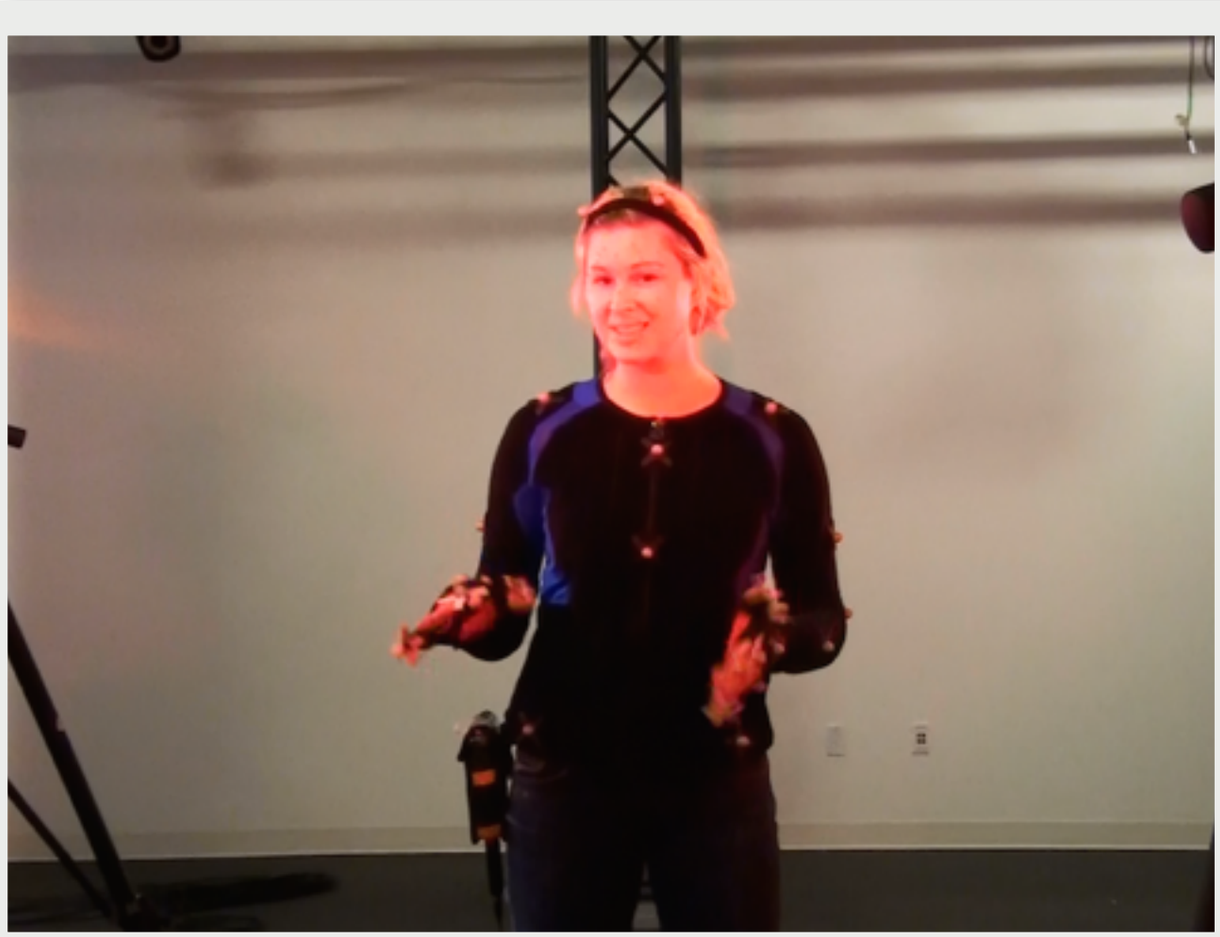
# MSP-AVATAR Corpus:

## Motion Capture Recordings to Study the Role of Discourse Functions in the Design of Intelligent Virtual Agents

NAJMEH SADOUGHI, YANG LIU AND CARLOS BUSSO



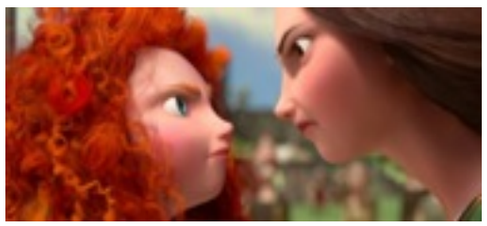
Multimodal Signal Processing (MSP) lab
The University of Texas at Dallas
Erik Jonsson School of Engineering and Computer Science

May 8th, 2015

busso@utdallas.edu

# Motivation

ICT-USC



maxresdefault.jpg

- Synthesizing human-like behaviors
  - Animation
  - Entertainment
  - Virtual reality
  - Tutoring/training systems
- Multiple verbal and nonverbal behaviors
  - Head motion, facial expressions, hand gestures, and body postures
  - Gesture coordinated, and synchronized with speech

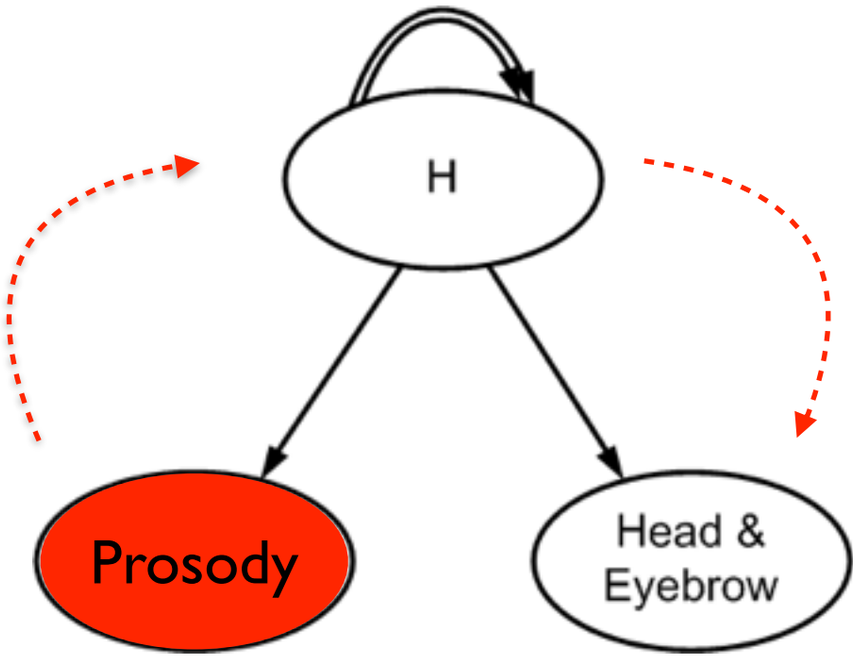
Complex relationship between gestures and speech has to be carefully considered

