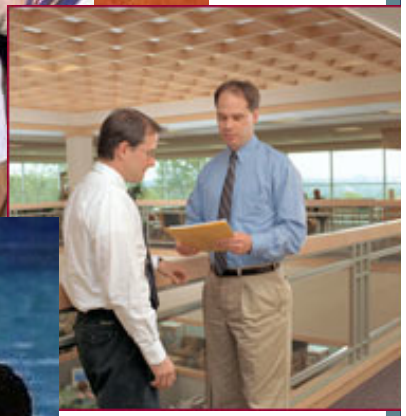
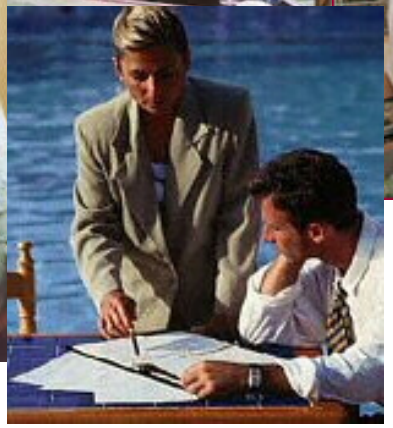
Understanding and Modeling Communication Scenes

Hervé Bourlard
IDIAP Research Institute, Switzerland
Swiss Federal Institute of Technology, Lausanne
bourlard@idiap.ch

IEEE/ACL 2006 Workshop on Spoken Language Technology
Aruba, December 10-13, 2006

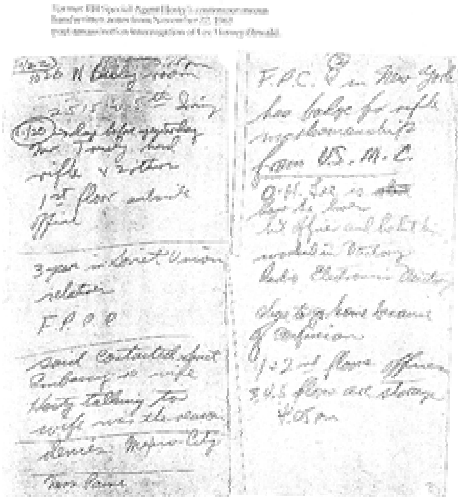
Communication Scenes





Meetings, meetings!!!!

- All important decisions are taken in (face-to-face) meetings.
- Resulting in many meetings
- >12 million business meetings daily!!!!
- 37% of employee time spent in meetings
- Often not as efficient as expected.



Context

- **EU Integrated Projects (15 partners)**
  - *AMI (Augmented Multi-party Interaction), Jan'04-Dec'06*
  - *AMIDA (Augmented multi-party Interaction with Distance Access), Nov'06-Oct'09*
  - *www.amiproject.org*
- **Swiss Research Network IM2**
- **DTO VACE-III**
  - *ROADMAP: RObust Automatic Detection of Meeting-events with Audiovisual Perception, Sep'06-Aug'07*

