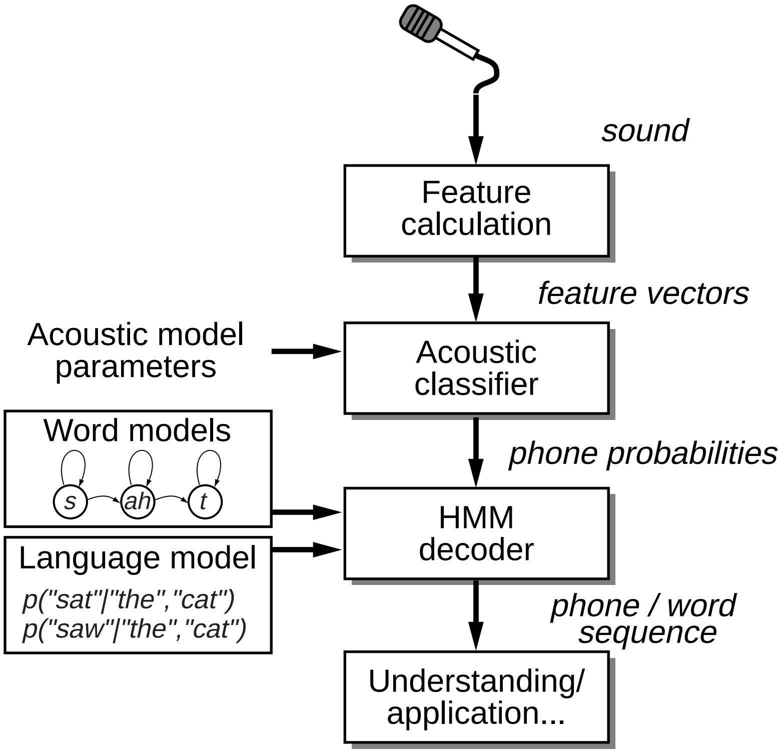
Speech interfaces:
A survey and some current projects

Dan Ellis & Nelson Morgan
International Computer Science Institute
Berkeley CA
{dpwe,morgan}@icsi.berkeley.edu

Outline

1. Speech recognition: the state of the art
2. Current projects at ICSI
3. Conclusions

Speech recognition



sound

Feature calculation

feature vectors

Acoustic model parameters

Acoustic classifier

phone probabilities

Word models

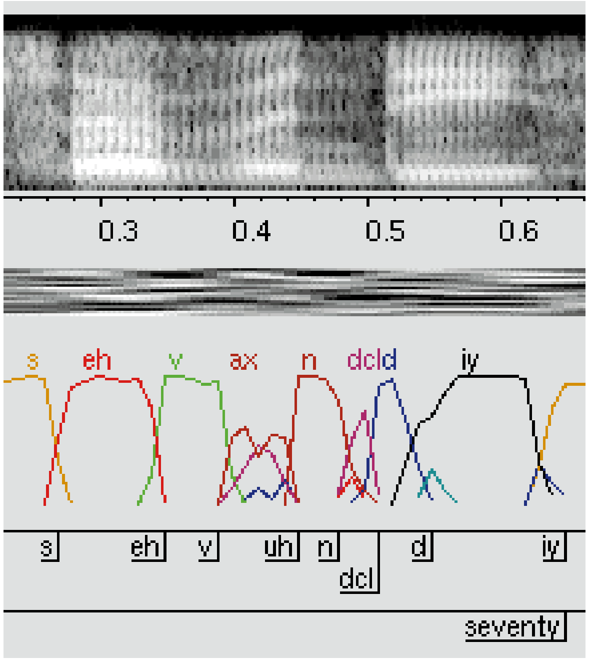
Language model

p("sat"|"the","cat")
p("saw"|"the","cat")

HMM decoder

phone / word sequence

Understanding/ application...



- **Elements of a recognizer:**
  - feature design
  - acoustic modeling
  - pronunciation/language modeling

} data!

How good is speech recognition?

- **Standard measure is word error rate (WER):**
  - dictation (close-mic): 2-5%
  - broadcast news: ~15%
  - telephone conversations: ~30%

F0: THE VERY EARLY RETURNS OF THE NICARAGUAN PRESIDENTIAL ELECTION SEEMED TO FADE BEFORE THE LOCAL MAYOR ON A LOT OF LAW

F4: AT THIS STAGE OF THE ACCOUNTING FOR SEVENTY SCOTCH ONE LEADER DANIEL ORTEGA IS IN SECOND PLACE THERE WERE TWENTY THREE PRESIDENTIAL CANDIDATES OF THE ELECTION

- **What are the problems?**
  - acoustic variability (noise, channel)
  - speech variability (accent, manner)
  - exploiting linguistic constraints
  - speech understanding...