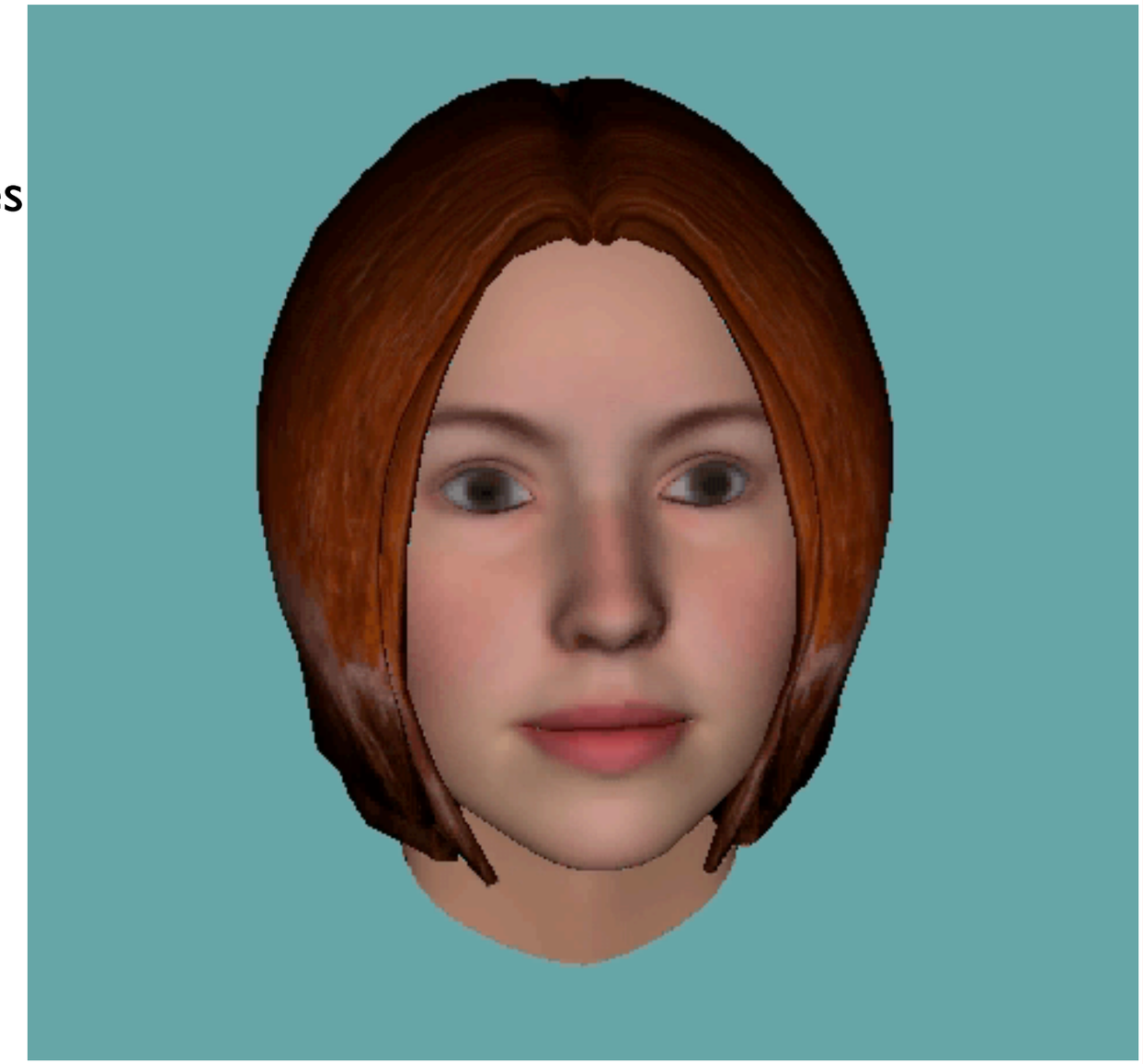
# Previous Studies on Talking Faces

- Rule-based systems [Cassell et al., 1994; S. Kopp 2006]
  - \+ Semantic meaning of behaviors (nodding)
  - \- Repetitive behaviors for a given gesture
- Data-driven methods [Levine et al., 2010; Busso et al. 2007]
  - Prosodic features are effective modalities to synthesize nonverbal human-like behaviors

# Speech-Driven Animation

- Dynamic Bayesian Models (DBNs)
  - Joint head-eyebrow discrete state variables
  - Joint continuous nodes for head and eyebrow



S. Mariooryad and C. Busso, "Generating human-like behaviors using joint, speech-driven models for conversational agents," IEEE Transactions on Audio, Speech and Language Processing, vol. 20, no. 8, pp. 2329-2340, October 2012.

# Previous Studies on Talking Faces

- Rule-based systems [Cassell et al., 1994; S. Kopp 2006]
  - \+ Semantic meaning of behaviors (nodding)
  - \- Repetitive behaviors for a given gesture
- Data-driven methods [Levine et al., 2010; Busso et al. 2007]
  - Prosodic features are effective modalities to synthesize nonverbal human-like behaviors
  - \+ Capturing more variability, and synchrony
  - \- Behaviors may not convey the semantics

# Our Vision

Rule-based systems

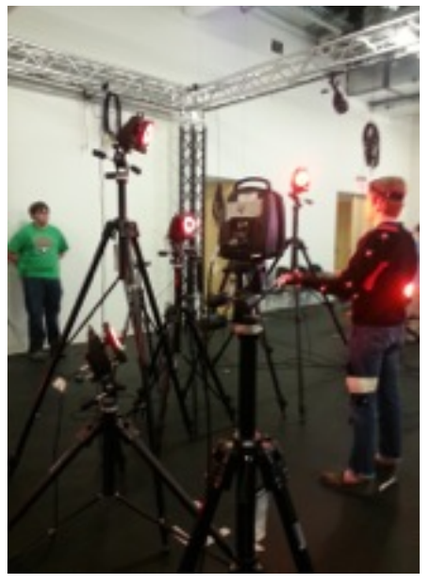
Data-driven systems

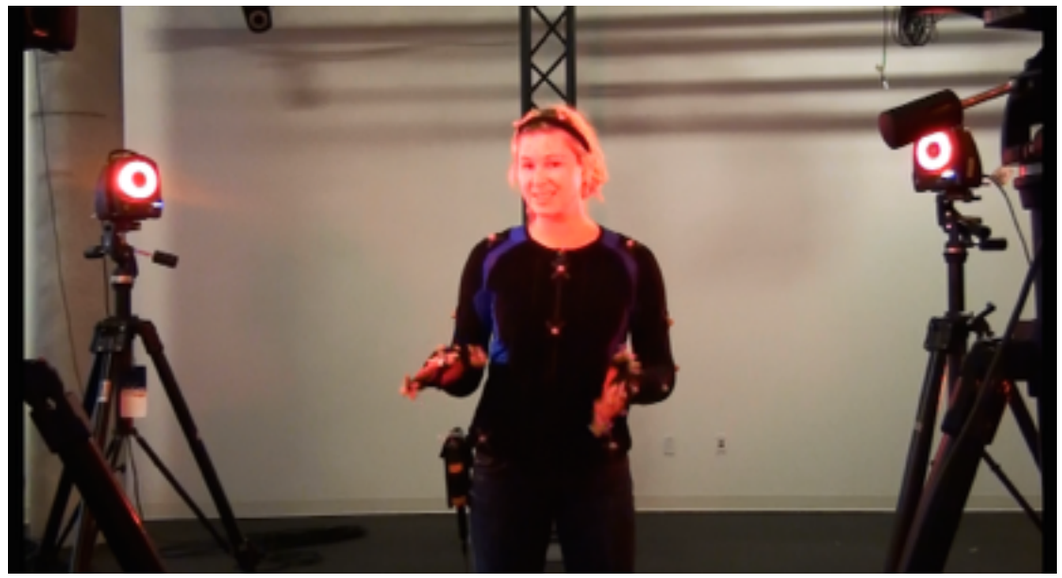
- Considering the underlying discourse function to bridge the gap between data driven and rule-based systems.
  - Synthesizing behaviors that are timely aligned with speech
  - Synthesizing behaviors that convey the right meaning

