A Preliminary Exploration of *Group Social Engagement Level Recognition in Multiparty Casual Conversation*

YUYUN HUANG, NICK CAMPBELL

Motivation

Recently, much attention is being paid to the concept of socially-intelligent human-robot interaction, which aims to enable social robots or agents to interact naturally with humans. Specifically, a robust engagement model and automatic engagement recognition system is needed to evaluate the success of social and task-based communication.

Research Question

In this work we report on studies we have carried out on the novel research topic about social group engagement in non-task oriented (casual) multiparty conversations.

- To test that a reliable multimodal way can be used to predict the social engagement level in multiparty scenarios.

What is Social Engagement?

The engagement is defined as: *the process by which two (or more) participants establish, maintain and end their perceived connection. This process includes: initial contact, negotiating a collaboration, checking that other is still taking part in the interaction, evaluating whether to stay involved, and deciding when to end the connection.* （Sidner 2005）

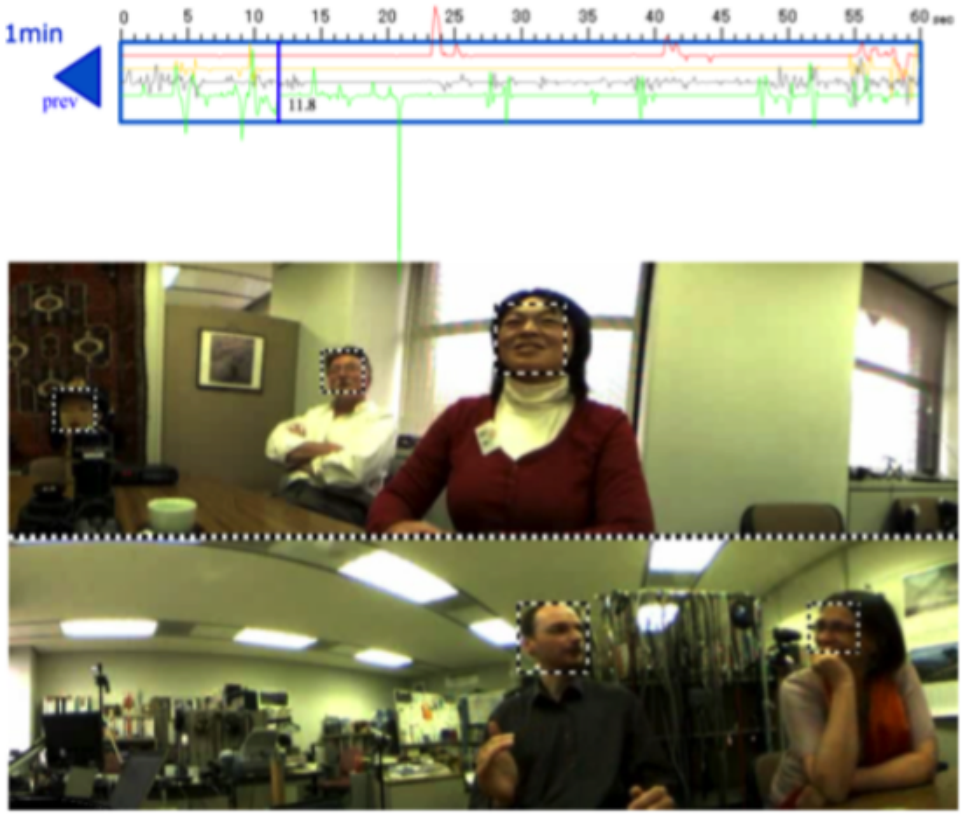
Multiparty Casual Conversation Corpus: TableTalk