AMI Consortium

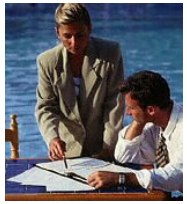
IM INTERACTIVE MULTIMODAL INFORMATION MANAGEMENT

DTO

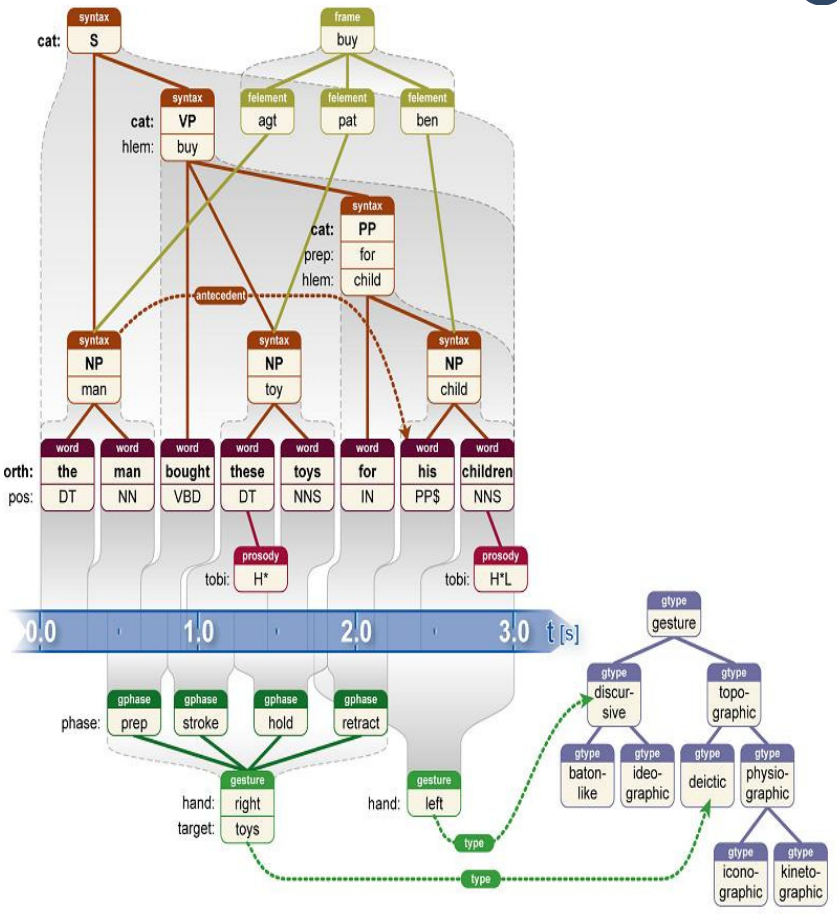
Current Framework: Instrumented meeting rooms

- Instrumented meeting/seminar rooms at several sites (IDIAP, UniEdin, Twente, ICSI)
- Typical meeting rooms
  - 4 close-, 2 wide-view cameras
  - 4 headset, 8 array microphones
  - Data projector capture
  - Whiteboard capture
  - Digital pen capture
  - Extra site-dependent devices (e.g., second microphone array, lapel mics)

AMI Corpus



- **100 hour corpus of multiparty meetings**, recorded in 3 instrumented meeting rooms
- **Annotated in terms of:**
  - Audio (checked) transcription
  - Named entities
  - Dialogue acts
  - Topic segmentation
  - Extractive and abstractive summaries
  - Hand gestures
  - (limited) Head gestures
  - Location of person on video
  - (limited) gaze direction
  - Movement around room
- Freely available under the Ceative Commons 'share alike' licence http://corpus.amiproject.org



Multiple layers/tiers, hierarchical, support for time aligned and general content

Inderdisplinary Research Problem

1. **Signal processing** and machine learning: making sense of communication scenes starting from the signals
2. **Multimodal group modeling**
3. **Linguistic and discourse modelling**: understanding the content of the recognized signals
4. Moving from qualitative to quantitative models of social dynamics
5. Applications that correspond to the needs and requirements of people
6. Links with many other aspect of social interactions (social networks)

I will try to briefly illustrate some of these aspects….

Current Evaluation Efforts

- **Component evaluations**
  - Signal processing: speech recognition, face detection, tracking, gesture recognition,....
  - Content abstraction: summarization, topic segmentation, dialogue act segmentation, ...
- **System-level evaluations**
  - Browser evaluation test
  - Task-based evaluation
- Participation in and provision of data to: NIST RT, CLEAR, PASCAL Speech Separation, CLEF

Signal Processing
Audio-video processing

- **Defined according to core problems:**
  - What are they saying? (speech recognition)
  - How are they saying it? (prosody)
  - What are they doing? (action recognition)
  - Where are they going? (location tracking)
  - How are they feeling? (emotional state)
  - Where are they looking? (focus of attention)
  - Who are they anyway? (person identification)
- **Challenges:**
  - Scalable models and online/real-time algorithms
  - Adaptation to new domains without extensive data collection

Signal Processing Evaluations

- Multiparty speech recognition (RT)
- Keyword spotting
- Face detection (CLEAR)
- Multi-person tracking
- Gesture/action recognition
- Focus of attention recognition (CLEAR)
- Speaker activity detection (RT)
- Speaker diarization ('who spoke when') (RT)

