MAESTRO: CONDUCTOR OF MULTIMEDIA ANALYSIS TECHNOLOGIES

Ze'ev Rivlin, Douglas Appelt, Robert Bolles, Adam Cheyer, Dilek Hakkani-Tür, David Israel, Luc Julia, David Martin, Greg Myers, Ken Nitz, Bikash Sabata, Ananth Sankar, Elizabeth Shriberg, Kemal Sonmez, Andreas Stolcke, Gökhan Tür

SRI International
Menlo Park, California 94025
http://www.chic.sri.com/projects/Maestro.html
Contact: julia@ai.sri.com

Although keyword-based queries are now a familiar part of any user's experience with the World Wide Web, they are of limited direct applicability to the vast and growing quantity of multimedia information becoming available in materials such as broadcast news, video teleconferences, reconnaissance data, and audio-visual recordings of corporate meetings and classroom lectures. Content-based indexing, archiving and retrieval would facilitate access to large databases of such materials.

For example, in the broadcast news domain, content-based archiving is particularly useful. Archiving can be done by exploiting the speech contained in the audio track, the images contained in the video track, and the text in video overlays. One application is to filter down huge volumes of raw news footage to create the nicely packaged news broadcasts that we watch on television. Another use is to create *a news-on-demand* system for viewing news more efficiently. We can create a database of news broadcasts annotated for later retrieval of news clips of interest. The query *"Tell me about the recent elections in Bosnia"* would bring up news clips related to the elections.

MAESTRO (Multimedia Annotation and Enhancement via a Synergy of Technologies and Reviewing Operators) is a research and demonstration system developed at SRI International for exploring the contribution of a variety of analysis technologies — for example, speech recognition, image understanding, and optical character recognition — to the indexing and retrieval of multimedia. Informedia [1] and Broadcast News Navigator [2] are similar projects that use these technologies for archiving and retrieval. The main goal of the MAESTRO project is to discover, implement, and evaluate various combinations of these technologies to achieve analysis performance that surpasses the sum of the parts. For example, British Prime Minister Tony Blair can be identified in the news by his voice, his appearance, captions, and other cues. A combination of these cues should provide more reliable identification of the Prime Minister than using any of the cues on their own.

MAESTRO is a highly multidisciplinary effort, involving contributions from three laboratories across two divisions at SRI. Each of these SRI technologies is described in more detail below. The integrating architecture makes it easy to combine these in different ways, and to incorporate new analysis technologies developed by our team or by others.

MULTIMEDIA ANALYSIS TECHNOLOGIES ON THE MAESTRO SCORE

On the MAESTRO Score (see Figure 1), on each line, similar to musical notation for a particular instrument, as seen on a musical conductor's score, we see the output from a particular multimedia analysis technology. All of these outputs share a common timeline. For example, the speaker tracking line (second from top) shows the audio segmented into speakers, numbered 1, 2, 3, and so on. MAESTRO accepts an input format that consists of pairs of information, each describing an event and the time interval in which it occurred. For example, for speaker tracking, an information pair could be speaker 4 speaking from time 1:12 to time 1:35 in the given news story. This appears on the MAESTRO Score as the marked segment on the speaker tracking line with the label "4" below it. MAESTRO provides the visual interface (the MAESTRO Score) for discovering synergistic strategies of combining analysis technologies for the best possible archiving and retrieving given all available data and relevant technologies.

