Real-time Monitoring of Participants' Interaction in a Meeting using Audio-Visual sensors

Carlos Busso, Panayiotis G. Georgiou, and Shrikanth S. Narayanan

Speech Analysis and Interpretation Laboratory (SAIL)
Viterbi School of Engineering,
University of Southern California,

Presented by Carlos Busso

Outline

- Introduction
- Smart room
- Multimodal fusion
- Participants' interaction
- Conclusions

Introduction

Motivation

- Meetings are important for any organization
- Automatic annotations of human interaction will provide better tools for analyzing teamwork and collaboration strategies
- Examples of application in which monitoring human interaction is very useful are summarization, retrieval and classification of meetings

Goals

- Infer meta-information from participants in a multiperson meeting
- To monitor and track the behaviors, strategies and engagements of the participants
- Infer interaction flow of the discussion

Introduction

Approach

- Use an smart environment equipped with audio-visual sensors to get the annotations
- Extract high-level features from automatic annotations of speaker activity (e.g. number and average duration of each turn)

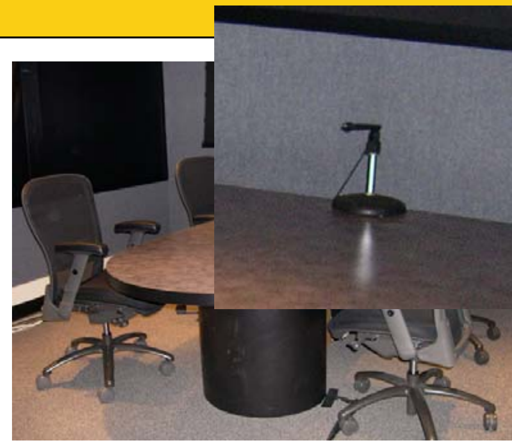
Related work

- Smart room [Checka,2004] [Gatica-Perez,2003] [Pingali,1999]
- Monitoring human interaction [McCowan,2005] [Banerjee,2004] [Zhang,2006] [Basu,2001]

Outline

- ✓ Introduction
- ✓ Smart room
- ➢ Multimodal fusion
- ➢ Participants' interaction
- ➢ Conclusions

Smart room



- Visual
  - 4 firewire CCD cameras
    - Participants' location
  - 360^{o} Omnidirectional camera
    - Angles of detected faces
- Audio
  - 16-channel microphone array
    - Speech source location
  - Directional microphone (SID)
    - Speaker identity