Frontiers of speech recognition

- **Acoustic modeling**
  - beyond head-mounted mics
  - background noise (mobile phones)
  - speech in mixtures (broadcast)
  - → robust feature design, better statistical models
- **Speaking styles**
  - coarticulation
  - pronunciation variability
  - speaking styles
  - → lump into acoustic model, more training data, better pron. models, context-dep. models
- **Linguistic constraints**
  - ‘inferred’ words
  - ambiguity
  - → higher-order n-grams (more training data), tree grammars

Applications of speech recognition

- **Command & control**
  - more or less constrained
- **Dictation**
  - large vocabulary
  - known, co-operative user
- **Voice response systems**
  - dialog & speech understanding
  - robustness!
  - human factors (timing, barge-in etc.)
- **Information extraction & retrieval**
  - multimedia archive retrieval
  - live ‘listener’

Outline

1 **Speech recognition: the state of the art**

2 **Current projects at ICSI**

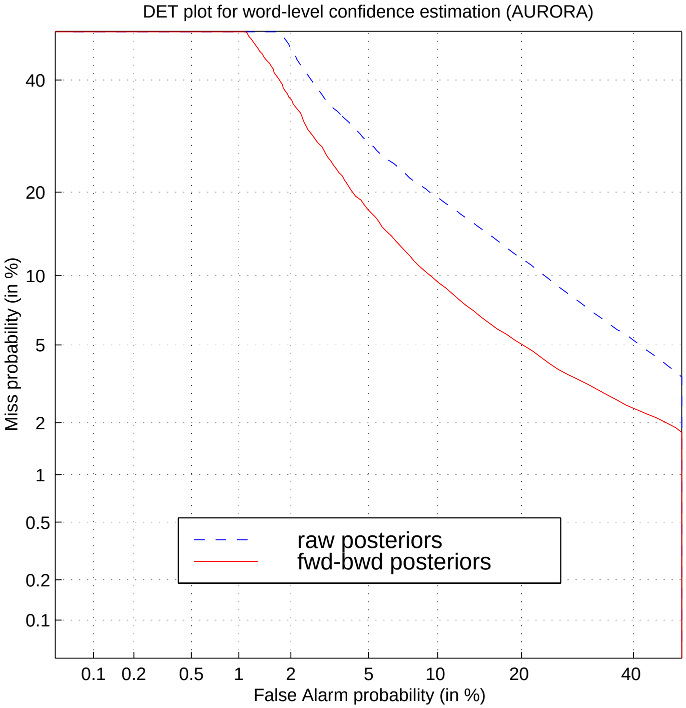
- Recognizer confidence measures
- Combining information sources
- The meeting recorder
- Audio content-based retrieval

3 **Conclusions**

Recognizer confidence measures

(Warner Warren, Andy Hatch, Eric Fosler + SRI)

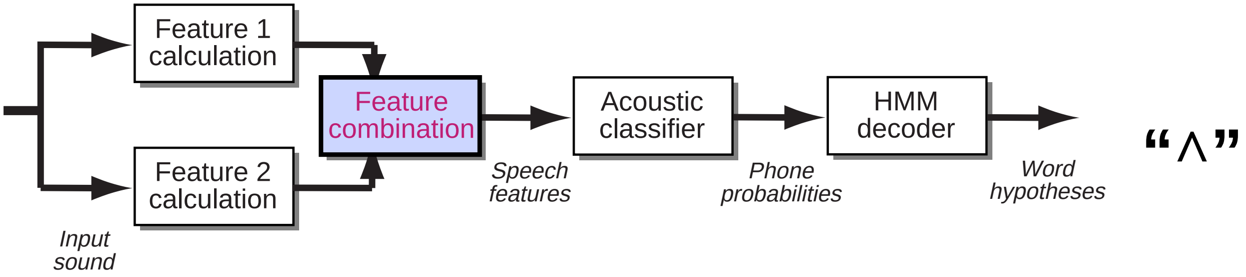
- **Knowing which words are wrong can help**
  - hard to tell because recognition only *just* works
- **Average per-phone entropy + re-estimation:**