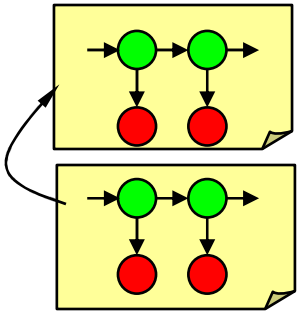
	Internal Evaluation	International Evaluation	Contributing Data
ASR	✓	NIST	✓
KWS	✓	Planned (Nov.2006)	
SEG	✓	NIST	✓
ID/LOC	✓	VACE (CLEAR) ARDA	✓
FOA	✓	Part of VACE-III (initiated and managed by IDIAP)	DTO ✓
GAA	✓		

- *ASR: Automatic speech recognition*
- *KWS: keyword spotting*
- *SEG: speaker segmentation*
- *ID/LOC: id. and localization/tracking*
- *FOA: focus of attention*
- *GAA: gesture and action recognition*

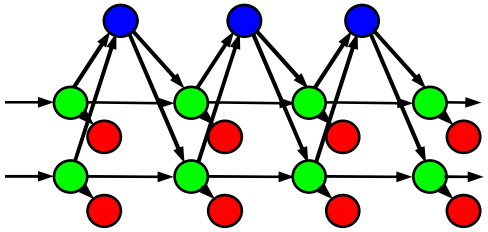
Content Abstraction

Multimodal Group Modeling

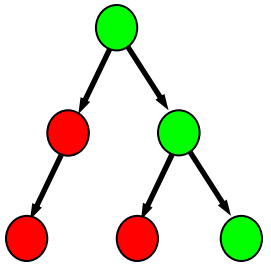
Group Action Modeling ⇨ **Multi-layer HMMs**