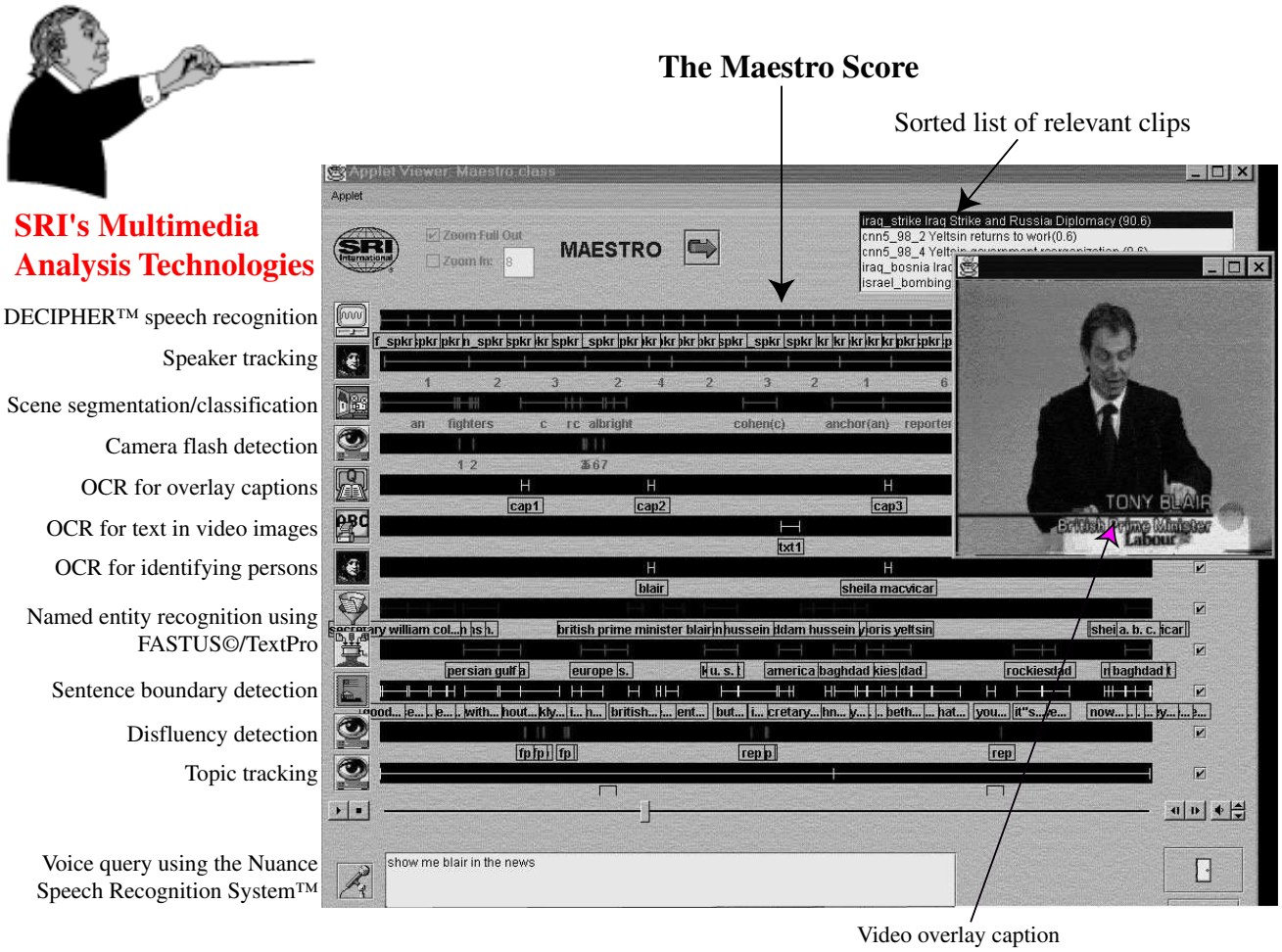

Figure 1: The MAESTRO Score

Up to now, we have incorporated into MAESTRO the SRI multimedia analysis technologies denoted to the left of the MAESTRO Score in Figure 1. In the MAESTRO video display (the inset frame in Figure 1), we see British Prime Minister Tony Blair speaking with his name in a video overlay caption. We will refer to the example of identifying the Prime Minister in the brief descriptions of each of the SRI analysis technologies below. For more detailed descriptions, we include references to related technical publications and relevant Web pages.

Speech Recognition

Speech recognition technology (for more information, see http://www.speech.sri.com) is used to automatically recognize the words that people speak in the news, including speech from news anchors, field reporters, world leaders, and passersby interviewed on the street. SRI's large-vocabulary speech recognition system, known as DECIPHER™, automatically transcribes the speech, creating corresponding text. DECIPHER™ can recognize natural, continuous speech without requiring the user to train the system in advance (*a speaker-independent* system). DECIPHER™ is distinguished by its efficient modeling approach (which results in very small models), its robustness to noise and channel distortion, and its multilingual capabilities. These features make the system fast and accurate in recognizing spontaneous speech of many styles, dialects, languages, and noise conditions. The DECIPHER™ recognizer was trained in the broadcast news domain [3], and its ability to deal with a wide variety of noise conditions (e.g., clean broadcast studio speech, noisy crowd background) and spontaneous speech (e.g., passersby interviewed on the street) was essential in this domain.