Data driven approach requires a corpus rich in discourse functions

# MSP-AVATAR Corpus

- Multimodal database comprising:
  - Motion capture data
  - Video camera
  - Speech recordings

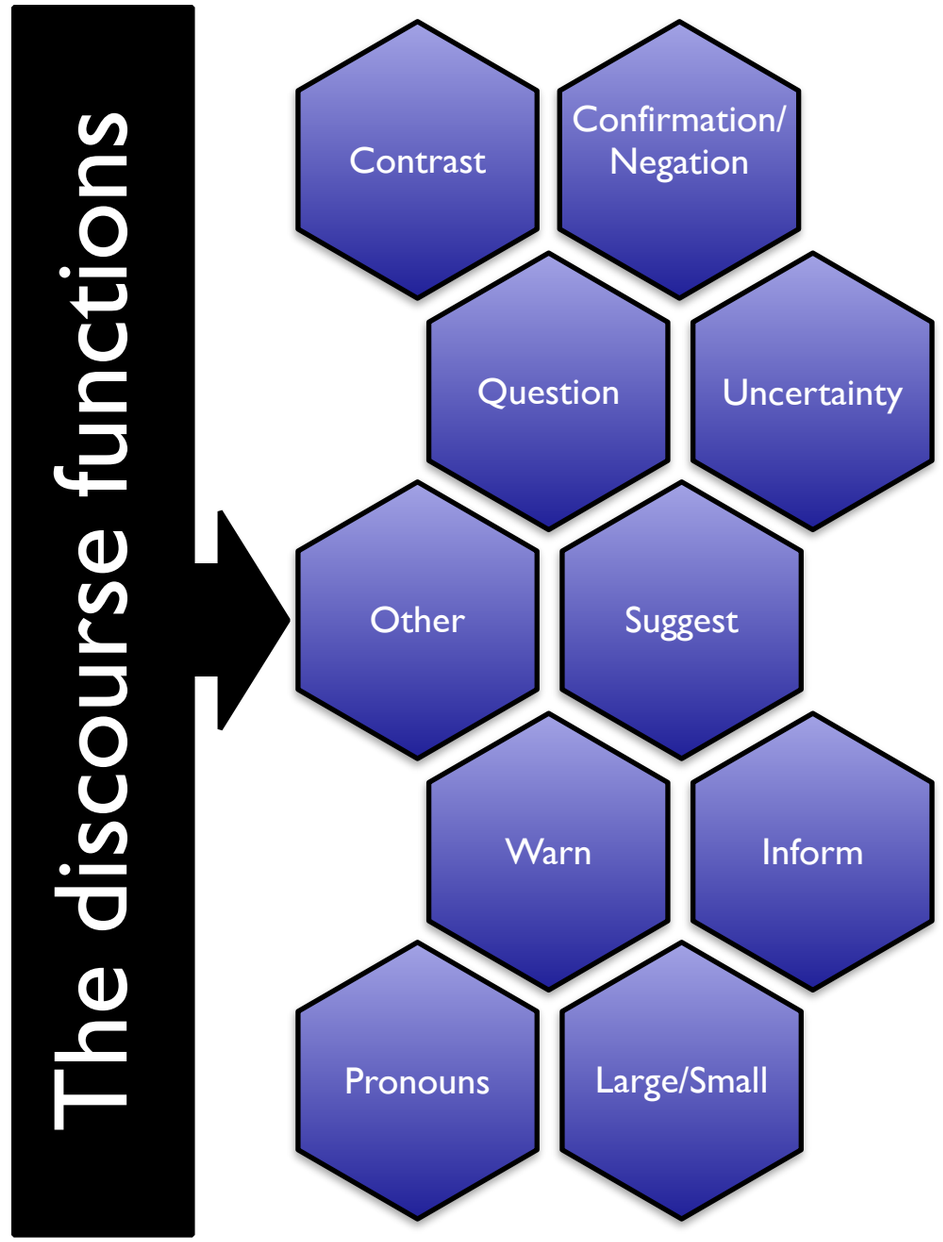




- Four dyadic interaction between actors
  - We motion captured one of the actors
- Database rich in terms of discourse functions
  - There are no corpora available to explore the role of discourse functions

# Selection of Discourse Functions

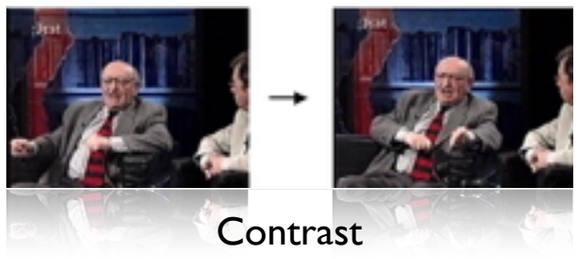
- We look for discourse functions that elicit specific gestural behaviors
- Selection guided by previous studies
  - Poggi et al [2005]
  - Marsella et al. [2013]
- 2-5 scenarios per discourse function



# Discourse Functions

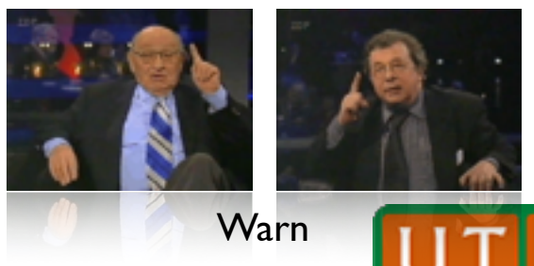
| | |
|---|---|
| CONTRAST | Contrasting two ideas, usually accompanied with contrast conjunctions such as *but, nevertheless, as, and as opposed to* |
| CONFIRMATION/ NEGATION | Showing agreement and disagreement, usually accompanied with phrases such as *Yes, No,* and *I don't think so* |
| QUESTION | Asking a question of any type: *Yes-No and Wh-questions* |
| UNCERTAINTY | Showing uncertainty in making a decision, might be accompanied by sentences such as *I really don't know what to do!* |
| SUGGEST | Suggesting ideas to the listener, e.g., *How about the new Japanese restaurant?* |
| GIVING ORDERS | Ordering any type of service, e.g. *ordering food in a restaurant* |
| WARN | Warning the listener of a danger, e.g. *Be careful about …* |
| INFORM | Inform something to the listener |
| LARGE/SMALL | The act of referring to something as small or large during speaking. These scenarios target iconic gesture usually accompany these two words or any of their synonyms |
| PRONOUNS | The act of referring to any pronoun (I/You/She/He/They). These scenarios target deictic gestures |


