INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Conversational Engagement Recognition Using Auditory and Visual Cues

Yuyun Huang, Emer Gilmartin, Nick Campbell

Speech Communication Lab,
School of Computer Science and Statistics, Trinity College Dublin
huangyu@tcd.ie, gilmare@tcd.ie, nick@tcd.ie

Abstract

Automatic prediction of engagement in human-human and human-machine dyadic and multiparty interaction scenarios could greatly aid in evaluation of the success of communication. A corpus of eight face-to-face dyadic casual conversations was recorded and used as the basis for an engagement study, which examined the effectiveness of several methods of engagement level recognition. A convolutional neural network based analysis was seen to be the most effective.

Index Terms: Conversational Engagement level Recognition, Human Social Behaviour

1. Introduction

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 for human-human and human-machine conversations in dyadic and multiparty interaction scenarios is needed to evaluate the success of social and task-based communication. This model would have a wide range of applications in areas including spoken dialogue systems, event detection in videos of conversations, detection of user satisfaction when using designated devices, and online monitoring of success in service conversations.

In this paper, we discuss the phenomenon of engagement and highlight relevant work. We then describe our current work on recognition of engagement levels in face-to-face dyadic conversations using audio and visual cues. We describe our methodology, and report an evaluation study based on a corpus of non-task oriented (casual) human-human dialogues collected in our laboratory.

1.1. Engagement

Engagement is defined 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."* [1], [2]. Inference of engagement level is an obvious way to learn social behaviour factors such as interest in the theme of a conversation, bonding between interlocutors, and level of social rapport. Gatica-Perez relates displayed level of engagement to *interest*, which he defines as a term used *"to designate people's internal states related to the degree of engagement displayed, consciously or not, during social interaction."* [3]. Non-verbal behaviours such as facial expression, gesture, and posture play a major role for humans when inferring information from partners [4], and such non-verbal cues have been proposed as perceptible factors that can be used to estimate engagement level [5].

1.2. Related Work

There has been extensive study of social engagement, involvement or social interest in both human-human and human-machine scenarios, focussing on areas including definition of related concepts, annotation, engagement model design, and engagement prediction, particularly using auditory and visual cues.

An early example of engagement detection is found in Yu et al. (2004), where speech emotion recognition was adapted to perform user conversational engagement estimation. A support vector machine (SVM) was used to classify users' emotion as expressed in individual utterances from two corpora - one of acted emotion and the other of social telephone conversations. The emotion labels obtained from the SVM were then used as inputs to coupled hidden Markov models for detection of engagement states. [6]. Hsiao et al. (2012) investigated engagement level estimation based on the idea that social engagement, seen through patterns of turntaking and speech emotion, is an observable form of inner social interest. They collected 11 dyadic conversations from 9 participants over two iPhone 3Gs and annotated the engagement level on a scale from 1 (strongly disengaged) to 4 (strongly engaged). A hierarchical model of speech and turntaking features was used to classify into two levels (low and high), with accuracy close to 80% [7].

Multimodal visual and audio cues including gaze, blinking, pitch level and intensity were used by Oertel et al. (2011) for involvement level prediction. They annotated involvement levels scaling from 0-10 (0 being the lowest level, 10 the highest) on the D64 human-human conversation corpus [8], and modelled involvement level using features drawn from manually annotated mutual gaze and blinking visual features, and automatically extracted acoustic features including pitch level and intensity [9]. Salam et al. (2015) explored engagement detection in conversations between 8 participants and a Nao Robot. Features such as head nods, head pose, face location, speaking and silence periods, and 'addressing the speech to someone' were extracted. The engagement level in their work is binary, distinguishing engagement and disengagement. Sanghvi et al. (2011) investigated detection of engagement with a game companion (a cat-like robot), they annotated engagement levels into yes or no with the detailed reasons that the user was engaged or not engaged during the segments. Motion features such as body lean, and silhouette were used for classification in Sanghvi's work [10].