Speech recognition contributes to our example task of identifying Tony Blair by recognizing the

news anchor announcing, “British Prime Minister Blair said...”.

Named Entity Detection

The named-entity detection system identifies names of people, companies, organizations, and locations in text. In broadcast news, this text can come from the output of the speech recognizer, or from video overlay captions. SRI’s FASTUS© system [4] uses an English lexicon, extensive lists of persons, companies and locations, as well as contextual information to identify and classify proper names. MAESTRO uses TextPro, (a version of which is publicly available over the Web at the URL http://www.ai.sri.com/~appelt/TextPro) which is derived from FASTUS©, but implemented in C++, and compliant with the architecture developed under the government sponsored TIPSTER program. The performance of TextPro has been evaluated on several hours of broadcast news transcripts, and it identifies approximately 91% of names in an uppercase-only input stream with no word recognition errors [5].

TextPro identifies “British Prime Minister Blair” (see Figure 1, center of the fourth line from bottom on the MAESTRO score) as the name of a person in the text output from the speech recognizer.

Disfluency Detection

Spontaneous speech frequently contains disfluencies (hesitations, self-correction, restarts) that we want to detect and eliminate before further processing. This is accomplished by SRI's hidden event modeling techniques that rely on a combination of lexical and prosodic cues to find the events of interest (for more information on this topic, see http://www.speech.sri.com/projects/hidden-events.html). These events are hidden in the sense that the speech recognizer typically just outputs text hypothesizing spoken words, but does not mark events such as the disfluencies mentioned above.

In certain cases, named entities have embedded disfluencies. For example, if the news anchor announced, “British Prime Minister uh Blair said...” then the interruption by a disfluency makes it difficult to recognize the named entity. The disfluency detector cleans up the text, resulting in “British Prime Minister Blair said...”.

Optical Character Recognition (OCR)

Text appearing in video imagery includes computer-generated captions overlaid on the imagery, and text that is part of the scene itself, such as signs and placards. Location and recognition of text in video imagery is more difficult than in many other OCR applications (e.g., reading printed matter) because of small character sizes and nonuniform backgrounds. SRI has developed an approach involving substantial preprocessing (binarizing individual color video frames, locating lines of text in each binarized frame, and binarizing the detected text regions again from the original color image data) before applying the OCR engine. The accuracy of the recognition result can be improved substantially by additional postprocessing using a lexicon of named entities automatically found by the named entity detector in text output from the speech recognizer. For example, Figure 2 shows part of the OCR recognition results of a video overlay caption. Without the help of the named entity lexicon, due to the background of the Prime Minister's hand behind the overlay caption, OCR recognizes ‘BLAKJB'. However, with the help of the lexicon from the named entity detector, OCR is able to correctly recognize ‘BLAIR'. The detection of instances of named entities in the video imagery, which often correspond to a person, place, or organization depicted in the video scene, is noted as a semantic event on the MAESTRO Score.

Without Lexicon: TONY BLAKJB
With Lexicon: TONY BLAIR

Figure 2: Name Recognition Helps OCR

Speaker Identification and Tracking

Speaker identification and tracking technology was developed at SRI for various domains including broadcast news. Identifying and tracking speakers is

of fundamental importance in archiving broadcast news since it allows the user to search the data stream by speaker identity. In SRI's approach to speaker tracking [6], portions of the audio stream marked as speech are processed to fit a speaker change model generated iteratively by grouping speech into unique speaker bins and training models for each hypothesized speaker. Fitting of the resulting speaker change model to the data determines the speaker turns as well as the acoustic models of speakers. Specifically, an existing set of models for speakers appearing frequently, such as news anchors and world leaders, can be trained with available data and subsequently used to verify these speakers' presence in the audio stream and assign labels accordingly.