Dominance Modeling ⇨ **Team-Player Influence Model**

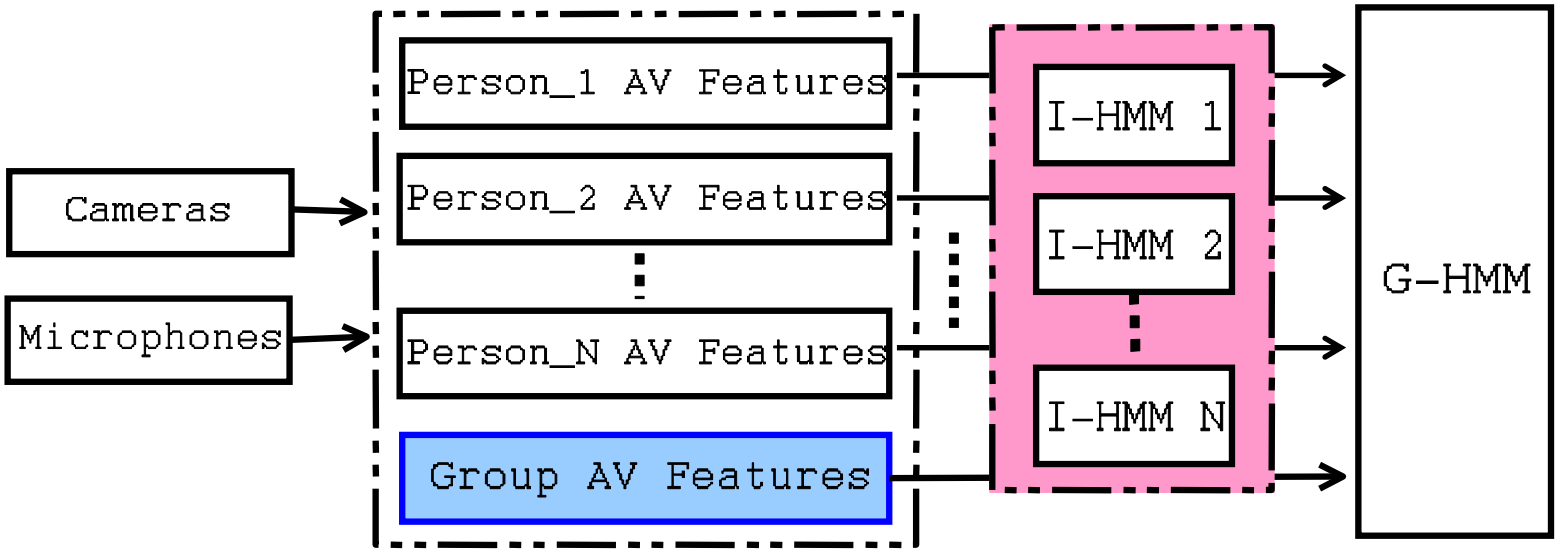


Unusual Event Modeling ⇨ **Semi-Supervised Adapted HMMs**



Group Action Modeling

Multilayered Approaches



- **Smaller** observation spaces.
- Individual layer HMM is **person-independent**.
- Group level HMM is **less sensitive** to low-level audio-visual features.
- Each layer trained **independently**

Group Action Modeling

Individual and Group Actions

Group actions

- Discussion
- Monologue
- Monologue + Note-taking
- Note-taking
- Presentation
- Presentation + Note-taking
- Whiteboard
- Whiteboard + Note-taking

individual actions

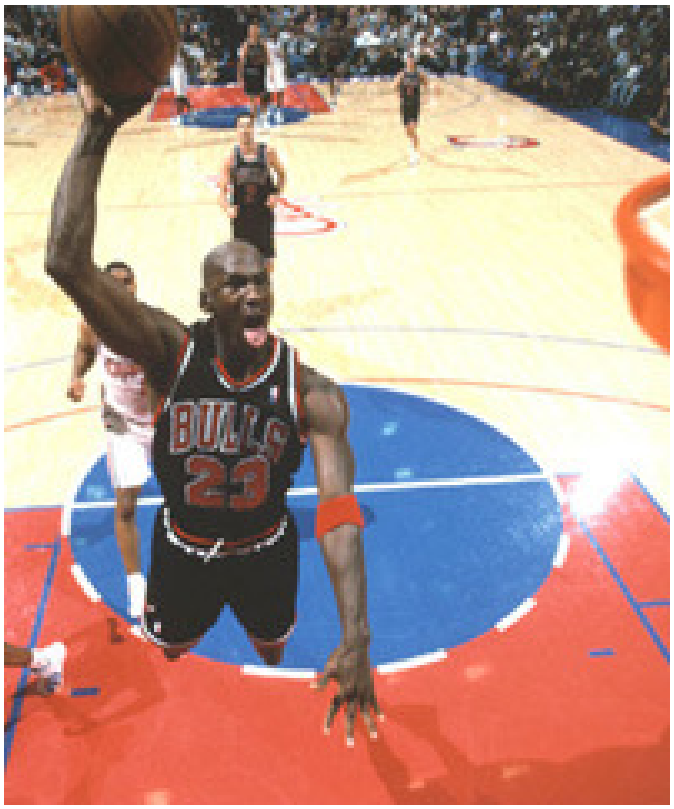
I = { ‘speaking’, ‘writing’, ‘idle’}

group actions

Person 1	S		S		W		W
Person 2		W		S		W		W
Person 3		W		S		S		W	
Person 4			S		W		S
Presentation				Used		
Whiteboard					Used
Group Action	Monologue1 + Note-taking	Discussion	Presentation + Note-taking	Whiteboard + Note-taking