Usage: Day 1 + Day 4

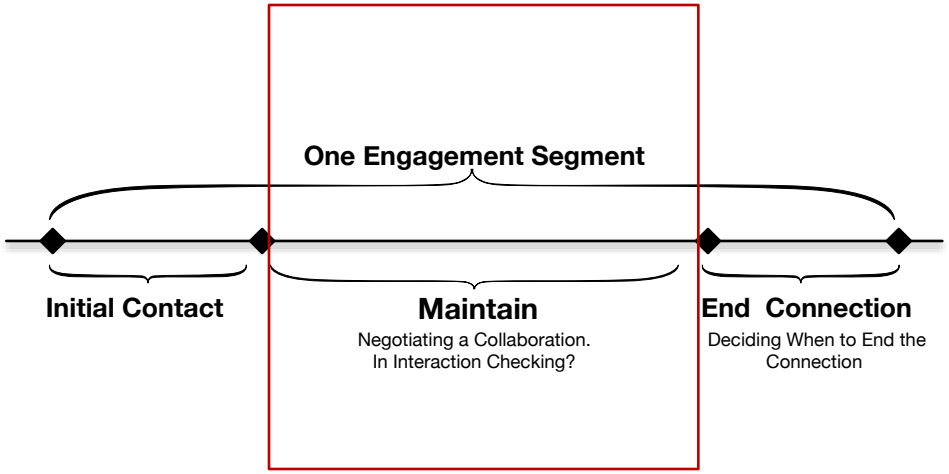
4 speakers

Length: 281 minutes

Topics change: Day 4: *71*

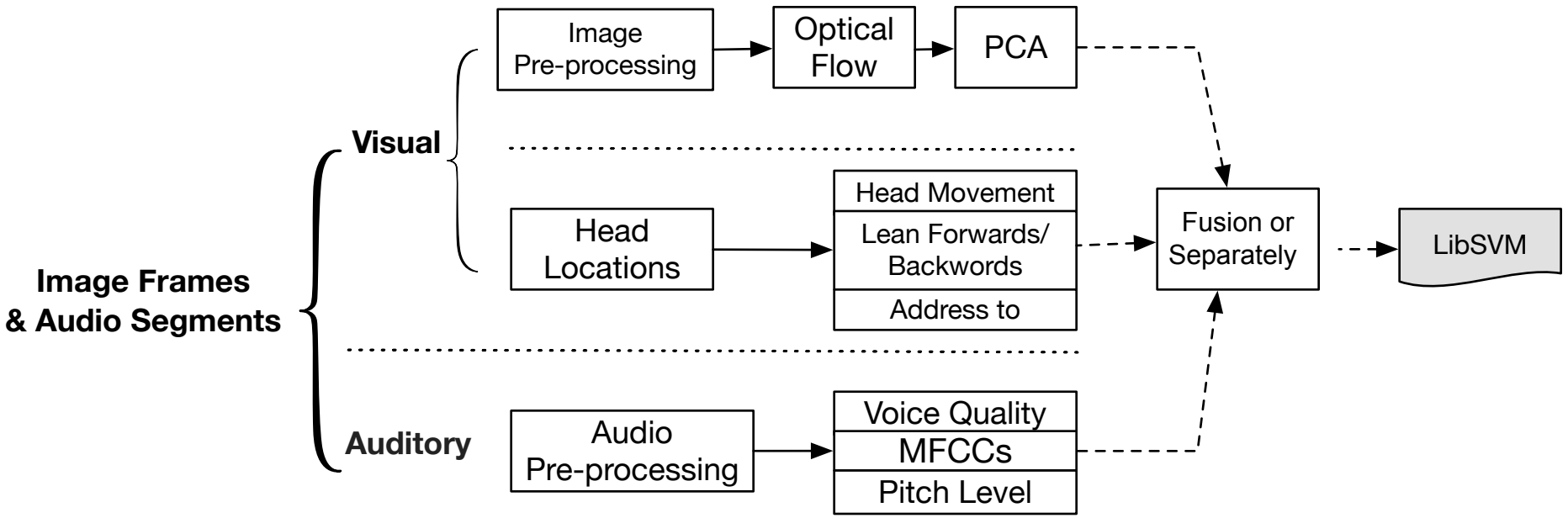


Engagement Level Annotation



5-level Engagement Annotation			
End of the previous segment			
Engagement Initialization			
Maintain	0. Strong Engaged	A. Engaged	Very engaged and strongly want to maintain the conversation
	1. Engaged		Interest but not very high, e.g willing to talking with no passion
	2. Neutral	B. Neutral	Neither show interest or lack of interest
	3. Disengaged	C. Disengaged	Less interest in the conversation
	4. Strong Disengaged		No interest to continue the conversation at all, want to leave the conversation
End Connection			