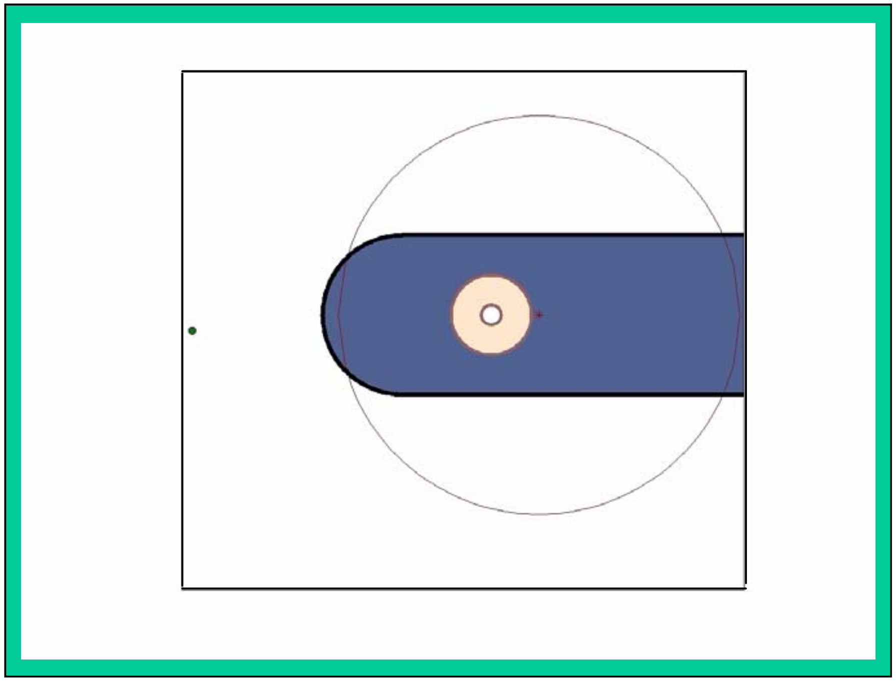
SAIL

Outline

- ✓ Introduction
- ✓ Smart room
- ✓ Multimodal fusion
- ➢ Participants' interaction
- ➢ Conclusions

Multimodal fusion

- Participants' location (visual modalities)
  - Participants are modeled with a Gaussian distribution
  - A background Gaussian model is adapted to the measurements to sequentially detect the participants
  - Position of the ceiling cameras are corrected using the location of the detected faces
  - Two participants cannot be too close
  - Participants are removed when measurements are not assigned to them



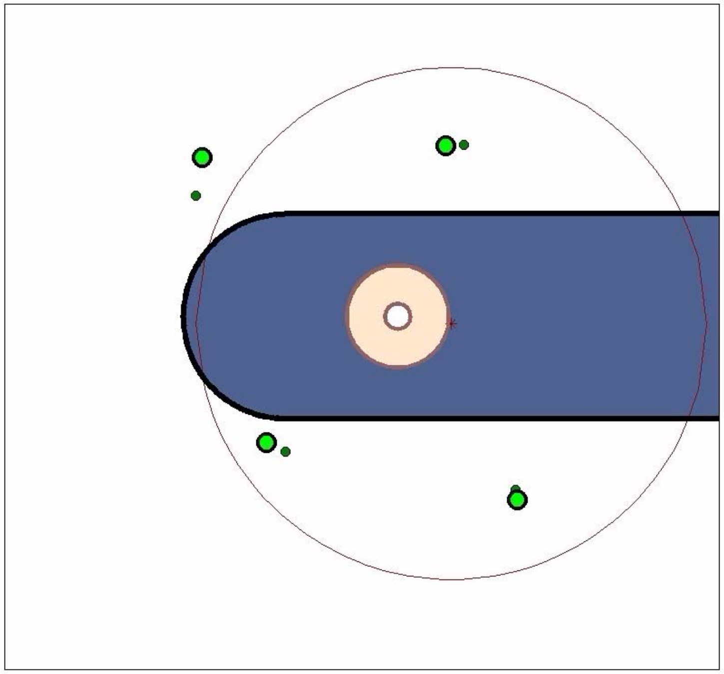
Detected participants
Measurements from ceiling cameras

Multimodal fusion

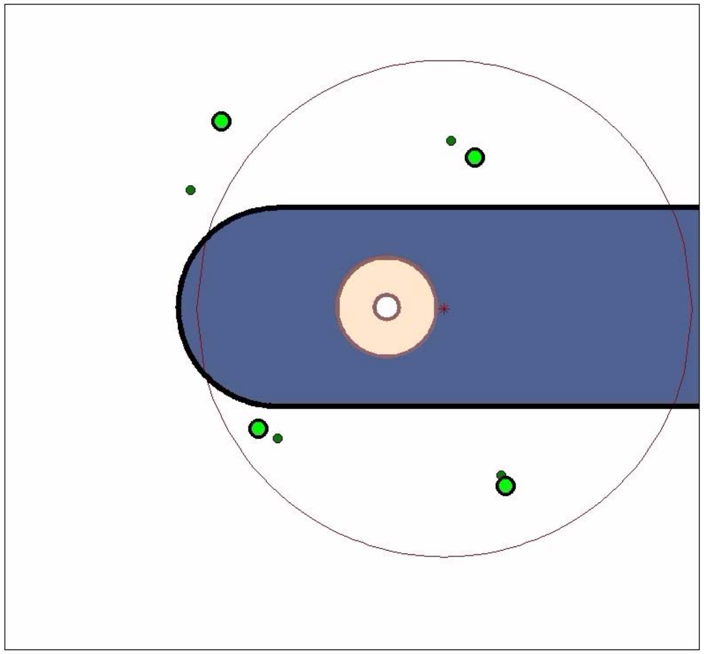
- Robustness of using multimodal sensors