Group Action Modeling

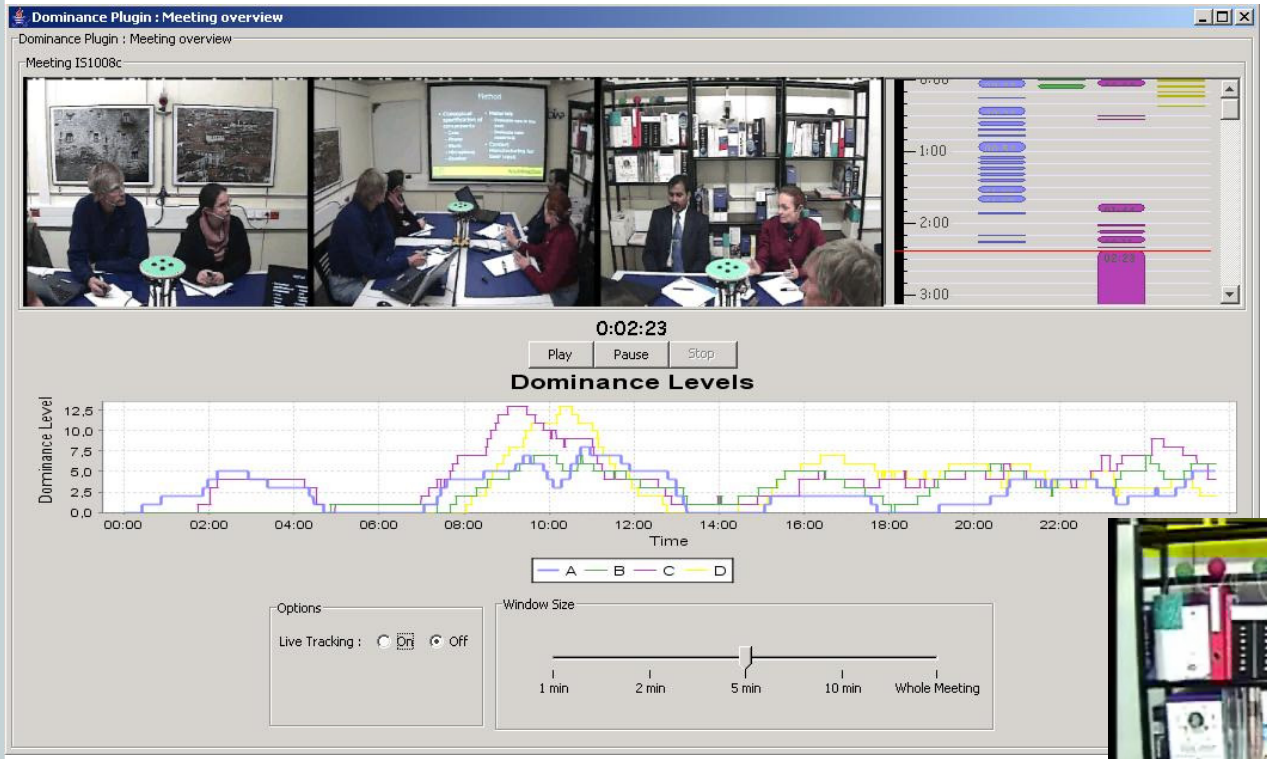
Learning Influence in Human Interactions

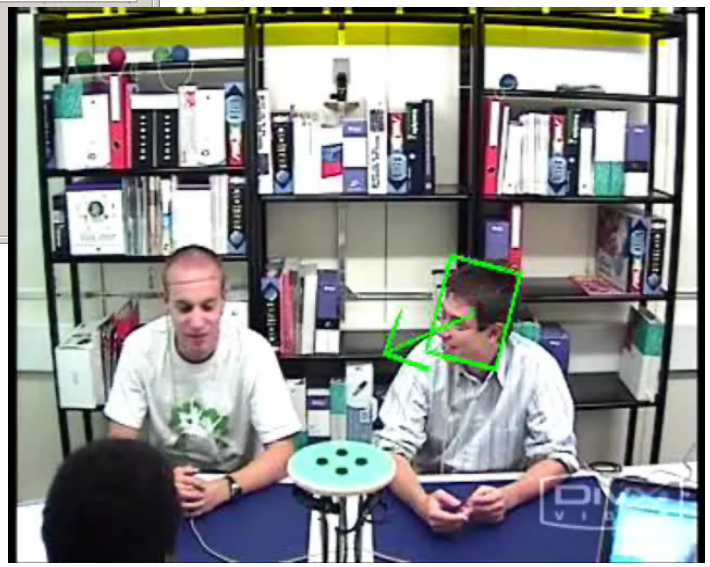


- What is the *influence* of each person in a meeting? Who is the *dominant* one? Who *drives the decisions* taken in a meeting?
- Team-player Influence Model (DBNs)

