Multimodal Technology Integration for News-on-Demand

SRI International

News-on-Demand Compare & Contrast
DARPA
September 30, 1998

Personnel

- **Speech:** Dilek Hakkani, Madelaine Plauche, Zev Rivlin, Ananth Sankar, Elizabeth Shriberg, Kemal Sonmez, Andreas Stolcke, Gokhan Tur
- **Natural language:** David Israel, David Martin, John Bear
- **Video Analysis:** Bob Bolles, Marty Fischler, Marsha Jo Hannah, Bikash Sabata
- **OCR:** Greg Myers, Ken Nitz
- **Architectures:** Luc Julia, Adam Cheyer

SRI International

SRI News-on-Demand Highlights

- Focus on technologies
- New technologies: scene tracking, speaker tracking, flash detection, sentence segmentation
- Exploit technology fusion
- MAESTRO multimedia browser

Outline

- Goals for News-on-Demand
- Component Technologies
- The MAESTRO testbed
- Information Fusion
- Prosody for Information Extraction
- Future Work
- Summary

SRI International

High-level Goal

Develop techniques to provide direct and natural access to a large database of information sources through multiple modalities, including video, audio, and text.

Information We Want

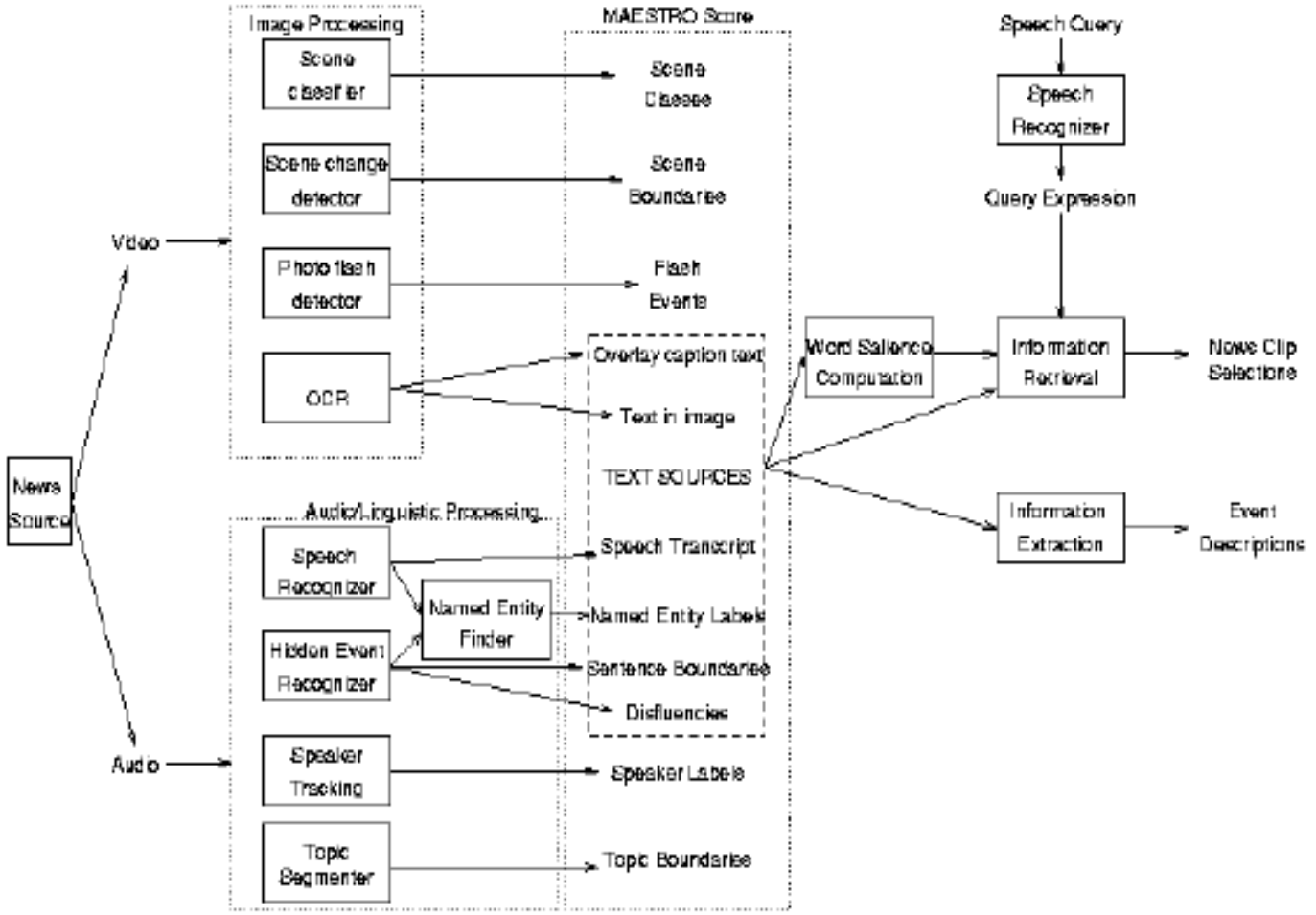
- Geographical location
- Topic of the story
- News-makers
- Who or what is in the picture
- Who is speaking