Detected participants
Measurements from ceiling cameras

Ceiling cameras



Ceiling and omnidirectional cameras



Multimodal fusion

- Speaker' detection (MA + Speaker ID)
  - Who is speaking?
  - Mahalanobis distance between acoustic source and position of the participants $P(S_i \mid X_{MA})$
  - Speaker ID is also used to detect the active speaker $P(S_j \mid X_{SID})$

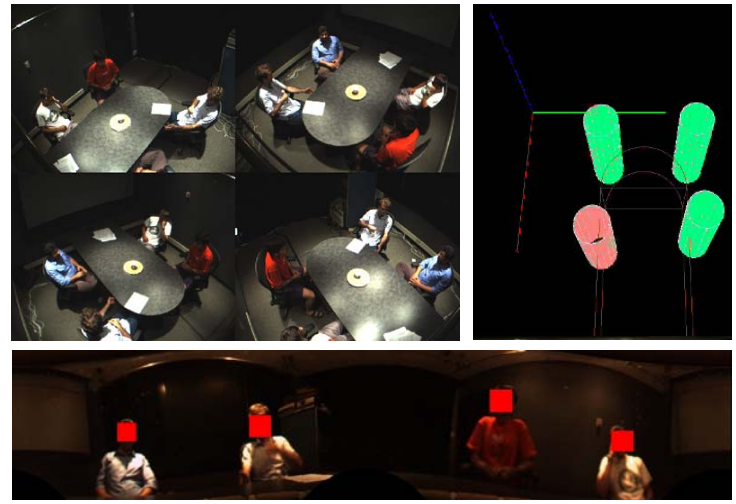
- Participants' identification (Speaker ID)
  - Participants' ID and seating arrangement
  - Correlation with physical constraints r_{ij}

$$P(S_j) = P(S_j \mid X_{SID}) \cdot \sum_{i=1}^{N} r_{ij} \cdot P(S_i \mid X_{MA})$$

Multimodal fusion

Localization and identification

- After fusing audio-visual stream of data, the system gives
  - Participants' location
  - Speaker identity
  - Seating arrangement
  - Active speaker segmentation
- Testing (~85%)
  - Three 20-minute meeting (4 participants)
  - Casual conversation with interruptions and overlap