For our identification example, we can build a speech model for Tony Blair using various news clips in which he speaks, and subsequently use this model to find new clips of the Prime Minister speaking.

Scene Segmentation and Classification

The first step in understanding the imagery content in broadcast news video is the segmentation of the video into contiguous camera shots. This is achieved by detecting scene changes in the video. We use the basic properties of the color distributions of individual video frames to detect the changes in the scene. Instead of past approaches where two adjacent frames' color distributions are compared, we use a novel robust multiscale approach to compute color discontinuity across scene change boundaries [7]. The procedure detects scene changes that occur because of cuts, fades or dissolves. Our method is very robust to the different noise phenomena in video and also not sensitive to the thresholds that are required by other methods.

The classification of the scenes can potentially provide important inputs to higher-level interpretations, such as distinguishing the anchor person camera shots from the other news item shots, or shots with text and graphics from other general scenes. Another application is that similar shots are used in similar or related stories over different news broadcasts. For example, we encountered cases in which Tony Blair appears in a few places in a particular news story, such as when the news anchor refers several times to comments made by the Prime Minister. Each time the Prime Minister is shown, the scene is similar — he is speaking from a white podium wearing a black suit, the background is gray. The scene classifier applies a hierarchical agglomerative clustering technique to the features of the color distribution in each frame to classify these segments of Tony Blair as being similar and assigns them to the same class. Once we have located Tony Blair in one segment of the news story, the scene classifier provides a cue that will be useful for identifying the Prime Minister in other video segments that were assigned to the same class.

Camera Flash Detection

Camera flashes are an important cue for understanding the content of scenes in broadcast news video, as they typically indicate the presence of important persons in news video as the press attempts to take good pictures. Our algorithm to detect camera flashes is a direct by-product of the above procedure to detect scene changes. Camera flashes are a cue that a person worth identifying, such as the Prime Minister, may be in the scene.

Topic Segmentation and Tracking

SRI has developed technology that enables MAESTRO to automatically divide speech into topic units and display automatically determined keywords characterizing each segment. To find the topic boundaries, the system combines prosodic information extracted from the speech waveforms (such as pause and pitch patterns) with word usage statistics. Both sources of information are integrated in a hidden Markov model to find the best segmentation [8].

For example, if we wanted to find the Prime Minister specifically in a story related to British economics, we could search for occurrences of Tony Blair in stories tagged with keywords *British* and *economics*.

SYNERGISTIC COMBINATIONS FOR IMPROVED ANALYSIS

The main objective of MAESTRO is to combine analysis technologies for the best possible performance in analysis, archiving, and retrieving. MAESTRO's multimedia interface allows the user to jump from one analysis technology line to another (Figure 1), to play video sequences and display output corresponding to the events on that line, and to evaluate the performance of each of these technologies. The display of analysis lines on the MAESTRO Score guides the viewer in the mental combination process between the lines --- that is, between the technologies. It even inspires some synergistic combination strategies and rules that would not have been obvious without this representation. This particular interface is effectively a “look under the hood” and has research and pedagogical use. A commercial

news-on-demand system may only present to the user the list of videos that match a query, without showing the underlying process that led to the retrieval result. For more information about intelligent human-computer interfaces developed at SRI, see http://www.chic.sri.com.

Through MAESTRO, we have learned how analysis technologies can be combined for extracting information of interest. For example, we can identify British Prime Minister Tony Blair by using the following cues:

- DECIPHER™ speech recognition recognizes the news anchor saying, "British Prime Minister Blair said...".
- Named entity detection using TextPro identifies "British Prime Minister Blair" as the name of a person.
- Disfluency detection can clean up a disfluent named entity "British Prime Minister uh Blair said..." yielding the recognizable name "British Prime Minister Blair."
- Optical character recognition, driven by a lexicon of names (including "BLAIR") from named entity detection, reads the video overlay caption recognizing "TONY BLAIR."
- Speaker identification and tracking identifies the Prime Minister's speech as 'speaker 4' (see Speaker Tracking line in Figure 1).

Conclusion: The person in the video and speaker 4 are identified as "TONY BLAIR."