Component Technologies

- Speech processing
  - Automatic speech recognition (ASR)
  - Speaker identification
  - Speaker tracking/grouping
  - Sentence boundary/disfluency detection
- Video analysis
  - Scene segmentation
  - Scene tracking/grouping
  - Camera flashes
- Optical character recognition (OCR)
  - Video caption
  - Scene text (light or dark)
  - Person identification
- Information extraction (IE)
  - Names of people, places, organizations
  - Temporal terms
  - Story segmentation/classification

SRI International

Component Flowchart



SRI News-On-Demand System

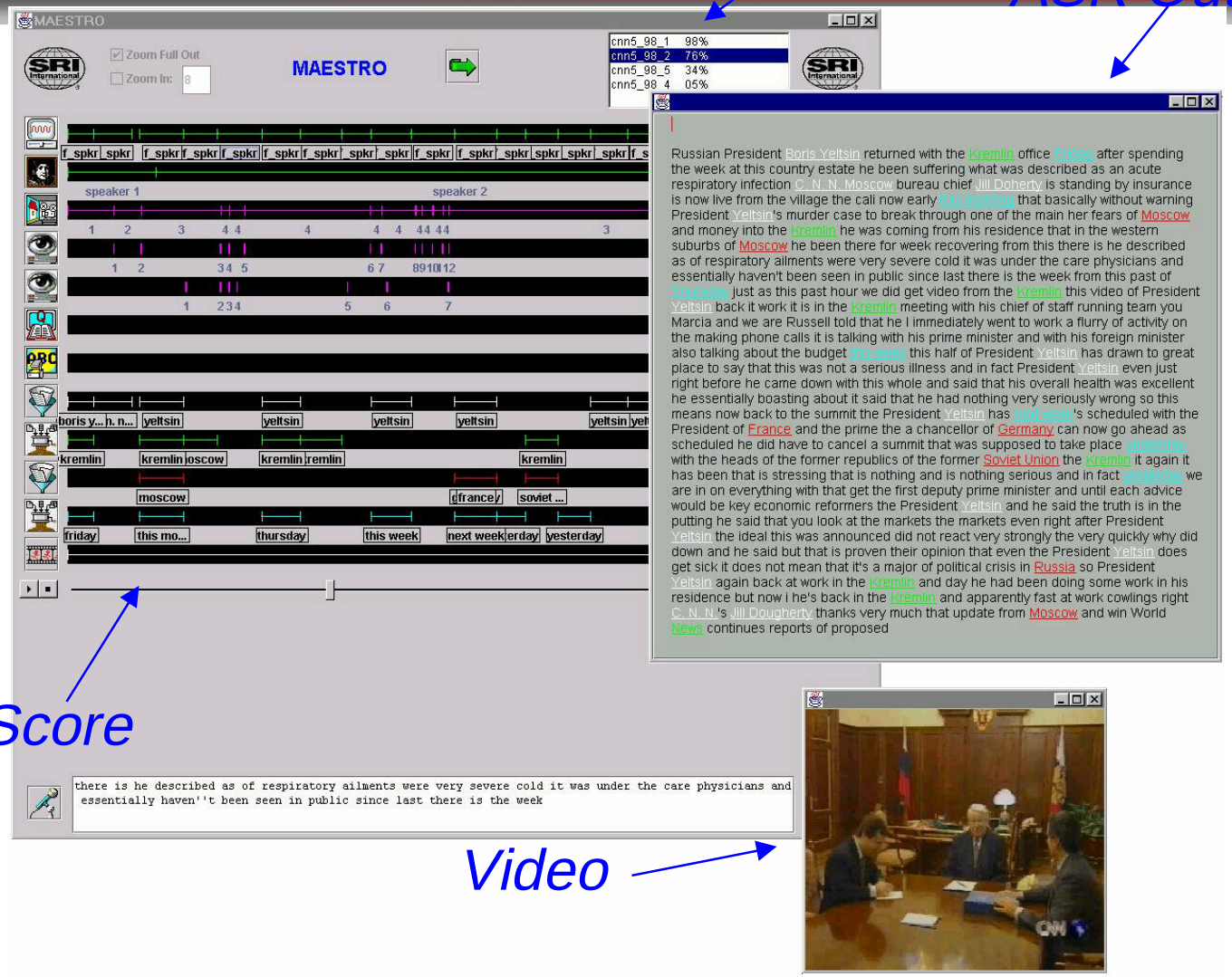
(MAESTRO)