Multimodal Approach



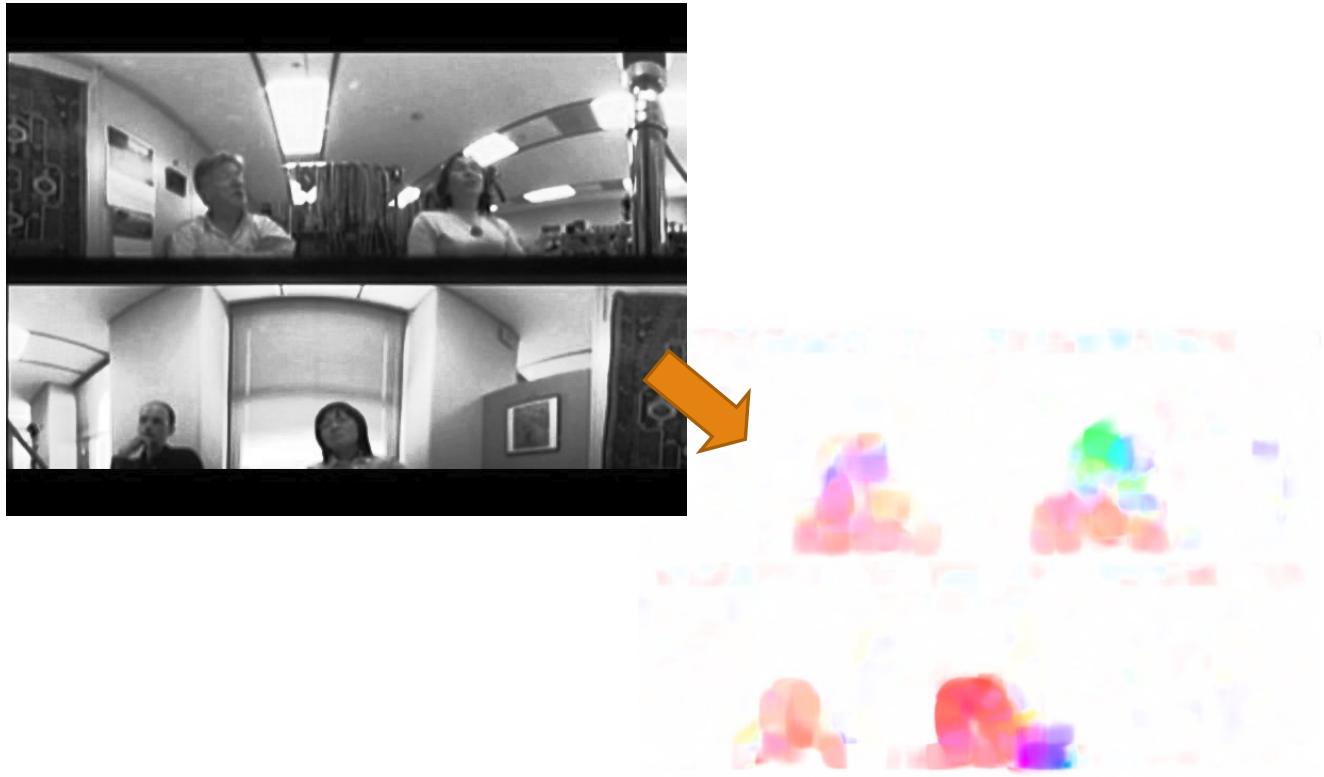
Head Related Features

The face detection and the yaw head position data were detected, Camshift tracking was also used tracking the detected faces.

Yaw angle range from −90 degrees to +90 degrees.

Backward or forward body movement (leaning) was computed by comparing the size of participants' faces across sequential frames in 10-frame steps on 30fps videos.

Optical flow with Principal Component Analysis

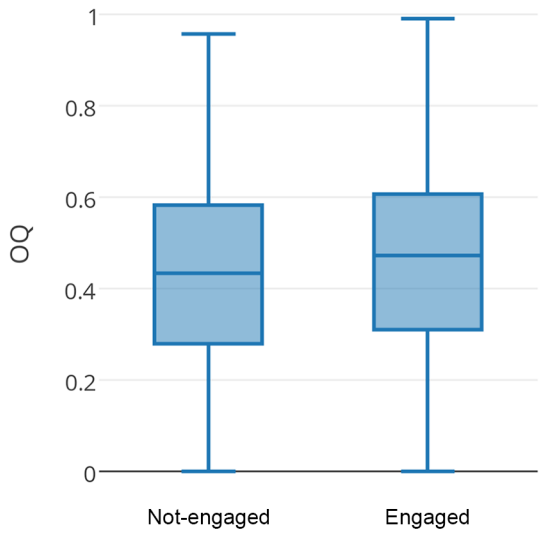


Optical flow is used to compute the motion of the pixels of an image sequence. It provides a dense (point to point) pixel correspondence over the entire scene, and thus provides an indication of how much movement is occurring overall. We used the algorithm proposed by Gunnar Farnback based on polynomial expansion, which provides all the motion of all the pixels between previous and current frames. PCA was also used for dimensionality reduction.