Apart from engagement level recognition, Bohus et al. (2009) introduced a machine learning approach to predict engagement intentions in the interaction between a human and

http://dx.doi.org/10.21437/Interspeech.2016-846

an avatar dialogue system. They used location features, face-frontal features and manually labelled attention features from the training feature sets [11]. Anzalone et al. (2015) evaluated engagement with Social Robots (Nao and iCub respectively), using recorded head pose and body gesture to generate a set of metrics and to show how engagement perception was sensed by a human, and how engagement levels changed during the interaction [12]. Hall et al. (2014) investigated the effects of a robot having natural human familiarity engagement responses of nodding, blinking, and gaze direction when interacting with participants [13]. Yu et al. (2015) created an engagement awareness dialogue system named TickTock [14], where engagement analysis constituted an important part of the dialogue system allowing the system sense the states of participants and guide the dialogue manager to decide a suitable conversation strategy.

In our work, we focus on engagement level recognition in human-human face to face conversation. We explore the use of several visual and auditory features familiar from emotion detection but previously unused in the recognition of engagement level, using training methods including Local Binary Patterns (LBP) with Principal Component Analysis (PCA), detailed facial movements, loudness shape features, and convolutional neural networks (CNN). We also perform comparisons with previously used features such as Mel Frequency Cepstral Co-efficients (MFCCs).

2. Methodology

In order to predict different levels of engagement using non-verbal cues, we explored a multi-modal visual and auditory feature-based learning method and deep learning using the convolutional neural networks.

2.1. Feature based "shallow" learning with Visual Cues

Both appearance and geometric hand-crafted visual features were used. Texture feature extraction was based on local binary patterns (LBP) with principal component analysis (PCA) for the dimensionality reduction. Geometric features were computed from 51 extracted facial landmarks.

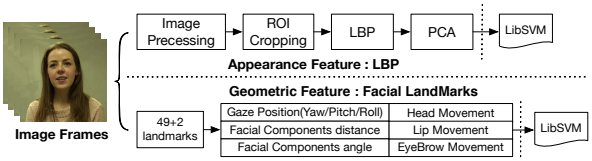

Figure 1: *Visual features based learning overview.*

2.1.1. LBP+PCA

Ojala et al [15] introduced the basic LBP operator and later Extended LBP (aka Circular LBP) [16]. Circular LBP can encode more details thanks to the flexible arbitrary radius and sample points, although computational cost is higher. A simple circular LBP with 8 neighbours and radius set to 1 was used for texture feature extraction in the work described here.

A simple Haar-like feature-based face detection [17], [18] was used for cropping the face images. All face images were resized to 98 x 115 pixels and converted to grey scale as inputs for computing the LBP operators. Each face image was divided into 4x5 sub-areas, the amount of LBP code is 4 x 5 x (2^8) = 5120. Some very minor Gaussian smoothing was also added. The obtained LBP feature was dimensionality reduced by applying PCA before classification.

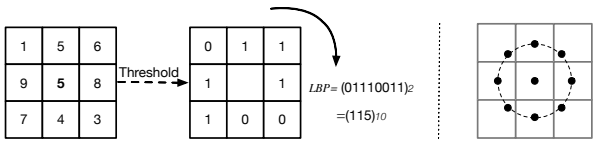

Figure 2: *Left: Basic LBP (3x3); Right: Circular LBP with radius = 1 and neighbours = 8*

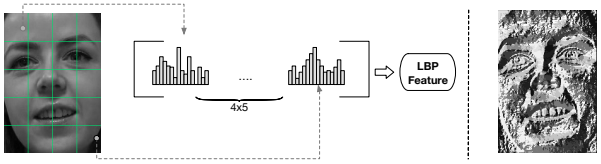

Figure 3: *Left: divide the face image into 4x5 sub-regions; Right: Visualization of Circular LBP with a tiny bit smooth*

2.1.2. Facial Landmarks + Head Pose

Face landmarks and head pose estimation were implemented based on work in [19] - 49 facial landmark locations with two eye centre locations plus head pitch/yaw/roll positions were extracted from each input frame. Eyebrow/lip movement, facial component angle and distance were calculated based on the landmarks. Facial angle/distance is motivated by Hernandez's work [20], as illustrated in Figure 4.