MAESTRO

- Testbed for multimodal News-on-Demand Technologies
- Links input data and output from component technologies through common time line
- MAESTRO “score” visually correlates component technologies output
- Easy to integrate new technologies through uniform data representation format

SRI International

MAESTRO Interface



The Technical Challenge

- Problem: Knowledge sources are not always available or reliable
- Approaches
  - Make existing sources more reliable
  - Combine multiple sources for increased reliability and functionality (fusion)
  - Exploit new knowledge sources

Two Examples

- Technology Fusion: Speech recognition + Named entity finding = better OCR
- New knowledge source:
  Speech prosody for finding names and sentence boundaries

Fusion Ideas

- Use the names of people detected in the audio track to suggest names in captions
- Use the names of people detected in yesterday's news to suggest names in audio
- Use a video caption to identify a person speaking, and then use their voice to recognize them again

Information Fusion

Text Recog

"Moore" +

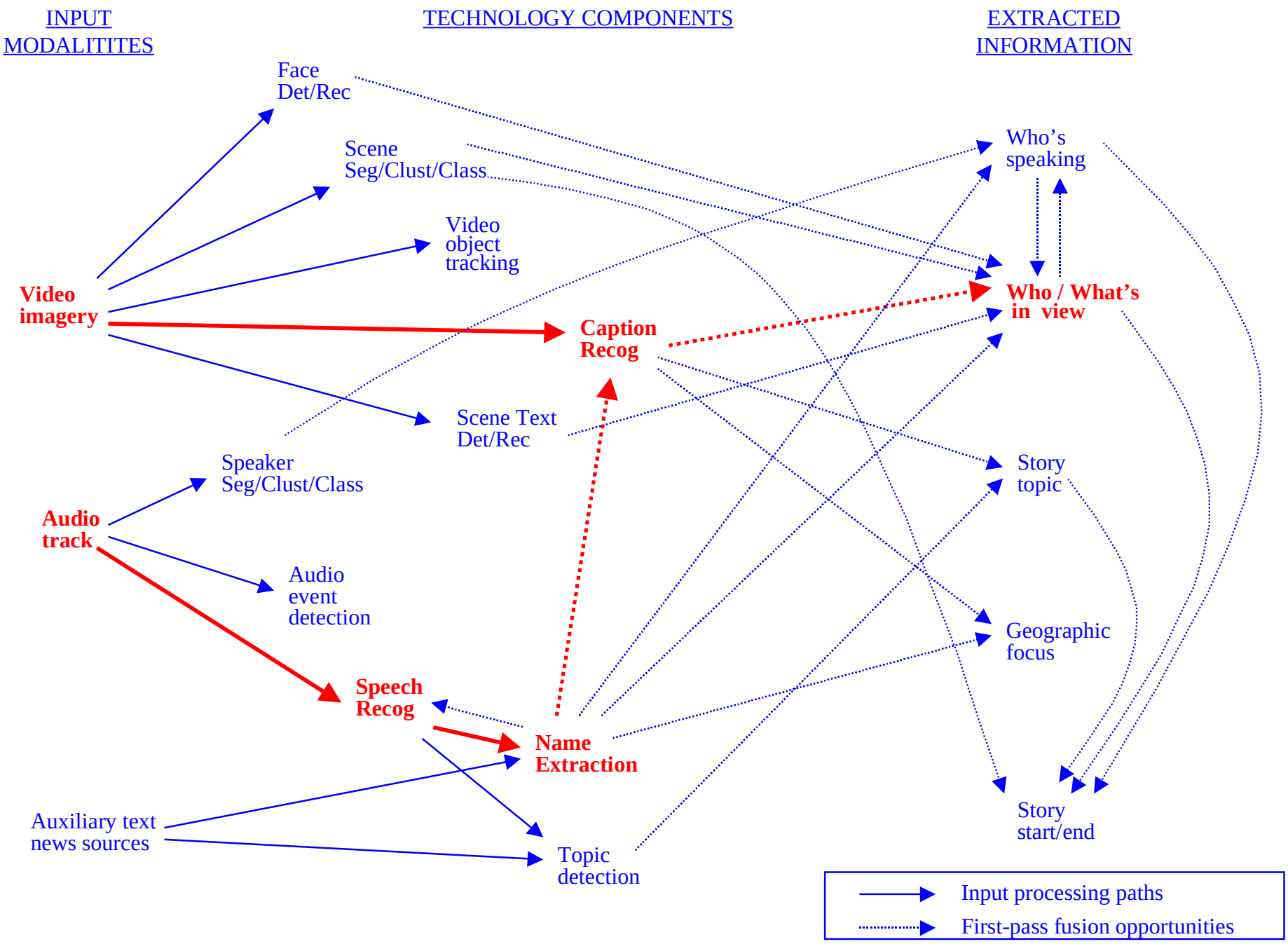
add to lexicon

ASR

moore

NL

"Moore"


INPUT MODALITITES
TECHNOLOGY COMPONENTS
EXTRACTED INFORMATION
Face Det/Rec
Scene Seg/Clust/Class
Video object tracking
Video imagery
Caption Recog
Scene Text Det/Rec
Speaker Seg/Clust/Class
Audio track
Audio event detection
Speech Recog
Name Extraction
Auxiliary text news sources
Topic detection
Who's speaking
Who / What's in view
Story topic
Geographic focus
Story start/end
Input processing paths
First-pass fusion opportunities

Augmented Lexicon Improves Recognition Results

Without lexicon: TONY BLAKJB WNITEE SIATEE

With lexicon: TONY BLAIR UNITED STATES

Prosody for Enhanced Speech Understanding

- Prosody = Rhythm and Melody of Speech
- Measured through duration (of phones and pauses), energy, and pitch
- Can help extract information crucial to speech understanding