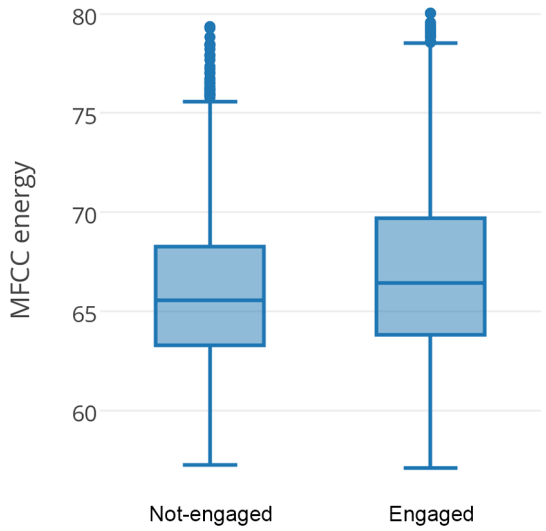
Auditory Cues

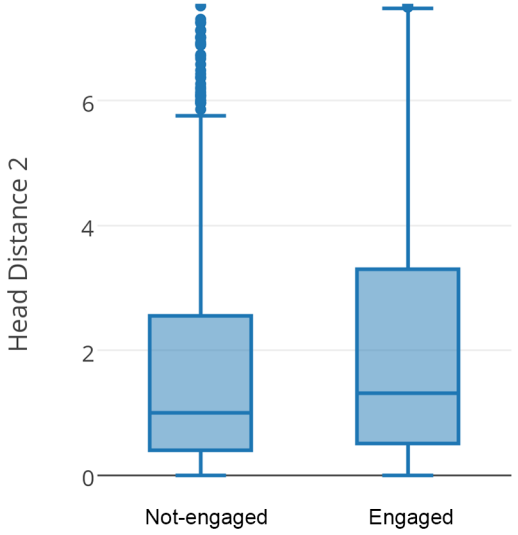
Audio recordings were down-sampled to 16 kHz for feature extraction in this work. The features extracted from the audio signal comprised ***pitch level, 12 MFCCs, MFCC energy, and glottal parameters***.

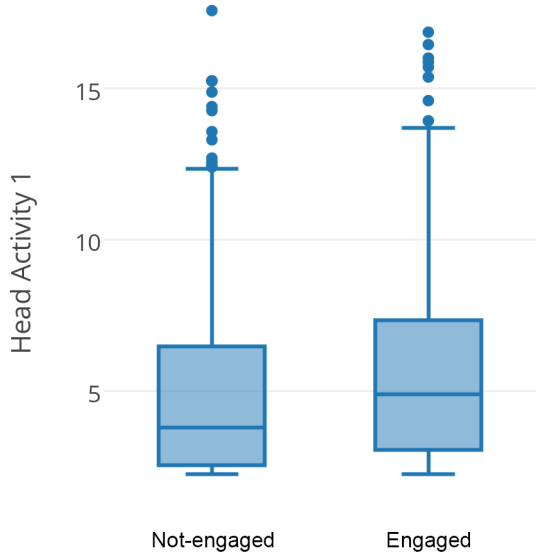
Feature Analysis-Box plots



Open Quotient (OQ)







Prediction Results – Visual

Feature Set	Accuracy
Head forward/backward	65.0024%
Head move distance	71.0875%
Head Pose	71.2081%
Optical Flow with PCA	73.4748%
Head Move Distance + Optical Flow + Head Yaw Angle	74.1741%

Prediction Results – Auditory

Feature Set	Accuracy
F0	60.9187%
Glottal	71.9074%
MFCCs	72.2691%
Glottal + MFCCs	72.679%

Prediction Results – Feature level fusion

Feature Set	Accuracy	Recall	Precision	F-Score
Auditory and Visual combined	82.23%	0.822	0.816	0.815

Conclusion

Low level visual and auditory cues of engagement have been analyzed in the TableTalk corpus. In general, the visual parameters performed slightly better than the auditory parameters in recognition of engagement in this work. We compared recognition results using feature fusion and using visual/audio features separately, and found that audio-visual fusion gave higher accuracy. As a shallow analysis, we believe that advanced detailed visual and audio features can definitely increase the prediction accuracy.

Future Work

1. Deep learning approaches
2. Other level fusion methods such as decision & modal level