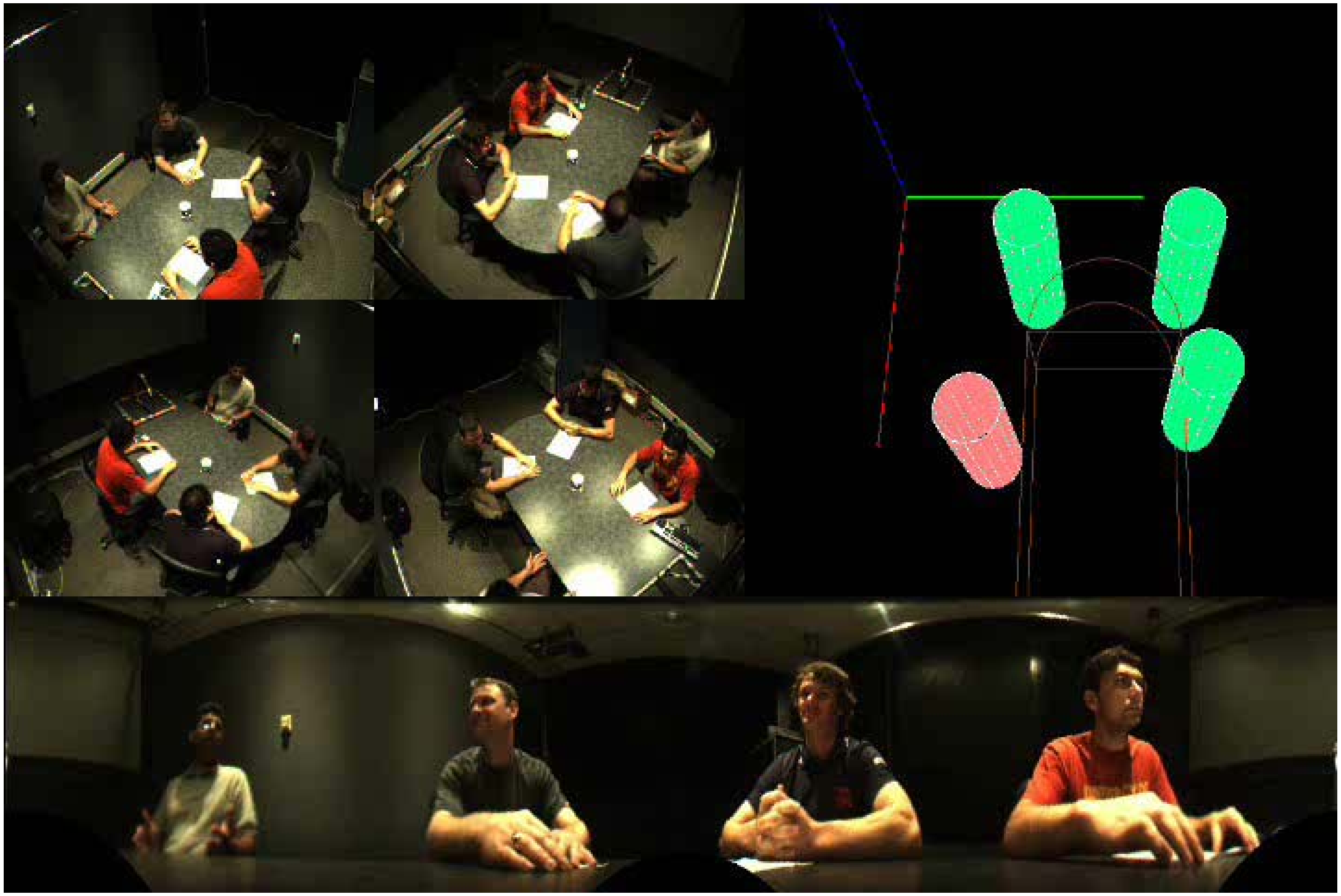


		Session	Strong Decision	Weak Decision
A	Speaker ID (GMM based)	1	66.13%	73.28%
		2	61.27%	68.51%
		3	60.10%	67.85%
B	Microphone Array + Video	1	81.26%	86.02%
		2	85.41%	92.86%
		3	83.03%	89.62%
C	Microphone Array + Video + Speaker ID (assumes known seating arrangement L)	1	81.55%	88.42%
		2	85.60%	93.56%
		3	82.49%	90.32%
D	Microphone Array + Video + Speaker ID (participant location (L) learned through data)	1	80.37%	87.34%
		2	78.77%	87.26%
		3	82.49%	90.24%
E	Seating arrangement automatically learned through data (L)	1	87.78%	
		2	74.60%	
		3	97.14%	

Multimodal fusion

USC | IMSC

- The system is been built to process data in real-time



Outline

- ✓ Introduction
- ✓ Smart room
- ✓ Multimodal fusion
- ✓ Participants' interaction
- ➢ Conclusions