Pronouns


Contrast

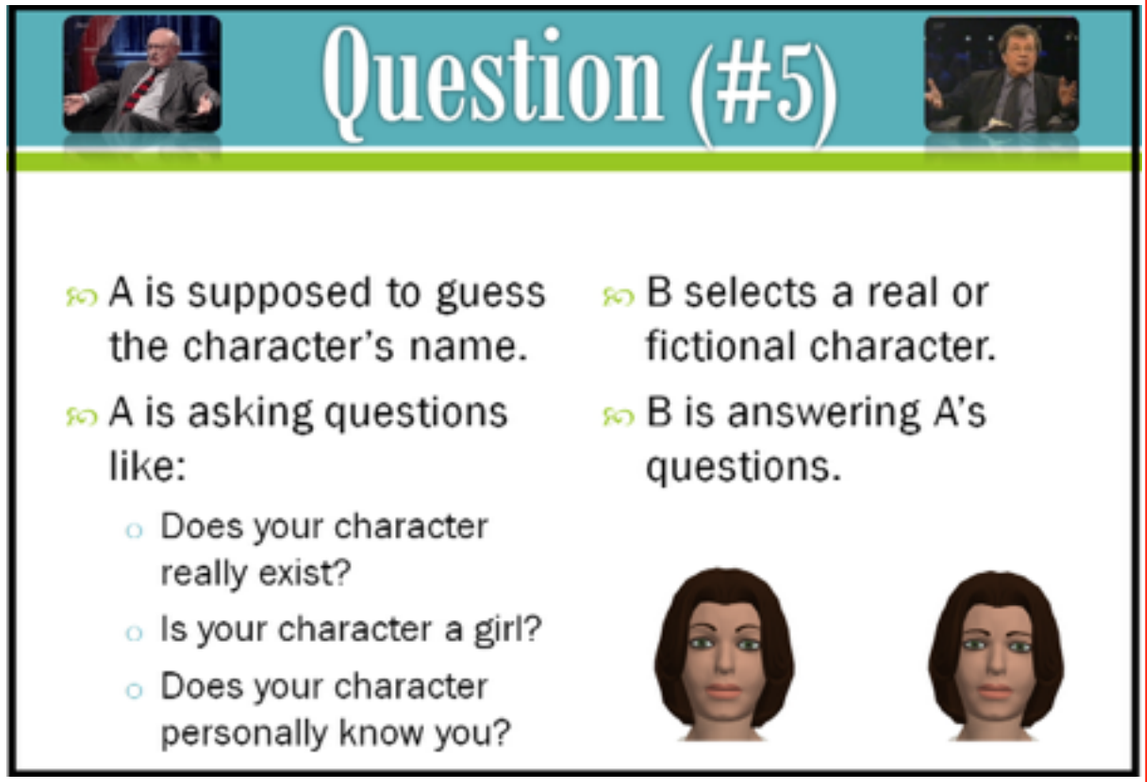

Question


Small/Big


Warn

# Scenarios

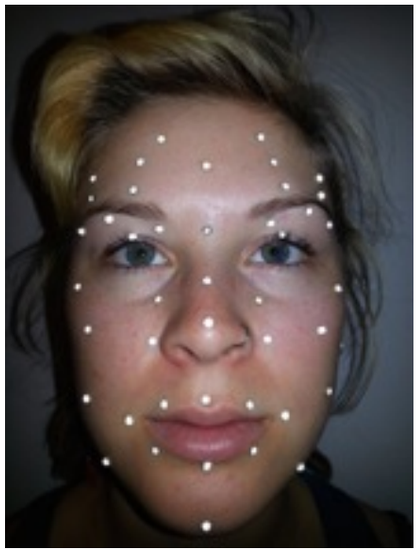
- Scenarios designed to elicit characteristic behaviors
  - Description of the scenario
  - Prototypical behaviors associated with target discourse function

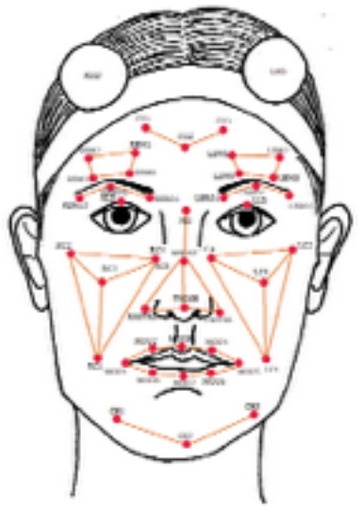


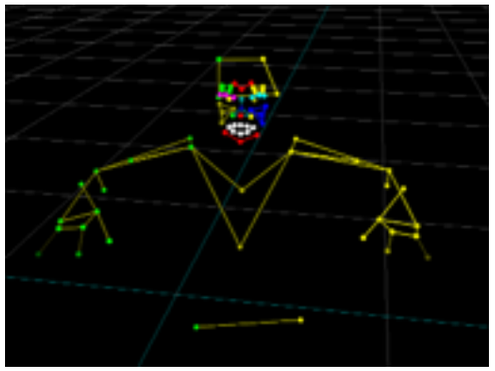

- Duration of the recorded scenarios:
  - MEAN = 143.1 sec, STD = 74.7 sec.
  - 21, 15, 22, and 16 scenarios (74 in total), based on pace of the actors

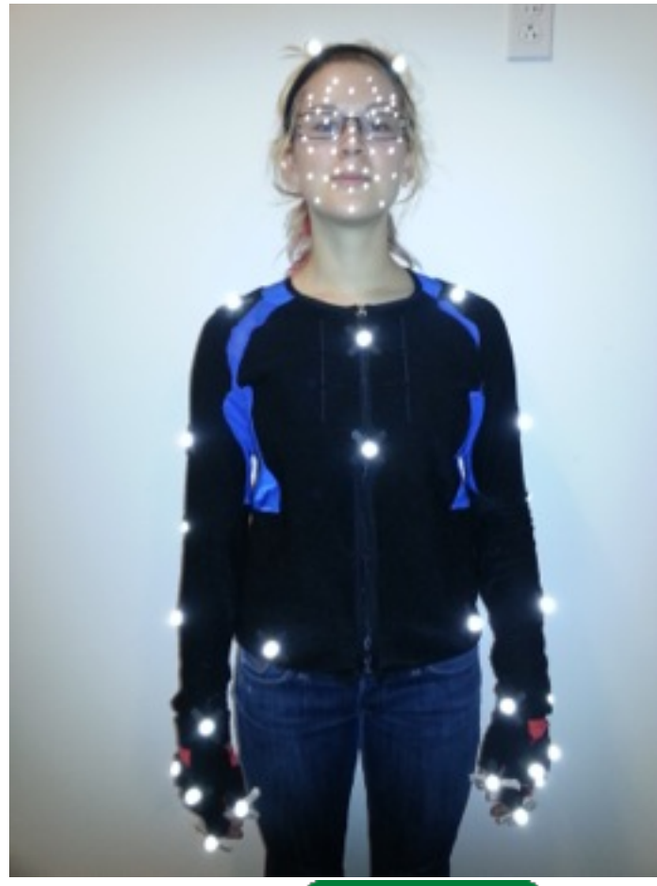
# Motion Capture

- 16 VICON cameras
- Face:
  - 43 reflective markers
- Upper-body joints:
  - Headband with 4 markers
  - Suite with 28 marker







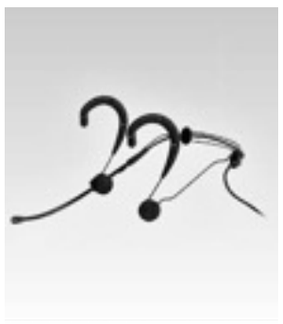


Recording "upper-body", and "facial"!

