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

Yuyun Huang[1(✉)], Emer Gilmartin[1], Benjamin R. Cowan[1,2], and Nick Campbell[1]

[1] Speech Communication Laboratory, School of Computer Science and Statistics, Trinity College Dublin, Dublin, Ireland
{huangyu,gilmare,nick}@tcd.ie
[2] SILS, University College Dublin, Dublin, Ireland
benjamin.cowan@ucd.ie

Abstract. Sensing human social engagement in dyadic or multiparty conversation is key to the design of decision strategies in conversational dialogue agents to decide suitable strategies in various human machine interaction scenarios. In this paper we report on studies we have carried out on the novel research topic about social group engagement in non-task oriented (casual) multiparty conversations. Fusion of hand-crafted acoustic and visual cues was used to predict social group engagement levels and was found to achieve higher results than using audio and visual cues separately.

Keywords: Acoustic and visual signal processing · Human social behaviours · Social engagement recognition

1 Introduction

Although engagement can be expressed through the voice and body gestures of interlocutors and easily perceived by human beings, machines have no ability to sense such human social cognitive behaviours. Levels of engagement are also referential parameters that can be used for conversation assessment and topic detection. In this paper, we describe engagement concepts and highlight relevant works in both group and dyadic conversational engagement. We then outline our proposed engagement recognition methodology and report on several evaluations based on a multiparty casual conversation corpus.

The most widely used definition of social engagement in human - human or human - machine conversation is that formulated by Sidner 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* [20]. In measuring engagement it is also vital to take account of auditory and visual non-verbal

A. Ronzhin et al. (Eds.): SPECOM 2016, LNAI 9811, pp. 75–83, 2016.
DOI: 10.1007/978-3-319-43958-7_8

cues, as they have been reported to contain much of the affective information transferred during conversations [9].

2 Related Works

There has been much valuable research into social engagement in various conversation scenarios. Many perceptible non-verbal cues have been analysed in social conversations. Eye gaze has been widely studied in terms of social engagement or interest during dialogues. Argyle and Cook (1976) [3] noted that the failure to attend other's gaze contact was evidence of having no interests and attention. Cassell et al. (1999) [7] examined the relationship between information structure and gaze behavior. They suggested that interlocutors' gaze behaviour served to integrate turntaking cues with the information structure of the propositional content of an utterance. They found that the beginnings of themes were frequently accompanied by a look-away from the hearer, while speakers frequently looked towards the hearer at theme endings. Rich et al. (2010) [19] built a computational model to recognize engagement by using manually annotated data on mutual facial gaze, directed gaze, adjacency pairs, and back-channels. Nakano and Ishii (2010) also used eye-gaze behaviours to estimate user engagement between human users and virtual agents [16].

Gustafson and Neiberg (2010) demonstrated that prosodic cues in Swedish, including change in syllabicity, pitch slope and loudness in non-lexical response tokens, could be used to detect engagement, and investigated prosodic alignment as a cue to engagement between speaker and listener [12]. Gupta et al. (2012) used speech cues to analyse childrens' engagement behavior, with results showing that vocal cues were informative in detecting children's engagement. [11] Hsiao et al. (2012) also investigated engagement level estimation using higher level speech cues like turntaking extracted from low level cues such as MFCCs and intensity [13].

Oertel et al. (2011) [18] used multimodal cues to predict the degree of group involvement during spontaneous conversation, extracting acoustic features including pitch level and intensity and visual features including eye blinking and mutual gaze from manually annotated data. The resulting automatic prediction was based on Support Vector Machines (SVM) with three classes of involvement. Oertel and Salvi (2013) [17] modelled individual engagement and group involvement in an eight-party dialogue corpus. Their results showed that engagement and involvement can be modelled by using gaze pattern. In order to describe engagement, they introduced presence, entropy, symmetry and MaxGaze features to summarize different eye-gaze pattern aspects. Their group involvement classification using Gaussian Mixture Models got accuracies of 71.0 % on training sets and 71.3 % on test sets. Lai et al. (2013) used turn-taking features to detect group involvement and used the involvement cues to predict extractive summary content in meeting segments; they concluded that automatically derived measures of group level involvement, like participation equality and turn-taking freedom, could help in identifying relevant meeting segments for summarization [15].