- **Use for combining recognizer outputs**

Combination schemes

(Mike Shire, Barry Chen + Michael Jordan)



Feature 1 calculation

Feature 2 calculation

Feature combination

Acoustic classifier

HMM decoder

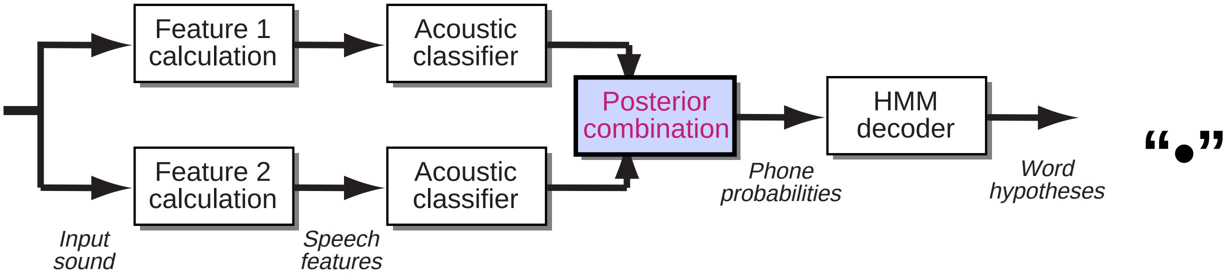
Input sound

Speech features

Phone probabilities

Word hypotheses

“^”



Feature 1 calculation

Feature 2 calculation

Acoustic classifier

Acoustic classifier

Posterior combination

HMM decoder

Input sound

Speech features

Phone probabilities

Word hypotheses

“•”

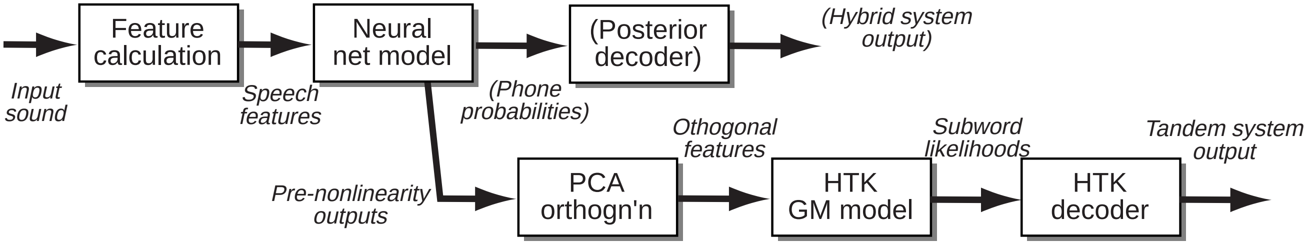
- **How best to combine different feature streams?**

Features	*Avg. WER*
plp	8.9%
msg	9.5%
plp ^ msg	8.1%
plp • msg	7.1%

Tandem acoustic modeling

(with Hermansky et al., OGI)

- **ICSI pioneered ‘hybrid-connectionist’ ASR; Can it be combined with conventional models?**