# Audio

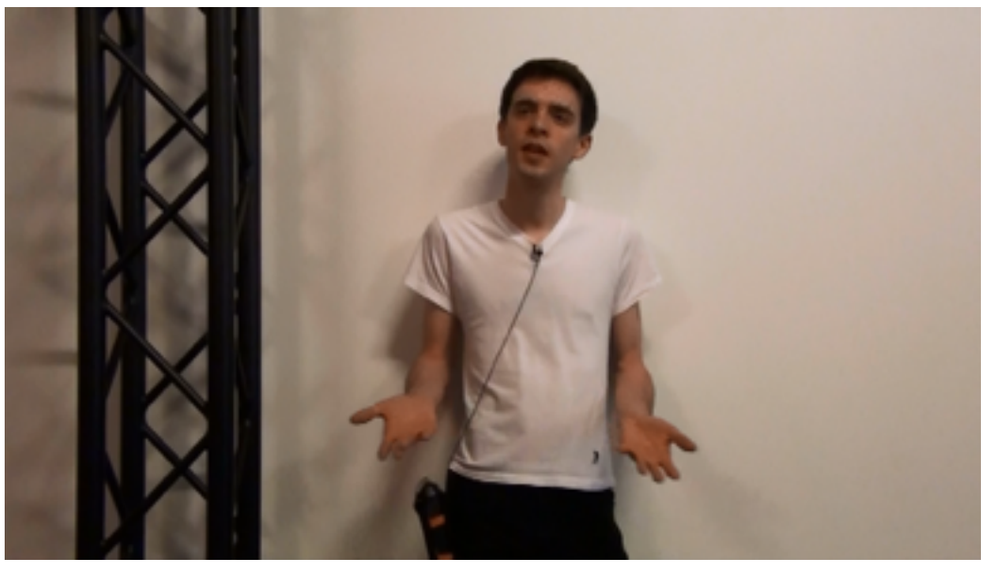
- Audio-visual data for all actors (6 people)
- Microphone connected to a digital recorder (TASCAM DR-100MKII)
  - 16 bit resolution
  - Sampling rate of 44.1 kHz

- First session: head-worn microphone (SHURE BETA 53)
  - Occluded some of the facial markers
  - Making the post-processing more difficult



- Next two session: lavalier microphone (SHURE MX150)

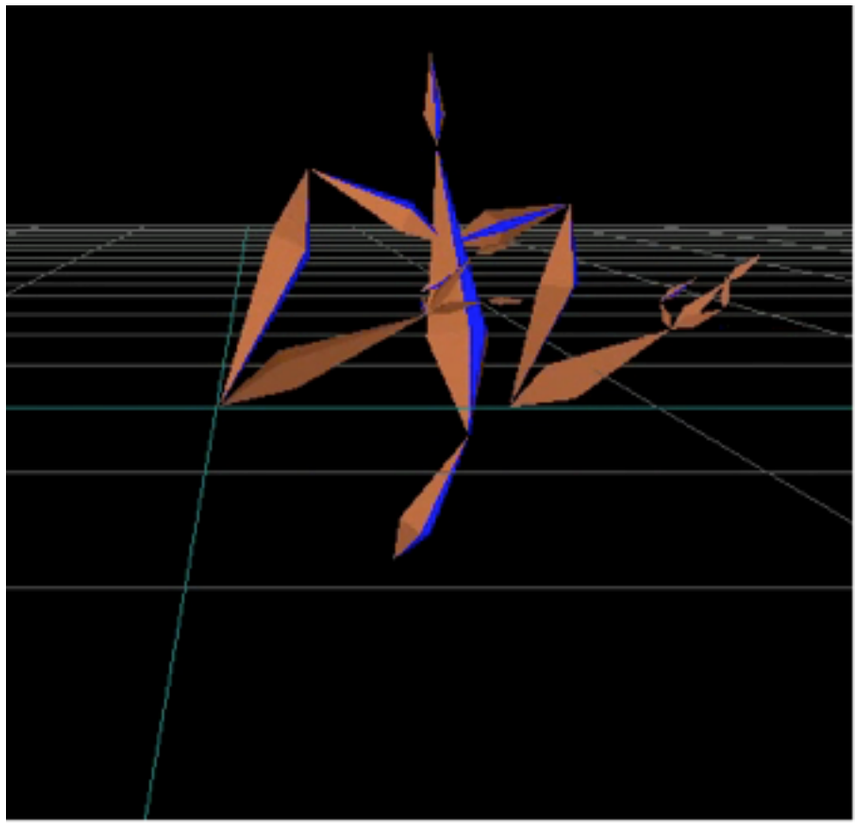
# Video

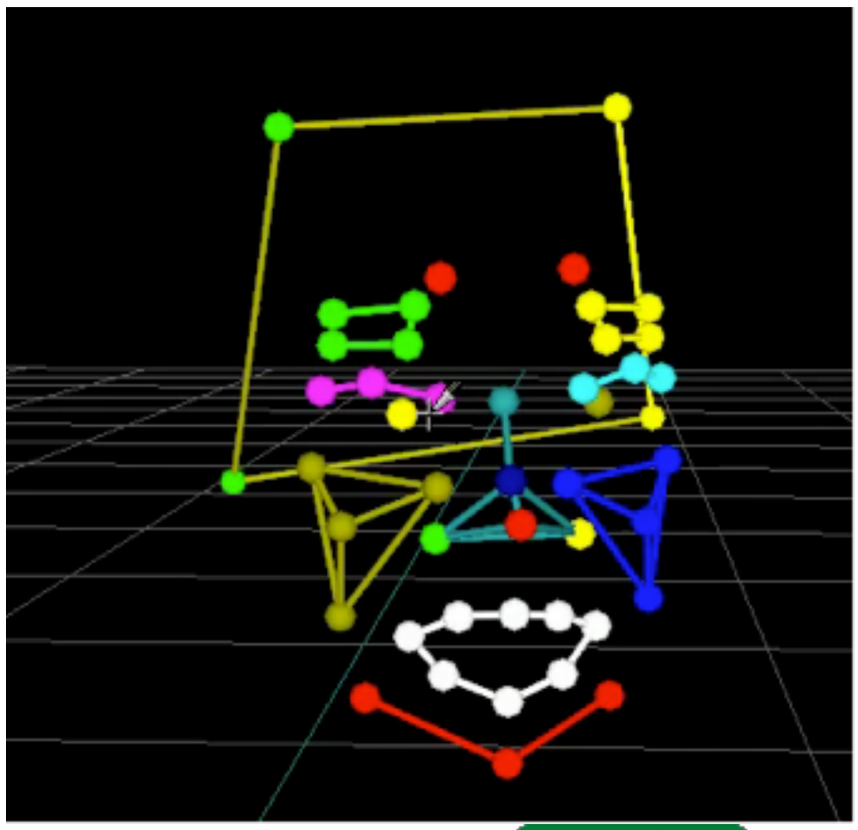


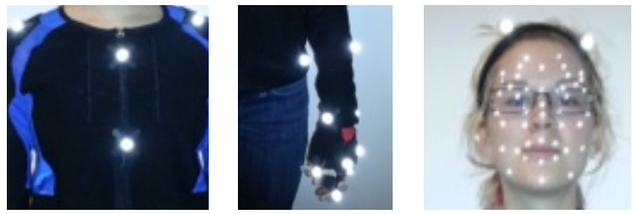
- Frontal view of both actors
- Two Sony Handycams HDR-XR100
  - 1920x1080 resolution in Full HD
- We use these videos to annotate the behaviors
- They can be used in extracting visual features from both actors
- We use a clapboard with two reflective markers to synchronize audio, video and motion capture data

