Speech-Driven Animation Constrained by Appropriate Discourse Functions

Najmeh Sadoughi[1], Yang Liu[2] and Carlos Busso[1]

1. Multimodal Signal Processing (MSP) Laboratory
2. Human Language Technology Research Institute

Erik Jonsson School of Engineering & Computer Science
University of Texas at Dallas

ICMI 2014
The 16th International Conference on
Multimodal Interaction
November 12-16, 2014 / Istanbul / Turkey

MOTIVATION

Background:

- Rule-based:
  + defining rules for behaviors based on the contextual information
  - repetitive behaviors
  - desynchronization between gestures and speech
- Speech-driven:
  + use of prosodic features to model behaviors
  + modeling emphasis, emotion, and timing of behaviors
  - may not properly respond to the underlying discourse functions in the dialog

Proposed Solution:

- Create a bridge to fill the gap between speech-driven and rule-based systems

METHOD

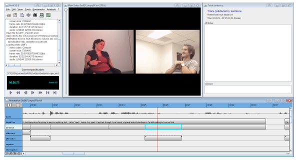
IEMOCAP corpus

- Dyadic interactions
- 1st session (1 male, 1 female)
- Motion capture data (head, and eyebrow motions)
- Audio: F0 contour, and Intensity

Statistical Analysis (MEAN)		
Question vs. Non-Question		
Pitch	F(1,452)=8.58	p=0.004
Roll	F(1,452)=7.05	p=0.008
Pitch Velocity	F(1,452)=7.05	p=0.008
Affirmation vs. Non-Affirmation		
LBRO3	F(1,464)=7.87	p=0.005
RBRO3	F(1,464)=10.42	p=0.001
Pitch Velocity	F(1,464)=6.74	p=0.0097
Negation vs. Non-Negation		
Yaw	F(1,419)=5.17	p=0.023
Pitch Velocity	F(1,419)=4.99	p=0.026
Statement vs. Non-Statement		
Pitch Velocity	F(1,470)=4.30	p=0.038

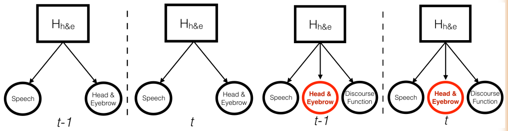
Annotation

- Selection of discourse function is inspired by previous studies [Poggi et al.,2005; Marsella et al., 2013]
- Discourse functions:
  - affirmation (90)
  - negation (53)
  - question (112)
  - statement (158)



Speech Driven Models Using DBN

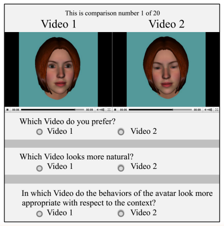
- Xface toolkit (compliant with MPEG-4 standard)
- Speech: prosody features
- Head & Eyebrow: Joint configuration of Head and Eyebrow [Mariooryad et al., 2013]
- Discourse function: A binary variable representing the discourse function
  - Training: full observation
  - Testing: partial observation



RESULTS

Subjective Evaluation (MTurk)

- Focus on question and affirmation
- Original, jDBN3, C-jDBN3
- 20 different videos
- Pairwise comparison (60)
- 3 evaluators per comparison



This is comparison number 1 of 20

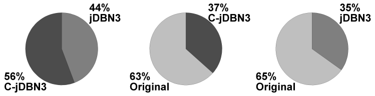
Video 1 Video 2

Which Video do you prefer?
○ Video 1 ○ Video 2

Which Video looks more natural?
○ Video 1 ○ Video 2

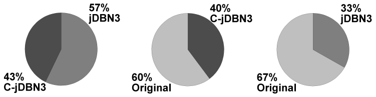
In which Video do the behaviors of the avatar look more appropriate with respect to the context?
○ Video 1 ○ Video 2

Constraint is "Question"



Which video do you prefer?

Constraint is "Affirmation"



Which video do you prefer?

"Question"

- 56% preferred C-jDBN3 over jDBN3
  - 95.5% probability that this proportion is greater than chance
- Similar results for other questions

"Affirmation"

- Direct comparison
  - 57% preferred jDBN3 over C-jDBN3
- Indirect comparison
  - C-jDBN3 closer to original videos
- Similar results for other questions

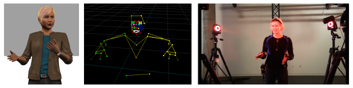
DISCUSSION

Conclusions:

- The statistical analysis demonstrated significant changes in behaviors across different discourse functions
- For "Question" we see more preference for CjDBN3, while for "Affirmation" the results are not conclusive
- Perception of head motion dominate the evaluation
  - "Affirmation" constraint is less effective since affects eyebrow

Future Work:

- We need more data to further explore this research direction
- Better talking heads



References:

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, 20(8): 2329-2340, October 2012.

This work was funded by NSF (IIS 1352950)