Participants' interaction

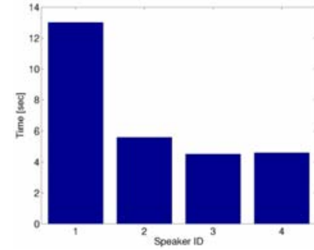
- High level features per participant
    - Number of turns
    - Average duration of turns
    - Amount of time as active speaker
    - Transition matrix depicting turn-taking between participants
- Evaluation
    - Hand-based annotation of speaker activity

SAIL

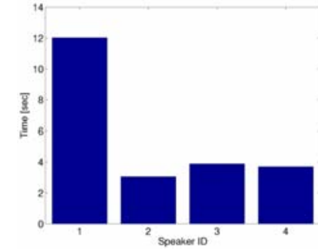
Participants’ interaction

- Automatic annotations are good approximation
- The distribution of time used as active speaker correlate dominance [Rienks,2006]
  - Subject 1 spoke more than 65% of the time
- Discussion are characterized by many short turns to show agreement (e.g. “uh-huh”) and longer turns taken by mediators [Burger,2002]
  - Subject 1 was leading discussion
  - Subject 3 was only an active listener

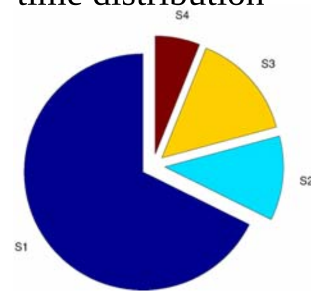
Ground-true duration



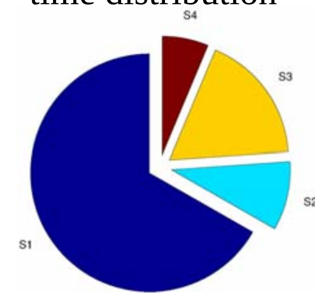
Estimated duration



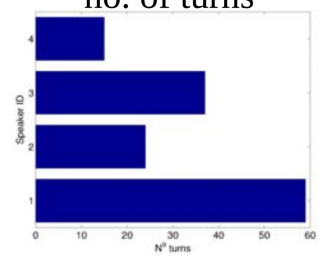
Ground-true
time distribution



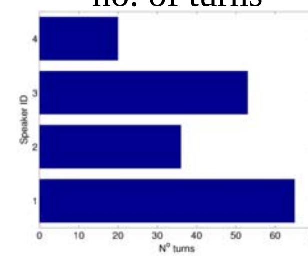
Estimated
time distribution



Ground-true
no. of turns



Estimated
no. of turns

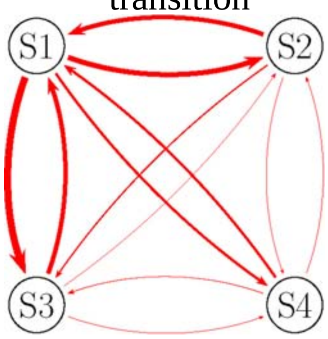


Participants' interaction

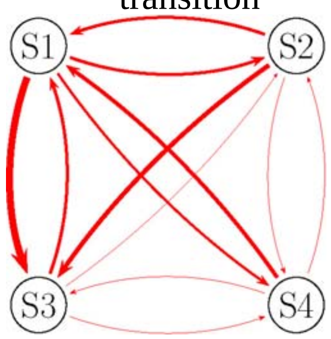
	Hand-based addressee Annotation			
	Sp1	Sp2	Sp3	Sp4
Sp1	0.00	0.31	0.44	0.25
Sp2	0.72	0.00	0.21	0.07
Sp3	0.69	0.18	0.00	0.13
Sp4	0.50	0.23	0.28	0.00

	Turn taking Transition Matrix			
	Sp1	Sp2	Sp3	Sp4
Sp1	0.03	0.34	0.46	0.17
Sp2	0.74	0.04	0.22	0.00
Sp3	0.76	0.08	0.05	0.11
Sp4	0.73	0.00	0.20	0.07