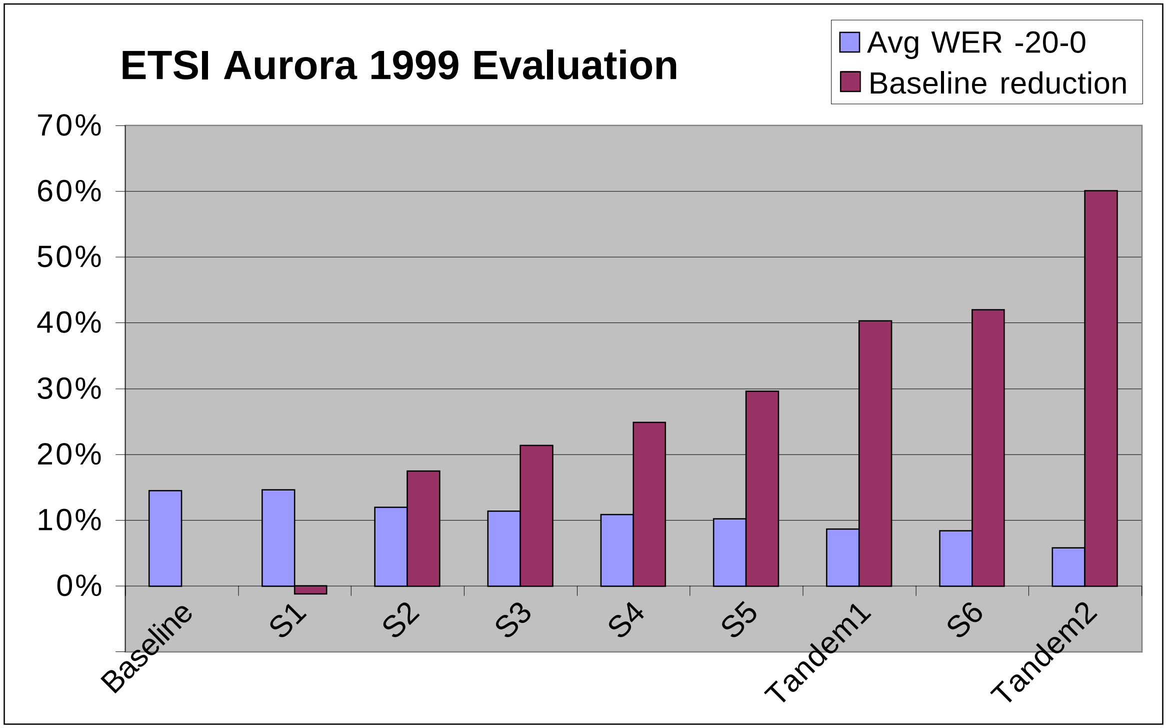


- **Result: better performance than either alone!**
    - neural net & Gaussian mixture models extract different information from training data

System-features	*Avg. WER*
HTK-mfcc	13.7%
Hybrid-mfcc	9.3%
Tandem-mfcc	7.4%
Tandem-plp+msg	6.4%

Aurora “Distributed SR” evaluation

- **Organized by ETSI (European Telecoms. Standards Institute)**



  - Tandem systems from OGI-ICSI-Qualcomm

The meeting recorder project

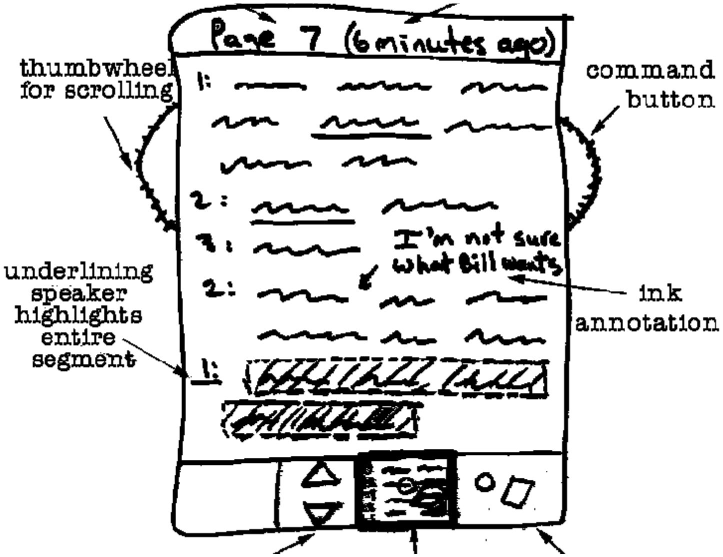
(Adam Janin, Eric Fosler + UW, SRI, UPM, James Landay)

- **Idea: PDA records meetings to replace / enhance note-taking**
- **First task: Collect a training corpus**