- Scene segmentation and classification classifies the video segment of the Prime Minister speaking with a later segment of the same news story based on scene similarity. We may hypothesize that this is again Tony Blair.
- Camera flash detection identifies camera flashes later in the news story, which is a cue of the presence of a famous person, possibly the Prime Minister again.
- Topic segmentation/tracking can be used to find the Prime Minister in news stories about a particular subject — for example, stories related to British economics.

We are currently exploring various ways of automatically combining the above cues to locate the Prime Minister in the news.

- Heuristic (rule-based): If certain cues are evident at certain times, we simply conclude presence of the Prime Minister.
- Statistical combinations: Weighting the cues based on confidence measures — e.g., we may be 70% sure it is Tony Blair's voice and 45% sure this scene matches another scene where he appeared. We can develop an optimal weighting scheme to make the final decision.
- Combinations of heuristic and statistical approaches.

ARCHIVING AND RETRIEVING IN MAESTRO

Each news story is archived via a statistical language model applied to text corresponding to the speech recognized from its audio track. Archiving and retrieval is based on content words, or words that have high salience for the task. The content words are *automatically* determined by computing a salience measure for each word using an algorithm that implements word occurrence statistics [9]. Salient keywords are those that are useful for discriminating between news stories. For example, the word 'the' may appear in all stories and is not useful for retrieval. However, the word 'Blair' is useful for discriminating between stories about the Prime Minister, and those not about him. Inspired by magnetic poetry on a refrigerator door, MAESTRO's "Fridge" (Figure 3) shows the automatically determined keywords for use in archive voice querying, with larger words being more salient.

Stories can be retrieved by voice queries using the Nuance Speech Recognition System™, developed by Nuance Communications based on SRI's DECIPHER™ technology. The query, *"Tell me about Blair and economics,"* brings up a sorted list of news clips ranked by numerical score in order of relevance to the query. The highest scoring news clips will be news stories most relevant to economics and the British Prime Minister.

MAESTRO is implemented using the Open Agent Architecture™ (OAA), a general-purpose framework for constructing systems composed of multiple software components written in different programming languages and distributed across multiple platforms. Similar in spirit to distributed object infrastructures such as OMG's CORBA or Microsoft's DCOM, OAA's delegation-based approach provides support for describing more flexible and adaptable interactions than the tightly bound method calls used by these architectures [10].

In the current MAESTRO implementation, OAA's role is to integrate all on-line components, which include the MAESTRO User Interface (Java), speech recognition and telephony agents (C), and retrieval scoring (Awk). For example, the user can

query MAESTRO from Washington D.C. over the telephone via the speech recognition agent that may reside remotely on a computer at SRI in Menlo Park, California. At the same time, the MAESTRO User Interface (Figure 1) can reside in Washington D.C. on a laptop comjuter for local display. In the future, it is expected that OAA's capabilities will help provide automated synergy of multimedia analysis technologies, inferencing across multiple on-line processing streams.

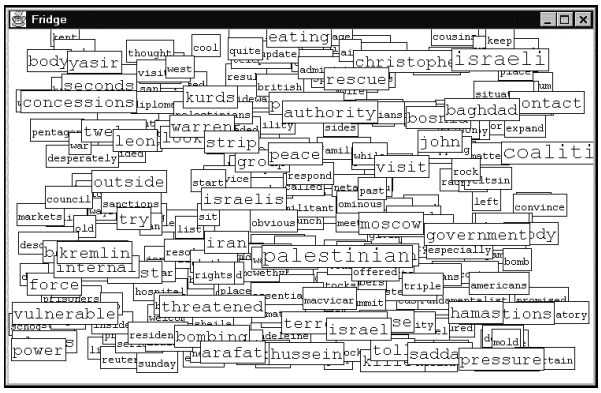

Figure 3: The MAESTRO Fridge

CONCLUSIONS AND FUTURE DIRECTIONS

MAESTRO is a research and demonstration testbed providing an integrating architecture and a visual interface (the MAESTRO Score). Through MAESTRO, the researcher or technical user can discover, develop, and test synergistic strategies of combining multimedia analysis technologies to achieve archiving and retrieving capabilities far exceeding that possible with any of the technologies individually. We have implemented a number of these strategies --- for example, using a lexicon of names from the named entity detector to enable correct recognition of video overlay captions by OCR. We have learned a great deal about information cues from studying the MAESTRO Score --- for example, cues for identifying newsmakers. And, most important, MAESTRO has motivated new ideas for combination strategies.