Bohus et al. (2009) introduced an approach to detect human participants' engagement intentions during dialogue with an avatar agent [4]. Yu et al. (2015) built an engagement awareness dialogue system named TickTock [21], which has a social engagement model to offer information to dialogue manager where conversational strategy was decided.

In this work, we focus more on studying group engagement level recognition, considering the group as a whole rather than individuals. We investigated features which can take all interlocutors into account and contribute to the whole conversation. Visual and acoustic cues like group head movement distance, optical flow, direction of head address (yaw), leaning forward or backward, voice quality and intensity were used for the recognition task.

3 Methodology

We propose a set of features which can represent group talking traits. These comprise visual and auditory visual and auditory cues, which are used in combination for engagement prediction. Figure 1 shows a flowchart overview of our methodology, while the features and steps are described in more detail below.

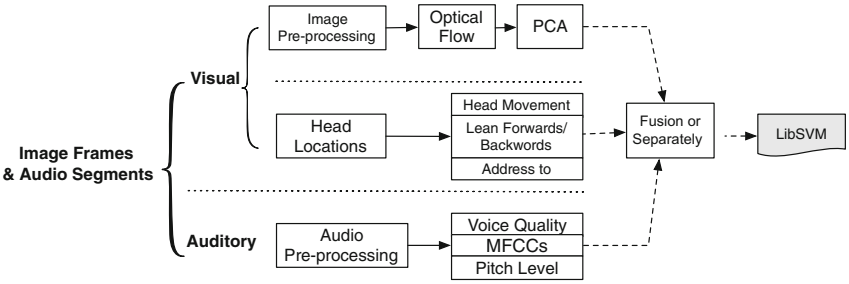


Fig. 1. Methodology overview

3.1 Optical Flow with Principal Component Analysis (PCA)

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 [10] 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.

3.2 Head Pose Related Features

The face detection and the yaw head position library were used from the work of [1], Camshift tracking [6] 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 30 fps videos.

3.3 Auditory Features

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.

3.4 Applying Additional Windows

The auditory features were extracted in a small window size, and the video data was recorded at 30 frames per second. However, changes in human cognitive state occur over a longer time frame, up to several minutes. To model these events more reliably, we tested additional window lengths. To make the visual results more reliable, we downsampled the video data from 30 frames per second to 3 frames per second. The method was motivated from previous studies [13]. For the audio, we calculated average feature values across longer window lengths.

4 Experiments and Results

To test the general performance of our engagement model, the LibSVM package [8] with RBF kernel was used for binary classification tasks with grid search method for best parameters selection, cost (set Cost, search from 2^{-5} *to* 2^{15})and gamma (set gamma, search from 2^{-15} *to* 2^{3}) to avoid overfit and underfit. The number of instances of each class used for training was balanced with an baseline accuracy of 50 %.

Fig. 2. Screenshot of TableTalk with face detection and head yaw angle

4.1 Data Sets and Annotation

TableTalk [2] is a 210-min corpus of group social conversations collected at the ATR Research Labs in Japan. A 360-degree camera was used to capture the frontal faces of participants chatting around a table. Audio was captured using a centre mounted microphone. Figure 2 shows a screen shot of the video of the corpus with face detection. The TableTalk corpus has been widely studied for social tasks e.g. Scherer [14] studied it for visual interaction management; Bonin F. investigated the engagement annotation study based on TableTalk corpus [5]. We annotated engagement levels on a 0–4 scale in maintain segment as shown in Table 1 and Fig. 3, and the binary classes of engaged (A) and not engaged (C) were analysed in this paper. The maintenance or central phase part engagement was annotated into different degrees, and was the focus of this analysis, rather than the initial phase or approach phase examined in other works [4].

4.2 Feature Analysis

Box plots of several features are shown in Fig. 4. The first two box plots from the left show the distributions of two selected visual cues - head pose (yaw)