- **Related to DARPA Communicator, SmartKom**

Meeting recorder: Research areas

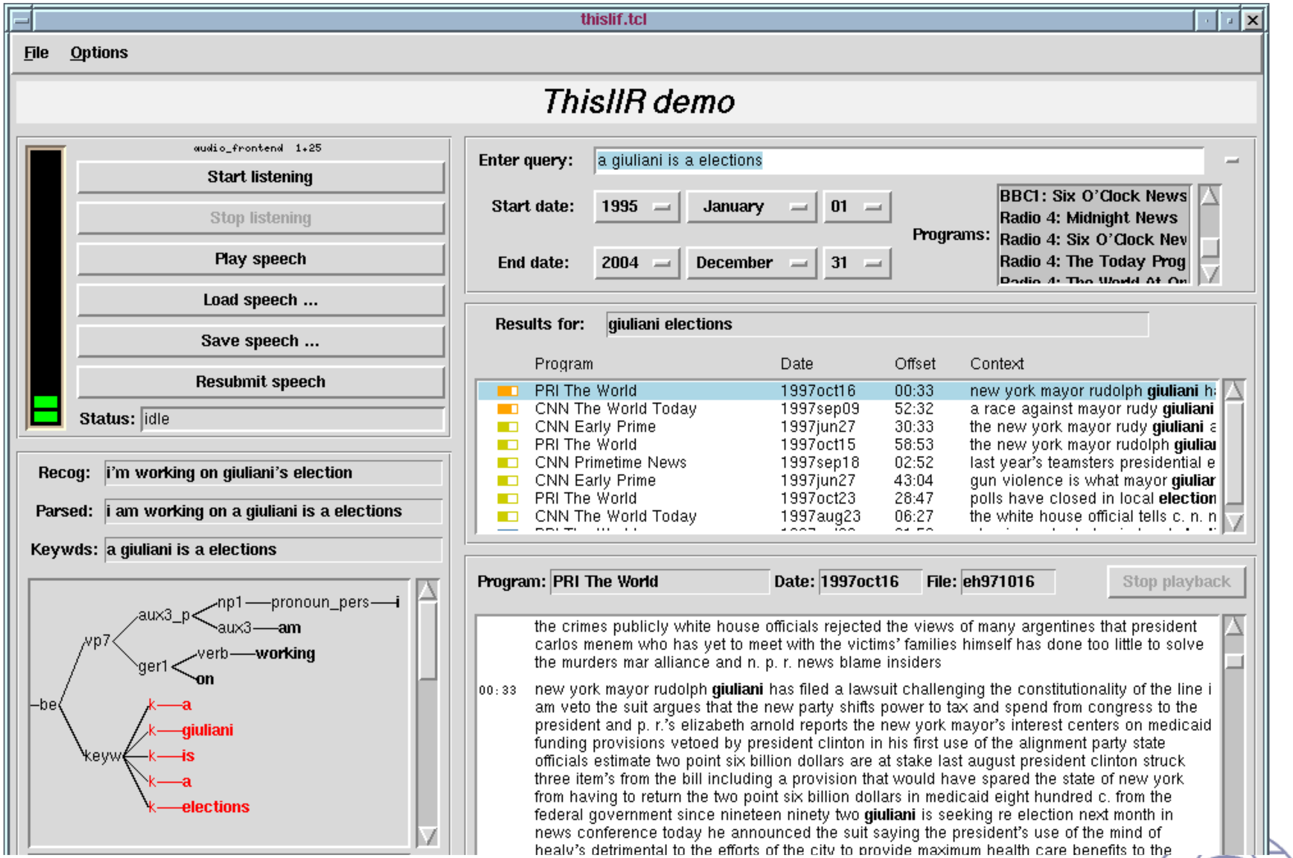
- **Audio recognition**
  - recognition from noisy microphones
  - speaker identification & tracking
  - nonspeech events
- **Indexing application**
  - understanding the structure of meetings
  - information retrieval
  - user interface



Audio content-based retrieval

(with Sheffield, Cambridge, BBC, Avideh Zakhor)

- **Idea: speech recognition output as indexes for broadcast news**
  - useful even with 15-30% WER



Status: idle

Recog: i'm working on giuliani's election

Parsed: i am working on a giuliani is a elections

Keywds: a giuliani is a elections

Enter query: a giuliani is a elections

Start date: 1995 January 01

End date: 2004 December 31

Programs: BBC1: Six O'Clock News; Radio 4: Midnight News; Radio 4: Six O'Clock Nev; Radio 4: The Today Prog; Radio 4: The World At On

Results for: giuliani elections

	Program	Date	Offset	Context
🟧	PRI The World	1997oct16	00:33	new york mayor rudolph **giuliani** h
🟧	CNN The World Today	1997sep09	52:32	a race against mayor rudy **giuliani**
🟨	CNN Early Prime	1997jun27	30:33	the new york mayor rudy **giuliani** a
🟨	PRI The World	1997oct15	58:53	the new york mayor rudolph **giuliai**
🟨	CNN Primetime News	1997sep18	02:52	last year's teamsters presidential e
🟨	CNN Early Prime	1997jun27	43:04	gun violence is what mayor **giuliar**
🟨	PRI The World	1997oct23	28:47	polls have closed in local **election**
🟨	CNN The World Today	1997aug23	06:27	the white house official tells c. n. n