- Examples: Sentence boundaries and Named Entities

Prosody for Sentence Segmentation

- Finding sentence boundaries important for information extraction, structuring output for retrieval
- Ex.: *Any surprises?*
  No. Tanks are in the area.
- Experiment: Predict sentence boundaries based on duration and pitch using decision trees classifiers

Sentence Segmentation: Results

- Baseline accuracy = 50% (same number boundaries & non-boundaries)
- Accuracy using prosody = 85.7%
- Boundaries indicated by: long pauses, low pitch before, high pitch after
- Pitch cues work much better in Broadcast News than in Switchboard

SRI International

Prosody for Named Entities

- Finding names (of people, places, organizations) key to info extraction
- Names tend to be important to content, hence prosodic emphasis
- Prosodic cues can be detected even if words are misrecognized: could help find new named entities

SRI International

Named Entities: Results

- Baseline accuracy = 50%
- Using prosody only: accuracy = 64.9%
- N.E.s indicated by
  - longer duration (more careful pronunciation)
  - more within-word pitch variation
- Challenges
  - only first mentions are accented
  - only one word in longer N.E. marked
  - non-names accented

SRI International

Using Prosody in NoD: Summary

- Prosody can help information extraction independent of word recognition
- Preliminary positive results for sentence segmentation and N.E. finding
- Other uses: topic boundaries, emotion detection

SRI International

Ongoing and Future Work

- Combine prosody and words for name finding
- Implement additional fusion opportunities:
  - OCR helping speech
  - speaker tracking helping topic tracking
- Leverage geographical information for recognition technologies

Conclusions

- News-on-Demand technologies are making great strides
- Robustness still a challenge
- Improved reliability through data fusion and new knowledge sources

SRI International