Ground-true transition



Estimated transition

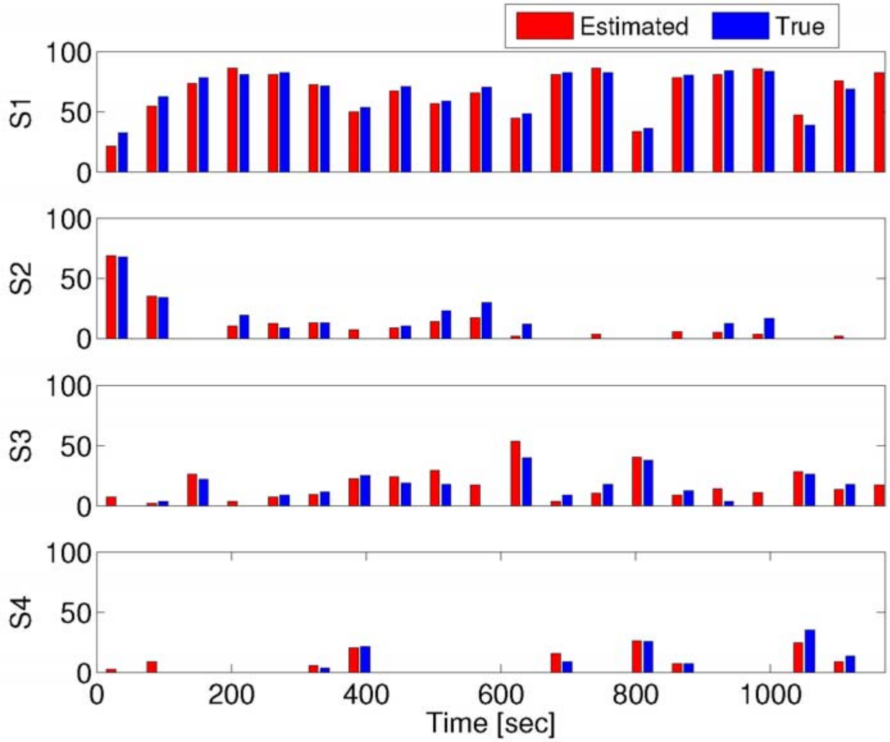


- The transition matrix gives the interaction flow and turn taking patterns
- Claim: transition between speaker ~ who was being addressed
  - To evaluate this hypothesis, addressee was manually annotated and compared with transition matrix
  - Transition matrix provides a good first approximation to identifying the interlocutor dynamics.
- Discussion was mainly between subjects 1 and 3.

Participants’ interaction

- These high-level features can be estimated in small windows over time to infer participants’ engagement
  - Subject 4 not engaged
  - Subjects 1, 2 and 3 engaged

Dynamic behavior of speakers’ activeness over time



Outline

- ✓ Introduction
- ✓ Smart room
- ✓ Multimodal fusion
- ✓ Participants’ interaction
- ✓ Conclusion

Conclusion

- Multimodal approaches to infer meta-information from speaker gives better performance than unimodal system
  - Robustness (redundant information)
  - Accuracy (complementary information)
- Participants' interaction can be estimated from automatic speaker segmentation
- Intelligent environments provide suitable platform to infer users' non-verbal messages

Conclusion

Future work

- Rough estimations of the participant gestures will be extracted
  - We propose to include this information as additional clue to measure speaker engagement
- Improve fusion algorithm
  - Particle filter based approach
- Smart room as training tool
  - Evaluate whether the report provided by the smart room can be used as training tool for improving participant skills during discussions

Smart room team

Thank you

Faculties

Shrikanth S. Narayanan
Panayiotis G. Georgiou
Ramakant Nevatia
Isaac Cohen
Scott Millward

Students

Kyu Jeong Han
Viktor Rozgic
Samuel Kim
Chi-Wei Chu
Anustup Choudhury
Tom Murray
Soonil Kwon
Sergi Hernanz
Anish Nair
Jinman Kang
Wei-Kai Liao
Sung Lee
Carlos Busso