Group Action Modeling

Dominance and Focus of Attention





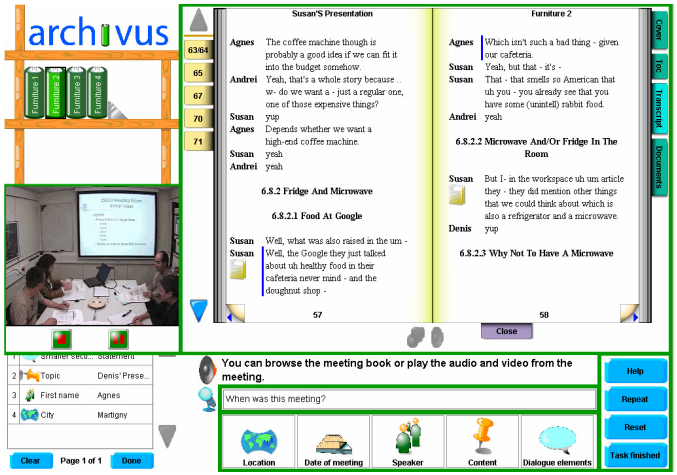
Content Abstraction Evaluations

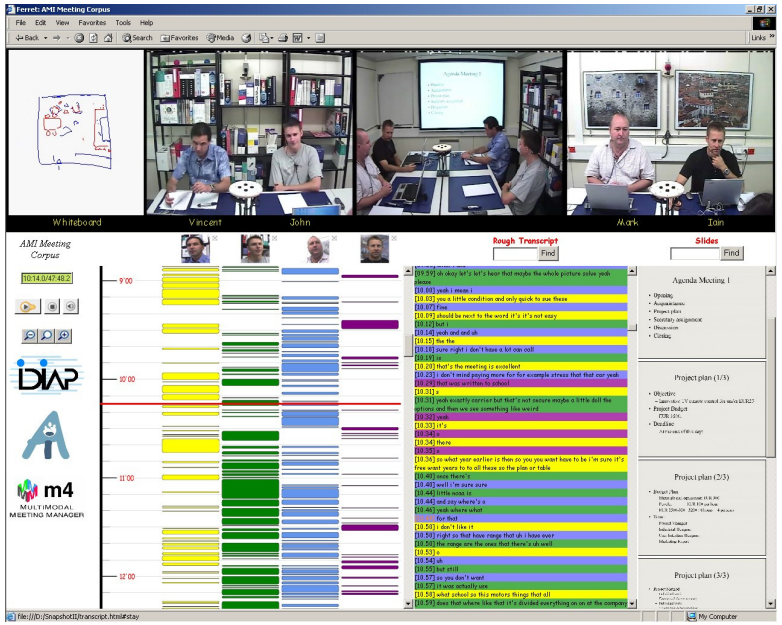
- Content abstraction (completed)
  - Dialogue act segmentation and tagging
  - Extractive summarization
  - Chunking (shallow parsing)
  - Named entity identification
  - Topic Segmentation
  - Addressee classification
- Evaluations under development include
  - Abstractive summarization
  - Decision points
- Evaluation metrics are not straightforward!