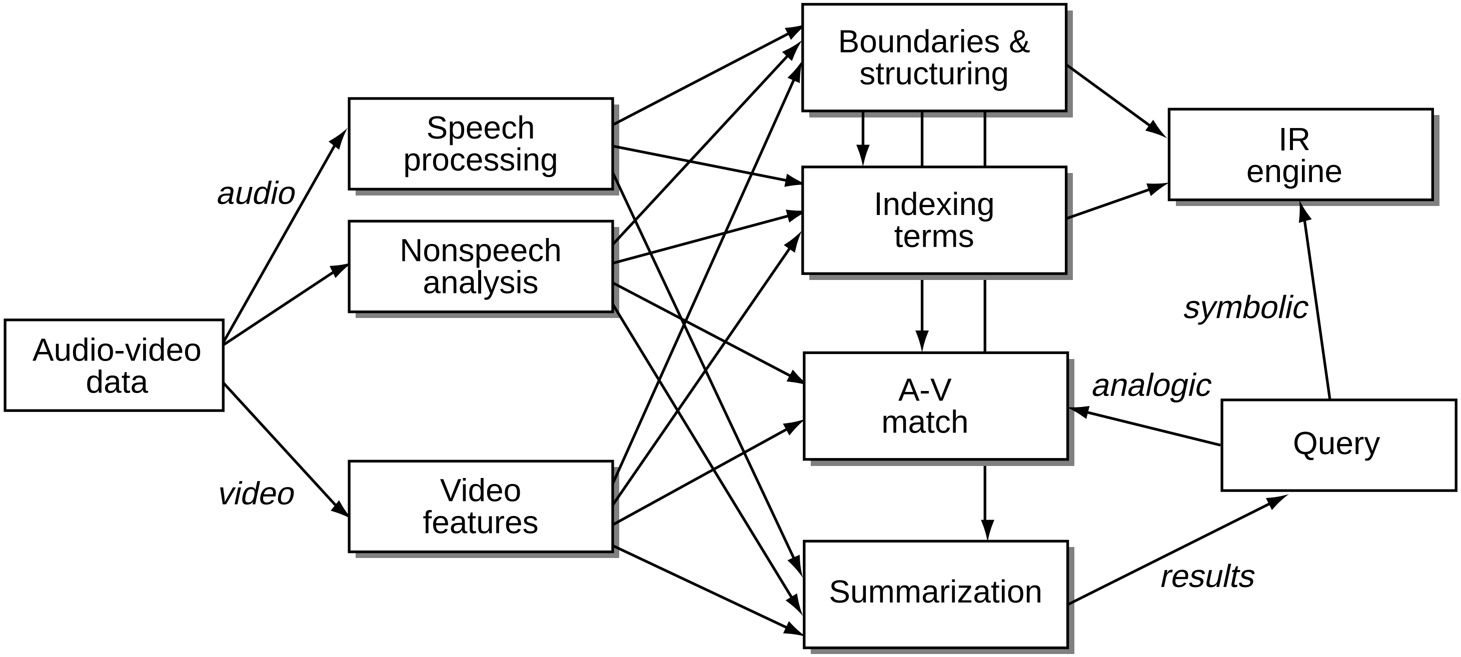
Program: PRI The World Date: 1997oct16 File: eh971016

the crimes publicly white house officials rejected the views of many argentines that president carlos menem who has yet to meet with the victims' families himself has done too little to solve the murders mar alliance and n. p. r. news blame insiders

00:33 new york mayor rudolph **giuliani** has filed a lawsuit challenging the constitutionality of the line i am veto the suit argues that the new party shifts power to tax and spend from congress to the president and p. r.'s elizabeth arnold reports the new york mayor's interest centers on medicaid funding provisions vetoed by president clinton in his first use of the alignment party state officials estimate two point six billion dollars are at stake last august president clinton struck three item's from the bill including a provision that would have spared the state of new york from having to return the two point six billion dollars in medicaid eight hundred c. from the federal government since nineteen ninety two **giuliani** is seeking re election next month in news conference today he announced the suit saying the president's use of the mind of healy's detrimental to the efforts of the city to provide maximum health care benefits to the

Audio-video organization & retrieval

- **Proposed project:**



- **Synergy between audio & video features**
- **Query by terms or by examples**
- **Recovering temporal structuret**

3

Conclusions

- **Speech recognition is now practical**
  - .. but still plenty of problems
- **Ongoing research in speech recognition**
  - recognition in demanding conditions
  - understanding / discourse a big issue
- **Multimodal information retrieval**
  - forgiving & fertile research area