilinguality (Gauvain, Lamel, & Adda 2000), another issue for future work, is of particular interest in our application, since major events are covered in different languages abroad and subtitled for the Greek audience.

In order to evaluate the effectiveness of speech recognition for multimedia retrieval, we have manually gathered 42 stories of interest that appeared in the news, current affairs, and talk shows during a period of five consecutive

days. Among them, 22 were in English and 20 were in Greek. A story is a segment of a news broadcast with a coherent news focus (Amaral & Trancoso, 2003). Since the system evaluation period coincided with the pre-election period for the parliamentary elections that were held on September 16, 2007, Greek stories fell into this major topic. The users extracted the set of keywords that best described the stories. The users have also determined the story boundaries within the videos. The system finally processed each story by automatic speech recognition. At this point we had a collection of story transcripts and a set of queries that reflected the users' information needs; in other words, we had created a small test document collection.

Next, the users were asked to search for the stories by any combination of topic keywords they judged to be suitable. They did not have to resort to specialized syntax or canonical forms of words in their query. For each query, the system returned the matching stories ranked by relevance. A document was judged by the users to be relevant if it addressed their information needs. Relevance is a standard assessment for each document–query pair in information retrieval systems. It is noteworthy that in our project we did not possess a standard test collection that we could run the testing against, such as the standard TREC, GOV2, and REUTERS data collections that are extensively used for ad hoc information retrieval evaluation.

We have used precision and recall, which are the basic measures of retrieval effectiveness. Precision is the fraction of retrieved stories that are relevant, while recall is the fraction of relevant documents that are retrieved. It depends highly on the application which of the two measures is more important than the other. Media analysts are concerned with setting as high recall as possible and will tolerate lower precision figures in order to achieve it. The system achieved very high recall, namely 41 out of 42 stories (98%), satisfying in this way one of the main requirements of the project. Precision, which constitutes the other major retrieval measure, was 41 out of 150 stories (27%), indicating that the majority of stories were judged as false positives—retrieved but not relevant. Analysts had to inspect and discard them as noneligible stories of interest for further annotation. It is important to note that although the retrieval system presents ranked results, we evaluated unordered sets of stories. It is also clear that our dataset is small and that our main purpose during evaluation has been to show that retrieval is successful despite imperfect speech recognition technology.

SUPPORTING MEDIA ANALYSIS

By means of the system that we have described above, the Section of Audiovisual Media at SGC-SGI monitors the broadcastings in order to compile information and news items pertaining to issues that lie within the competence of SGC-SGI. It also prepares bulletins to inform other divisions of the SGC-SGI, as well as state agencies and public services. The Section keeps the political leadership posted by means of clippings and stories from Greek television and Greek radio. It also maintains archives in a manner that secures the timely retrieving and correlating of the required information. The Section informs the hierarchy and delivers news, official statements, and information on state activity issues to the mass media, and it releases announcements to refute or reject inaccurate news items. In addition, it contacts the media representatives on a regular basis and supports the Daily Briefing by the Government Spokesman.

The QualiArc media monitoring system offers the possibility of analyzing the collected data. Users conduct thematic content analysis to news stories and other media products. They explore the main topics to which the news stories are referring. Thematic analysis reveals which aspects of a news story are highlighted by the media and which are less mentioned or omitted (e.g., which audiovisual programs covered a certain topic in a given period of time, or which news stories—and to what extent—mentioned government activity or the activity of the opposition parties). The emphasis given to particular aspects depends on multiple factors, such as the political and ideological orientation of the media, the existing information on the issue, the influence of the sources, the anticipated interest of the

public, etc. Therefore, political and social actors, wishing to impose their own agenda, need to take into consideration which issues get prominent attention and which aspects of these issues are made salient by the media in order to adapt their strategy or try to change the terms of the debate. This is of great importance to the SGC-SGI, which serves the communication needs of the state administration.

With the advent and growth of multimedia computing technologies, users are able to store and retrieve great amounts of information. Networks, technology, and an efficient internal structure enable SGC-SGI to fully exploit the inherent value of audiovisual data recordings from television and radio sources. The system provides a ten-day archive of the full broadcast of twelve TV and four radio channels. All video and audio is instantly available. A graphical browsing function allows easy navigation through time. It also enables clipping of video/audio segments that are subsequently imported into QualiArc for annotation and characterization. SGC-SGI users produce around 100 new annotations per day, ranging from short video segments of a few minutes' duration to full transcriptions of one-hour interviews accompanied by sentiment analysis, summary, and topics. The internal media department now has the necessary technology and networks capable of creating, storing, and distributing annotated media segments corresponding to news, events, and stories. Unstructured and unconnected content is transformed into structured information that is stored in a large relational database and can be retrieved by means of sophisticated search algorithms. We expect these digital media assets to accumulate and to be leveraged for multiple users and that they will substantially change the internal and external communications capabilities of SGC-SGI.