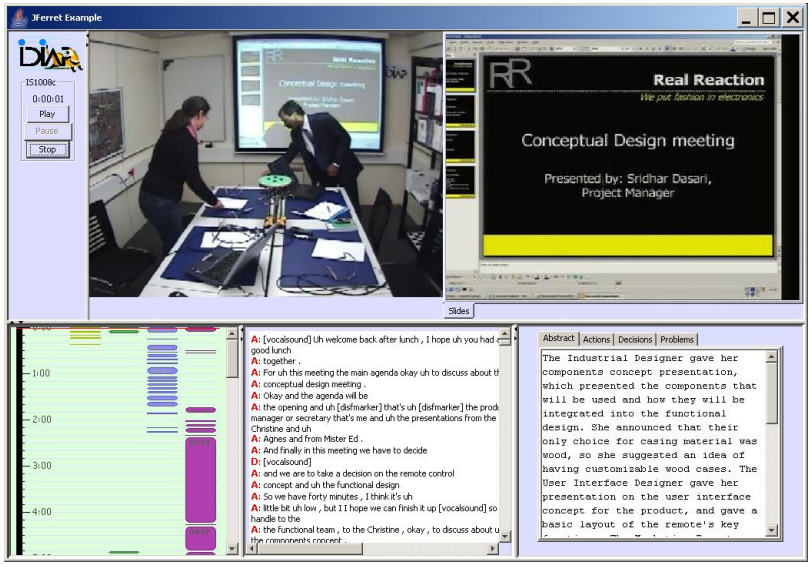
Interface and Information Retrieval

Meeting Browsers

- JFerret: extremely flexible, enabling almost any user interface to be composed, using combination of plug-in modules
- 13 applications currently using JFerret
- **Standard evaluation protocol:**
  BET (Browser Evaluation Test)



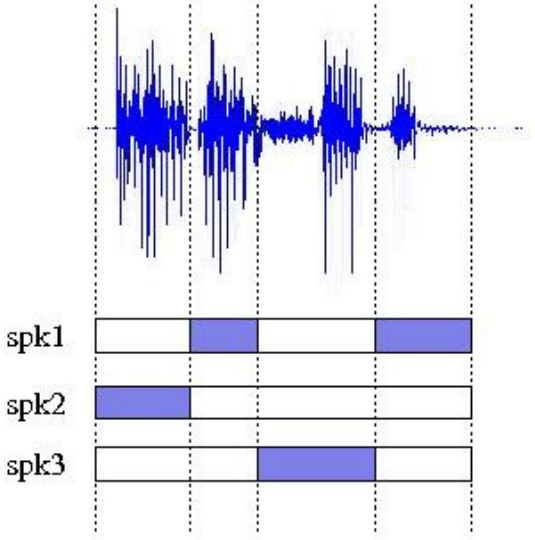


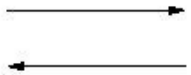


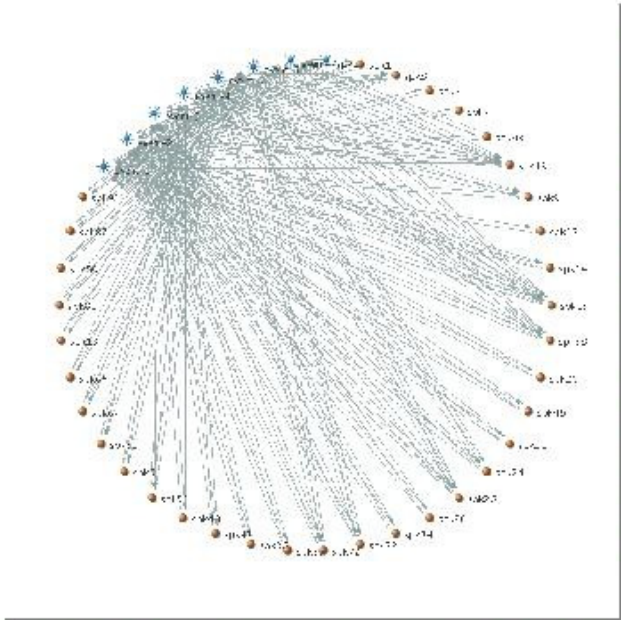
Looking into to future…

Joint (AV) Participant identification and Social Network (Role) Inference:

- Application of Social Network Analysis (SNA) for role and social identity recognition.
- Joint use of social and audio-vidual features for mutual enhancement of participant and social role identification.







Current State

- **Automatic processing of communication scenes in constrained environments**
  - Speech recognition from distant microphones
  - Multimodal tracking of people in meeting rooms
  - Automatic segmentation by speaker, dialogue acts, topic, meeting phase
  - Automatic summarization
- **Integration into systems**
  - Indexing search, browsing of archives
  - Limited online processing (see AMIDA)

Future Work

- Technology to create archives
  - Efficient access (online and offline) to multimodal meeting recordings
- Technology to create presence
  - Presence as realtime communication of state
  - Shared multimodal workspaces
- Technology to create context
  - Automatic incorporation of multiple information sources during a meeting

Thank you for your attention