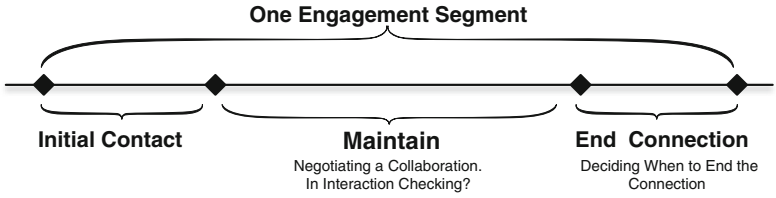
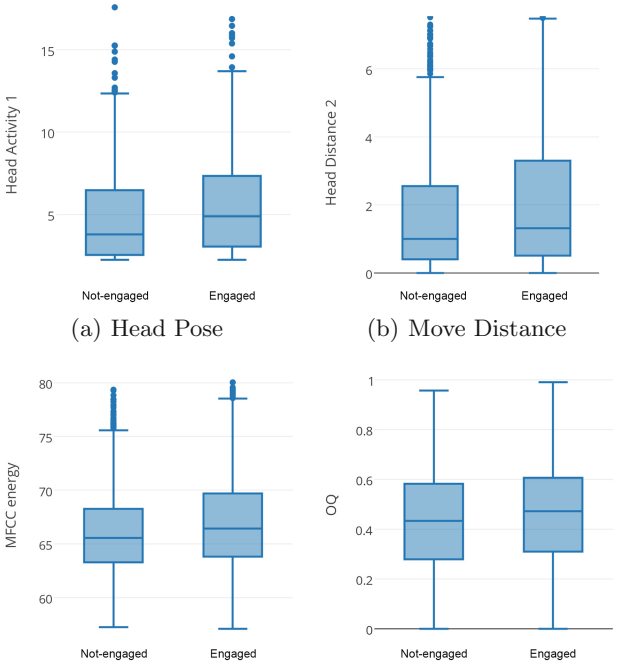


Fig. 3. One engagement segment

Table 1. Annotation rules

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			

and move distance. We observe that for these visual cues, as expected, the non-engaged category has lower values ($p<0.005$). The two plots on the right show the distributions of MFCC energy and Open Quotient (OQ). Again, non-engaged has lower values. Optical Flow visualization using Munsell Color System is shown in Fig. 5.



(a) Head Pose (b) Move Distance

(c) MFCC energy (d) OQ

Fig. 4. Box plots of selected features

4.3 Visual Cues Results

Table 2 shows the results for different combinations of visual features. The head backward/forward movement obtained the lowest accuracy rate. A higher result was obtained when head movement distance, optical flow and head yaw angle were considered together.

4.4 Acoustic Cues Results

Table 3 shows the results for the auditory features. Glottal and MFCCs features achieve a higher accuracy of 71 % than other acoustic feature sets.

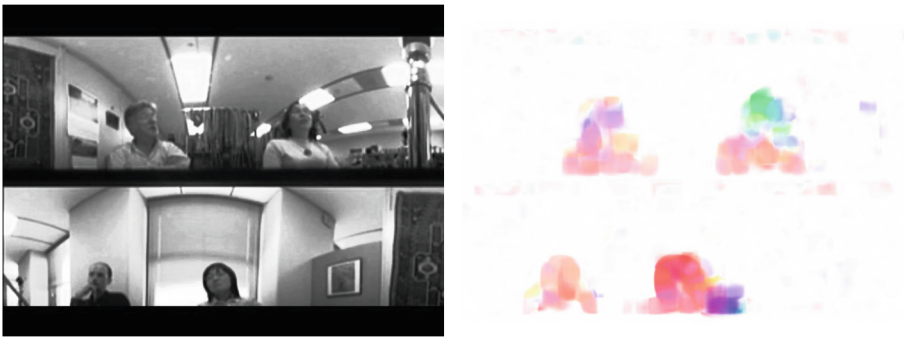


Fig. 5. Optical flow visulization

Table 2. Classification results of group engagement using visual features

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 %

Table 3. Classification results of group engagement using auditory features

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

4.5 Fusion Feature Set Results

A 'fusion' feature set consisting of both audio and video features was obtained by concatenating the visual and auditory vectors which had been time-aligned. The mean values of head move distance and head pose for the four speakers were calculated. The auditory features were extracted from a single recorded audio file containing all participants. Table 4 shows the results of the combination of these audio-visual feature sets with 82.23 % prediction rate. These results indicate that the combined audio-visual feature is better for detecting engagement than using the auditory and visual feature sets separately.

Table 4. Prediction results of combined features

Feature set	Accuracy	Recall	Precision	F-Score
Auditory and visual combined	82.23 %	0.822	0.816	0.815

5 Conclusion

Low level visual and auditory cues of engagement have been analysed 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, deep learning may also increase the results, which is conducted in the future works. Model-level and decision-level fusion will also be investigated in the future.