We presently are focusing our attention on implementing more synergistic multimedia-analysis-technology combination strategies. We are exploring various methods of combining cues that help answer questions like *"Is Tony Blair in the news?"* We are also interested in porting to new domains, such as meeting archival. Finally, we continue to further improve the individual core multimedia analysis technologies.

ACKNOWLEDGEMENTS

We gratefully acknowledge the efforts of the following people in developing MAESTRO: John Bear, Marty Fischler, Marsha Jo Hannah, Jerry Hobbs, Ray Perrault, Patti Price, Eric Rickard, and David Scott.

Support for this work from DARPA contract number N66001-97-C-8544, from DARPA through the Naval Command and Control Ocean Surveillance Center under contract number N66001-94-C-6048, through NSF grant IRI-9619921, and other Government agencies is gratefully acknowledged. The Government has certain rights in this material. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the Government funding agencies.

REFERENCES

[1] Hauptmann, A., and M. Whitbrock, "Informeida: News-on-demand multimedia information acquisition and retrieval," *Intelligent Multimedia Information Retrieval,* MIT Press, 1997, pp. 215–240.

[2] Merlino, A., D. Morey, M. Maybury, "Broadcase News Navigation using Story Segmentation," in *Fifth ACM International Multimedia Conference,* 1997, pp. 381–392.

[3] Sankar, A., R.R. Gadde, and F. Weng, "SRI's Broadcast News System — Toward Faster, Smaller, and Better Speech Recognition," in *Proceedings of the DARPA Broadcast News Workshop,* 1999, pp. 281–286. `http://www.speech.sri.com/papers/darpa99-fbs.ps.gz`

[4] Hobbs, J., D. Appelt, J. Bear, D. Israel, M. Kameyama, M. Stickel, and M. Tyson, "FASTUS: A Cascaded Finite-State Transducer for Extracting Information from Natural-Language Text," in Roche and Schabes (eds.) Finite State Devices for Natural Language Processing, MIT Press, 1996, 383–406.

[5] Appelt, D., D. Martin, "Named Entity Extraction from Speech: Approach and Results using the TextPro System," in *Proceedings of the 1999 DARPA Broadcast NEWS Workshop,* 1999, pp. 51–54. Also available at URL `http://www.ai.sri.com/~appelt/SRIIEN`

E.HTM.

[6] Sonmez, K., L. Heck, M. Weintraub, "Speaker Tracking and Detection with Multiple Speakers," in Proceedings of *EUROSPEECH 99,* 1999, pp. 2219–2222. Also available at URL http://www.speech.sri.com/papers/eurospeech99-tracking.ps.gz

[7] Sabata, B. and M. Goldszmidt, "Fusion of Multiple Cues for Video Segmentation," in *Proceedings of the Second International Conference on Information Fusion,* 1999.

[8] Stolcke A., E. Shriberg, D. Hakkani-Tür, G. Tür, Z. Rivlin, and K. Sonmez, "Combining Words and Speech Prosody for Automatic Topic Segmentation," in *Proceedings of the 1999 DARPA Broadcast NEWS Workshop,* 1999, pp. 61–64. http://www.speech.sri.com/papers/darpa99-topicseg.ps.gz

[9] Gorin, A. S. Levinson, and A. Sankar, "An Experiment in Spoken Language Acquisition, *IEEE Transactions on Speech and Audio Processing,* pp. 224-240, Volume 2, Number 1, 1994.

[10] Martin, D., A. Cheyer, and D. Moran, "The Open Agent Architecture: A Framework for Building Distributed Software Systems*," Applied Artificial Intelligence: An International Journal,* Volume 13, Number 1-2. January-March 1999. Also see the World Wide Web site http://www.ai.sri.com/~oaa/

For more information, see the MAESTRO web page at
http://www.chic.sri.com/projects/Maestro.html