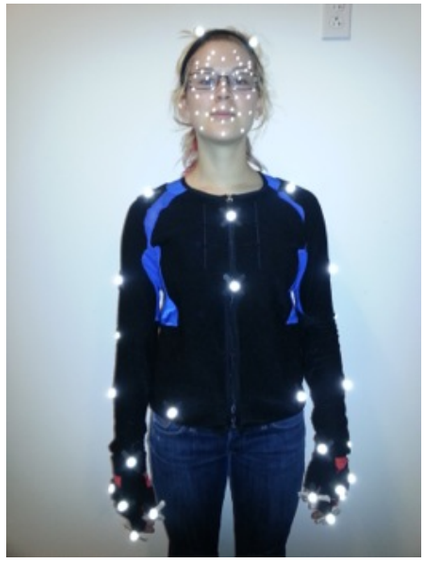
# Post Processing

- Cleaning the motion capture data
  - Blade (VICON)
- Skeleton personalized for each actor
- Upper-body skeleton
  - 74 scenarios (all sessions)
- Facial markers
  - Only 3 scenarios



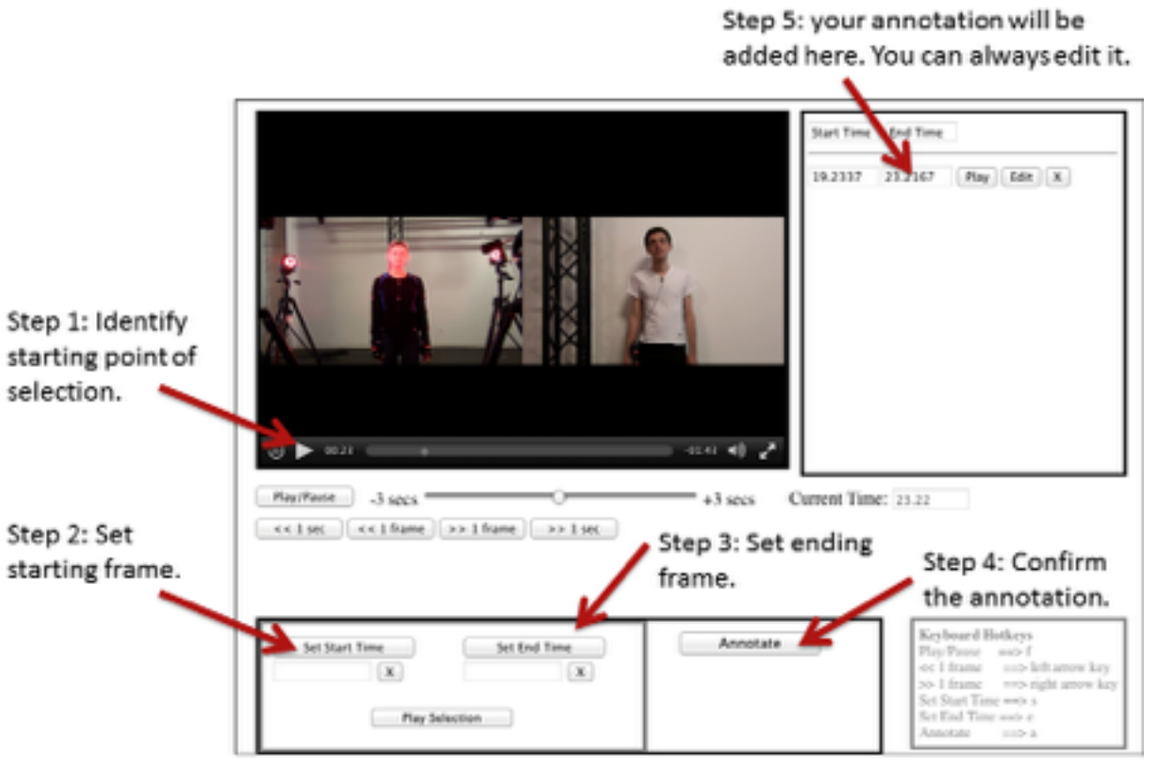






# Discourse Functions

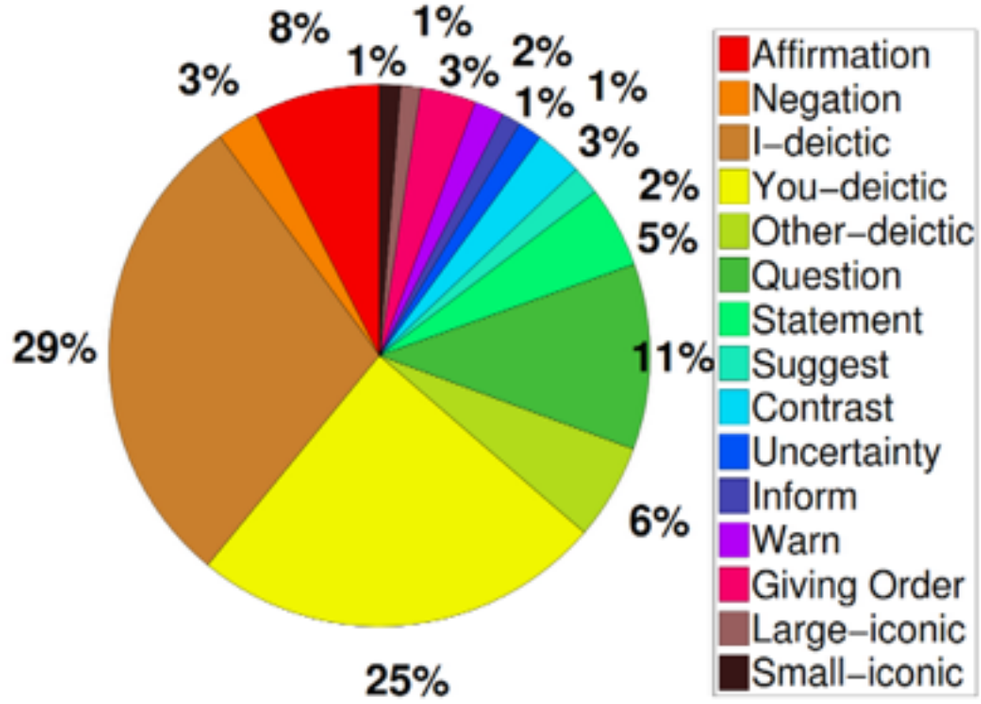
- Annotate discourse functions within speaking turns
- All annotations are done by one person, using Audacity
- We have 1751 samples
  - UPDATE: We are currently getting annotations by more people using AMT with OCTAB [Sunghyun et al., 2014]




1 . Park, Sunghyun, Philippa Shoemark, and Louis-Philippe Morency. "Toward crowdsourcing micro-level behavior annotations: the challenges of interface, training, and generalization." *Proceedings of the 19th international conference on Intelligent User Interfaces*. ACM, 2014.