Acknowledgement. This research is supported by Science Foundation Ireland through the CNGL Programme (Grant 12/CE/I2267) in the ADAPT Centre and CHISTERA-JOKER project at Trinity College Dublin.

References

1. Libfacedetection. https://github.com/ShiqiYu/libfacedetection
2. Campbell, N.: The freetalk database. http://freetalk-db.sspnet.eu/files/
3. Argyle, M., Cook, M.: Gaze and mutual gaze (1976)
4. Bohus, D., Horvitz, E.: Learning to predict engagement with a spoken dialog system in open-world settings. In: Proceedings of the SIGDIAL 2009 Conference: The 10th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pp. 244–252. Association for Computational Linguistics (2009)
5. Bonin, F., Böck, R., Campbell, N.: How do we react to context? annotation of individual and group engagement in a video corpus. In: Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom), pp. 899–903. IEEE (2012)
6. Bradski, G.R.: Computer vision face tracking for use in a perceptual user interface. Intel Technol. J. **Q2**, 214–219 (1998)
7. Cassell, J., Bickmore, T., Billinghurst, M., Campbell, L., Chang, K., Vilhjálmsson, H., Yan, H.: Embodiment in conversational interfaces: Rea. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 520–527. ACM (1999)
8. Chang, C.C., Lin, C.J.: LIBSVM: a library for support vector machines. ACM Trans. Intell. Syst. Technol. (TIST) **2**(3), 27 (2011)
9. Ekman, P., Friesen, W.V.: The repertoire of nonverbal behavior: categories, origins, usage, and coding. Semiotica **1**(1), 49–98 (1969)
10. Farnebäck, G.: Two-frame motion estimation based on polynomial expansion. In: Bigun, J., Gustavsson, T. (eds.) SCIA 2003. LNCS, vol. 2749, pp. 363–370. Springer, Heidelberg (2003)

11. Gupta, R., Lee, C.C., Bone, D., Rozga, A., Lee, S., Narayanan, S.: Acoustical analysis of engagement behavior in children. In: WOCCI, pp. 25–31 (2012)
12. Gustafson, J., Neiberg, D.: Prosodic cues to engagement in non-lexical response tokens in swedish. In: DiSS-LPSS, pp. 63–66. Citeseer (2010)
13. Hsiao, J.C.y., Jih, W.r., Hsu, J.Y.: Recognizing continuous social engagementlevel in dyadic conversation by using turntaking and speech emotion patterns. In: Activity Context Representation Workshop at AAAI (2012)
14. Jokinen, K., Scherer, S.: Embodied communicative activity in cooperative conversational interactions-studies in visual interaction management. Acta Polytech. Hung. **9**(1), 19–40 (2012)
15. Lai, C., Carletta, J., Renals, S., Evanini, K., Zechner, K.: Detecting summarization hot spots in meetings using group level involvement and turn-taking features. In: INTERSPEECH, pp. 2723–2727 (2013)
16. Nakano, Y.I., Ishii, R.: Estimating user's engagement from eye-gaze behaviors in human-agent conversations. In: Proceedings of the 15th International Conference on Intelligent User Interfaces, pp. 139–148. ACM (2010)
17. Oertel, C., Salvi, G.: A gaze-based method for relating group involvement to individual engagement in multimodal multiparty dialogue. In: Proceedings of the 15th ACM on International Conference on Multimodal Interaction, pp. 99–106. ACM (2013)
18. Oertel, C., Scherer, S., Campbell, N.: On the use of multimodal cues for the prediction of involvement in spontaneous conversation. In: Interspeech 2011, pp. 1541–1544 (2011)
19. Rich, C., Ponsler, B., Holroyd, A., Sidner, C.L.: Recognizing engagement in human-robot interaction. In: 2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI), pp. 375–382. IEEE (2010)
20. Sidner, C.L., Lee, C., Kidd, C.D., Lesh, N., Rich, C.: Explorations in engagement for humans and robots. Artif. Intell. **166**(1), 140–164 (2005)
21. Yu, Z., Papangelis, A., Rudnicky, A.: Ticktock: a non-goal-oriented multimodal dialog system with engagement awareness. In: 2015 AAAI Spring Symposium Series (2015)