CONCLUSION AND FUTURE WORK

In this article, we have described a media monitoring system that we have developed and implemented for the Secretariat General of Communication and Secretariat General of Information in Greece (SGC-SGI). The system applies emerging technologies for audiovisual recording, speech recognition, news annotation, language processing, multimedia indexing, and retrieval, all integrated into a large video and audio library that covers broadcast news and current affairs in Greek and in English. 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 and language technology. We have also delved into the problems that large organizations face when they adopt new systems and technologies. Most important, we have learned that the use of such technologies can support government operations and services, journalists, and political analysts with their work and can engage the government in delivering applications and services of the highest quality.

The success of the project led SGC-SGI to make the decision to further develop it by applying a holistic concept to modeling media monitoring public services. At the conceptual level, we constantly aim at merging the public service model with state-of-the-art technologies of indexing and searching in multimedia information. We plan to enrich the number of information sources by including local TV and radio channels in Greece and by extending the number of English-speaking broadcasts. Speech recognition in new languages is also foreseen. Through the integration of intelligent engines from the fields of language processing and understanding, speech recognition, and image processing, the video and audio library system will assist the user in exploring multimedia data in depth. Speaker recognition, key-frame extraction, and retrieval techniques that go beyond the Boolean model (i.e., learning weights by machine learning techniques) are among the technologies that will be integrated in the future system extension. Moreover, a Web information extraction system will visit the sites of broadcasts and news agencies in order to collect information that will be analyzed and mined centrally, in order to help analysts to find online information that matches their needs. SGC-SGI will continue to implement information technologies in order to improve its operations and services and to enhance the role that stems from its mandate.

NOTES

1. Former Greek Ministry of Press and Mass Media; http://www.minpress.gr.

2. NewsML, now at version 2.0: http://www.iptc.org/G2-Standards/newsml-g2.php

3. International Press Telecommunications Council at: http://www.iptc.org

4. We have identified three zones of great interest: the morning zone 6–10 a.m., the afternoon zone 2–4 p.m., and the evening zone peaking at 8–9 p.m. Naturally, there are variations according to the medium and the channel.

5. tf*idf (term frequency—inverse document frequency) is a measure used in information retrieval to evaluate the importance of a term to a document in a collection.

6. Martin Porter extended his original work by building Snowball, a framework for writing stemming algorithms, available at http://snowball.tartarus.org.

7. For the computation of acoustic probabilities, we have used three-state triphone cross word HMMs with 20 Gaussians per state. We trained a trigram-based LM. The vocabulary consisted in 65,000 words.

8. We use gender-dependent quinphone and triphone models in the AM. We used four different language models and a 35,000 words dictionary from Switchboard and CallHome data for LM training.

9. DARPA EARS project at: http://projects.ldc.upenn.edu/EARS/

REFERENCES

Ajmera, J., McCowan, I., & Bourlard, H. (2002). Robust HMM-based speech/music segmentation. *Proceedings of ICASSP 2002* (pp. 297–300). Orlando, FL: Springer Verlag.

Amaral, R., & Trancoso, I. (2003). Topic indexing of TV broadcast news programs. In N. Mamede, J. Baptista, I. Trancos., & M. Nunes (Eds.), *Computational Processing of the Portugese Language: 6th International Workshop*, PROPOR 2003 (pp. 219–226). Berlin: Springer.

Colthurst, T., Kimball, O., Richardson, F., Shu, H., Wooters, C., Iyer, R., & Gish, H. (2000). The 2000 BBN Byblos LVCSR system. *ICSLP-2000, Vol. 2.* (pp. 1011–1014). Beijing: China Military Friendship Publishing.

Gauvain, J. L., Lamel, L., & Adda, G. (2000, February). Transcribing broadcast news for audio and video indexing. *Communications of the ACM, 43*(2), 64–70.

Hauptmann, A. G., Jin, R., & Ng, T. D. (2003, January). Video retrieval using speech and image information. *Storage and Retrieval for Multimedia Databases 2003, 5021,* 148–159.

Nguyen, L., Abdou, S., Afify, M., Makhoul, J., Matsoukas, S., Schwartz, R., Xiang, B., Lamel, L., Gauvain, J. L., Adda, G., Schwenk, H., & Lefevre, F. (2004, November). The 2004 BBN/LIMSI 10xRT English broadcast news transcription system. *Proceedings of DARPA RT04.* Palisades, NY: Springer Verlag.

Papageorgiou, H., Antonopoulos, V., Demiros, I., & Gkiokas, A. (2006, May). Thematic classification and intelligent indexing of broadcast news using speech recognition and image analysis. Paper presented at the *EuroITV 2006*, Athens, Greece.

Porter, M. (1980). An algorithm for suffix stripping. *Program, 14*(3), 130–137.

Wilson T., Wiebe, J. & Hwa, R. (2006). Recognizing strong and weak opinion clauses. *Computational Intelligence, 22*(2), 73–99.