# Discorse Functions in MSP-AVATAR

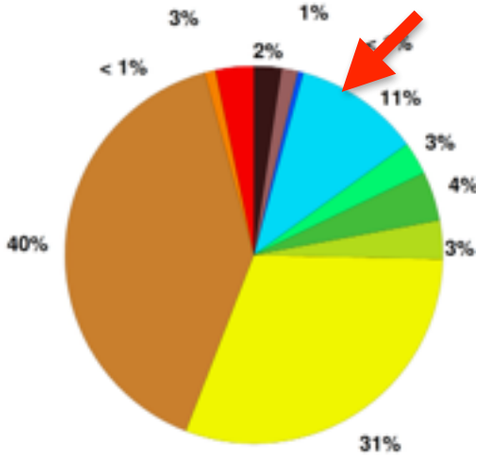
- Some discourse functions appears in most scenarios
  - I-deictic
  - you-deictic
  - questions



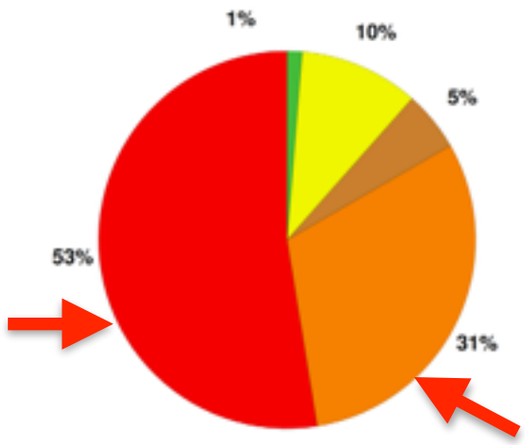

- We have multiple instances of the most relevant discourse functions

# Discorse Functions in MSP-AVATAR

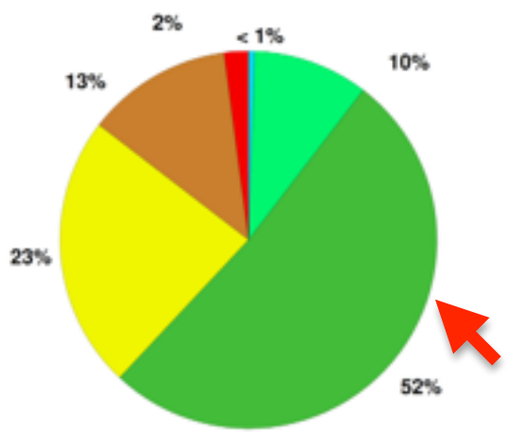
- The scenarios successfully elicited the target discourse functions



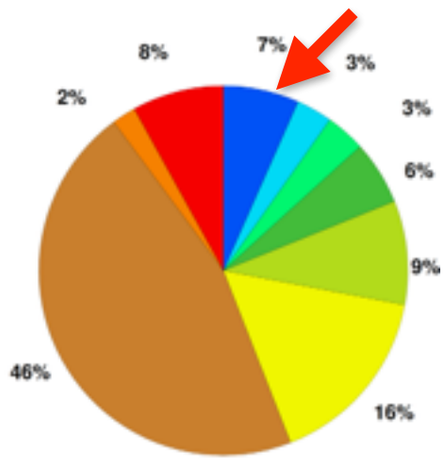

(a) Contrast



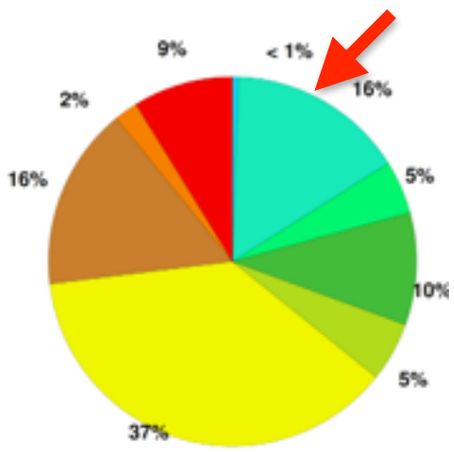

(b) Affirmation/Negation



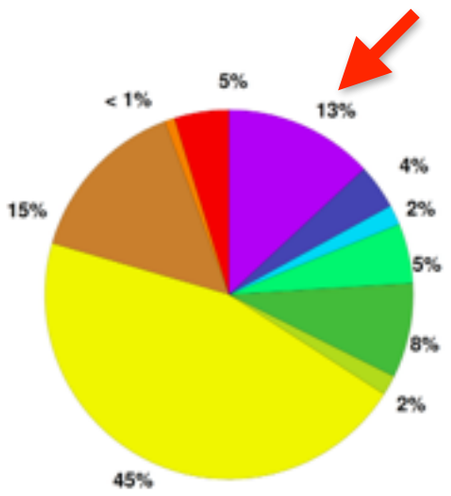

(c) Question




(d) Uncertainty



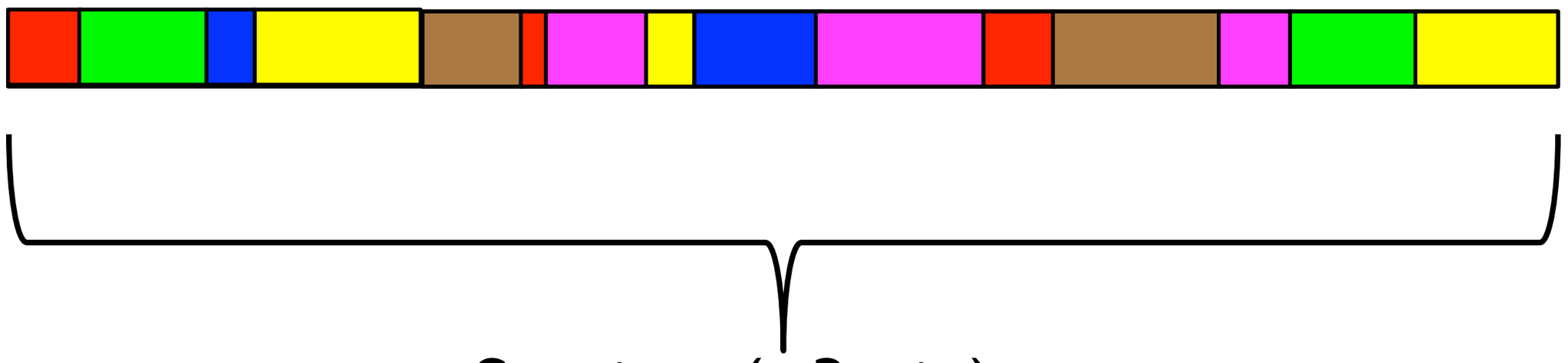

(e) Suggest




(f) Warn

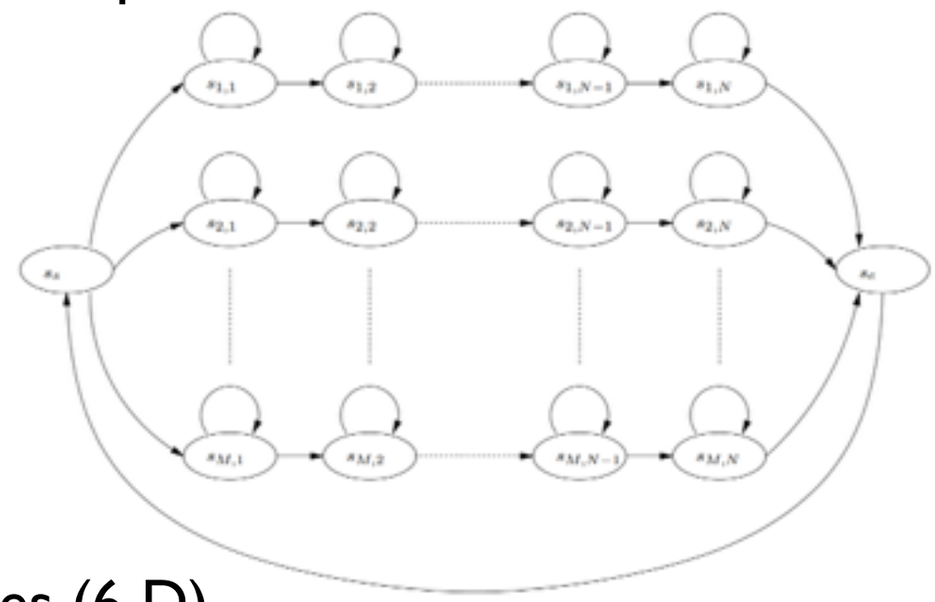
# Analysis of body movements

- Are the gestures different across discourse functions?
- We focus on hand and head gestures
  - We only have facial features of 3 sessions
- We rely on automatic unsupervised segmentation of gestures
  - Data-driven mid-level classes



# Parallel Hidden Markov Models

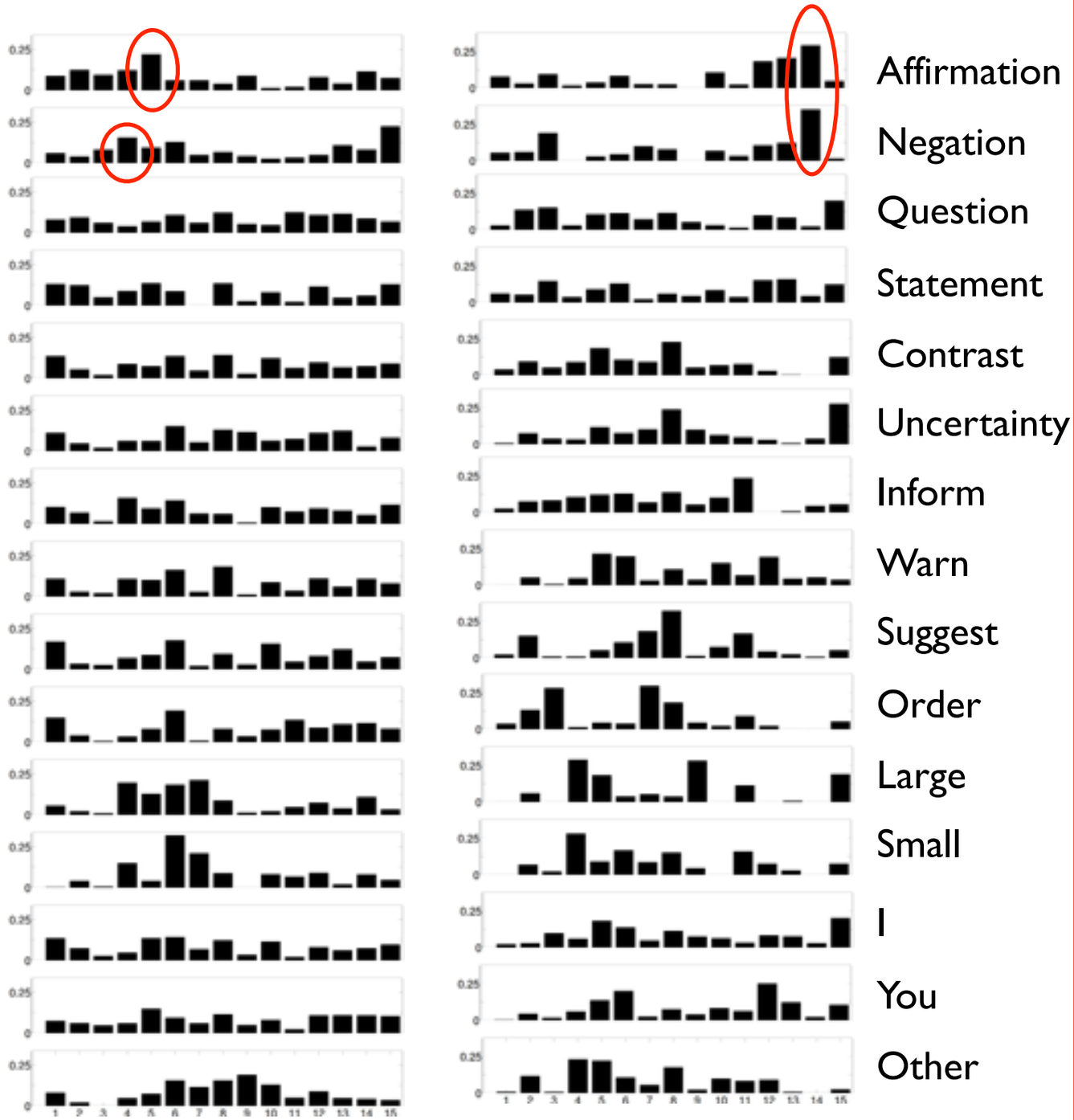
- Cluster the behaviors using PHMM
  - Individual left-to-right HMMs
  - States within the branches have self transitions
  - Model dynamic behaviors having different temporal durations
- #states per branch = 15
  - Minimum duration: 125 ms
- 15 branches for both head and hand
- Analysis on head, and arms movements
  - Head: three angular rotations + derivatives (6 D)
  - Arms: angular rotations + derivatives (20D)



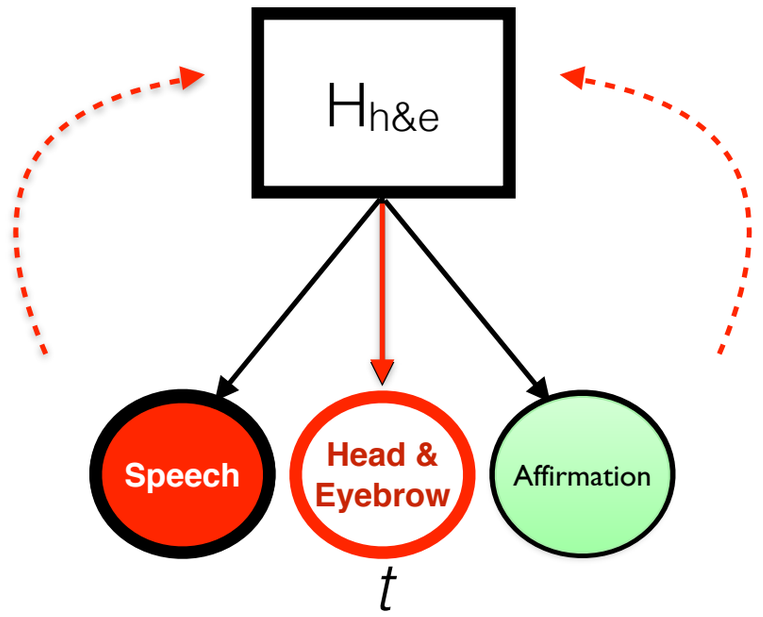
M. E. Sargin, Y. Yemez, E. Erzin, and A. M. Tekalp. Analysis of head gesture and prosody patterns for prosody-driven head-gesture animation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 30(8):1330–1345, 2008.

# Analysis of Head Motion and Hand Gestures

- Different gestures across discourse functions
- Hand: *affirmation* and *negation* have a distribution with a peak in cluster 14
  - Hand gesture less active
- Head: peak in *affirmation* relate to head nod (cluster 5)
- Head: *negation* include head shakes (cluster 4)



# Constraining DBNs with Discourse Functions




Constrained on Affirmation

# Conclusions

- This paper introduces the MSP-AVATAR database
- This corpus comprises video and audio of 6 actors and motion capture recording of 4 actors in dyadic conversations
- Scenarios are designed such that they elicit the behaviors associated with discourse functions
- One drawback of this corpus is small number of subjects that were motion captured
- We expect to release the database after finishing the correcting process of motion capture recordings

http://ecs.utdallas.edu/research/researchlabs/msp-lab/MSP-AVATAR.html

# Multimodal Signal Processing (MSP)

- Questions?


**http://msp.utdallas.edu/**