APPENDIX A: XML SAMPLE OF THE ASR MODULE

```
<?xml version="1.0" encoding="ISO-8859-7" ?>
<SpeechAnnotation project="MinPress">
<Header type="SpeechRecognition">
    <CreationTime>Mon Oct 1 09:36:48 2007</CreationTime>
    <LastUpdate>Mon Oct 1 09:36:48 2007</LastUpdate>
    <Comment>Speech Processing word_transcription_xml_handle</Comment>
    <Creator>QUALIA</Creator>
</Header>
<ASR>
<Source>
<Name>ET3</Name>
    <Speech Language="EL" StartingDT="2007-11-05 07:00:00">
<segment id="s_0" name="110507_ET3_NEWS_19-00.1" start="13.18" end="18.11"
    type="SPEECH" length="491" duration="4.91">
    <transcript wseq="<S> χαίρεται κυρίες και κύριοι λύση διαφαίνεται στο ζήτημα
    της καταβολής των αναδρομικών στους δικαστικούς </S>" />
    <word WordId="1" WordText="χαίρεται" start="13.54" end="13.95" />
    <word WordId="2" WordText="κυρίες" start="13.96" end="14.22" />
    <word WordId="3" WordText="και" start="14.23" end="14.35" />
    <word WordId="4" WordText="κύριοι" start="14.36" end="14.67" />
    <word WordId="5" WordText="λύση" start="14.68" end="14.95" />
    <word WordId="6" WordText="διαφαίνεται" start="14.96" end="15.44" />
    <word WordId="7" WordText="στο" start="15.45" end="15.58" />
    <word WordId="8" WordText="ζήτημα" start="15.59" end="15.88" />
    <word WordId="9" WordText="της" start="15.89" end="16" />
    <word WordId="10" WordText="καταβολής" start="16.01" end="16.47" />
    <word WordId="11" WordText="των" start="16.48" end="16.61" />
    <word WordId="12" WordText="αναδρομικών" start="16.62" end="17.21" />
    <word WordId="13" WordText="στους" start="17.22" end="17.43" />
    <word WordId="14" WordText="δικαστικούς" start="17.44" end="18.03" />
</segment>
 <segment id="s_1" name="" start="18.11" end="18.26" type="MUSIC" length="15"
duration="0.15" />
 </segment>
```

APPENDIX B:
ANNOTATED STORY SAMPLE IN THE NEWSML METADATA SCHEME

```
<?xml version="1.0" encoding="UTF-8"?>
<NewsML>
    <NewsEnvelope type="NewsMetadata">
            <DateAndTime>20070123T161535</DateAndTime>
            <NewsService FormalName="SGC-SGI" />
            <NewsProduct version="4.0">QualiArc</NewsProduct>
            <Priority FormalName="3" />
    </NewsEnvelope>
 <NewsItem>
    <NewsIdentifier>
            <ProviderId>ERT</ProviderId>
            <NewsItemId>{BF002C77-A0BF-4C00-8A74-47D6A931DF30}</NewsItemId>
    </NewsIdentifier>
    <NewsManagement>
            <NewsItemType FormalName="Radio News" />
            <FirstCreated>20070123T150000</FirstCreated>
            <Status FormalName="Usable" />
            <Urgency FormalName="3" />
    </NewsManagement>
  <NewsComponent>
    <HeadLine>Νέος αρχηγός ΓΕΑ</HeadLine>
    <SummaryLine></SummaryLine>
    <Topic Id="EPT_1">
            <ClassificationScheme>IPTCSubjectCodes</ClassificationScheme>
            <ClassificationCode>11000000</ClassificationCode>
            <ClassificationLabel>Politics</ClassificationLabel>
    </Topic>
    <Polarity>Neutral</Polarity>
    <AdministrativeMetadata>
            <Provider>
             <Party FormalName="EPT-NET" />
            </Provider>
            <Source>
             <Party FormalName="NET 105,8" />
            </Source>
    </AdministrativeMetadata>
    <DescriptiveMetadata>
            <Language FormalName="en" />
            <OfInterestTo FormalName="Analyst" />
    </DescriptiveMetadata>
    <ContentItem Href="./NETRadio_News_20070123T140000.wav">
     <MediaType FormalName="Audio" />
     <Format FormalName="WAV" />
     <Characteristics>
            <SizeInBytes>104728</SizeInBytes>
            <Property FormalName="FileExtension" Value=".wav"/>
            <Property FormalName="AudioCoder" Value="PCM" />
            <Property FormalName="TotalDuration" Value="60"/>
            <Property FormalName="SampleRate" Value="16.000"/>
            <Property FormalName="AudioChannels" Value="2"/>
     </Characteristics>
     <DataContent>Αλλαγές στην ηγεσία της Πολεμικής Αεροπορίας αποφάσισε το ΚΥΣΕΑ που συνεδρίασε την
Τρίτη υπό την προεδρία του πρωθυπουργού, Κώστα Καραμανλή.</DataContent>
    </ContentItem>
  </NewsComponent>
 </NewsItem>
</NewsML>
```