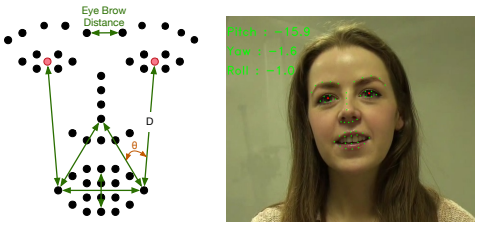

Figure 4: *Facial landmarks and shapes*

2.2. Features based "shallow" learning with Auditory Cues

Low-level features like pitch level, MFCCs, and loudness were extracted for auditory cues. Pitch features have been studied before, with researchers reporting different recognition results - Yu's work found that pitch is insignificant for engagement [21], in contrast to Voigt's reported results [22]. An overview of the auditory features contribution is shown in Figure 5.

2.2.1. Pitch, MFCCs and Loudness

The encoded wav file was downsampled to 16kHz, features were extracted in a frame window 500 samples or 500/16k = 31.25ms wide. The step size was half of the window size containing 250 samples and causing 250 samples overlap. Every second, there were 16k/250 = 60 feature vectors. The window size of basic feature extraction is small, while human social,

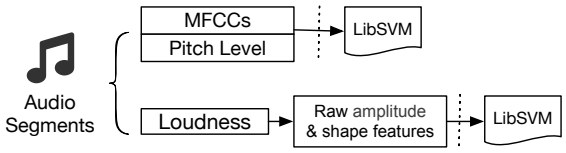

Figure 5: *Auditory features based learning overview*

emotional and cognitive states may last for several minutes. Therefore, after the basic features were obtained, a window size of half a second was applied to the feature set to generate an averaged feature set, using the method employed in previous work by Hernandes and by Hsiao [20], [7]. The newly generated feature set contains: pitch level, 12 MFCCs, and loudness.

2.2.2. *Auditory Shape and Angle Features*

A shape and angle method was performed on the loudness curve from our data, inspired by the method used by Hernandes[23] to generate features related to heart rate, which were then used in classification of body movements or positions such as sitting and lying. In our work, the two maximum and two minimum values in a period were selected. The four descriptive points marked in green showed in Figure (6) were used to generate features related to the distance and angle in the loudness curve. These features may be useful in inferring other information like duration of high or low speech and the gradient of changing voice.

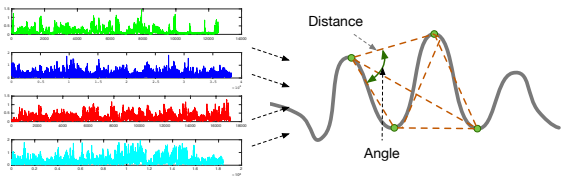

Figure 6: *Auditory Shape and angle feature*

2.3. Deep learning

We used a traditional 3 convolutional layer convolutional neural network with 64, 64, 128 filters in each layer. The activation function was Rectified Linear Unit (ReLU). The convolutional layer was followed by a max-pooling layer, using dropout of 0.2. The CNN codes were extracted for transfer learning for future work.

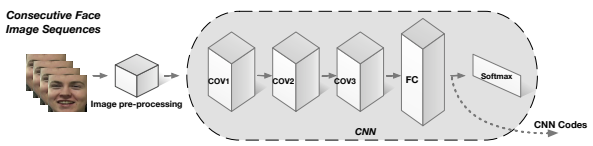

Figure 7: *Convolutional neural networks structure*

3. Evaluation

3.1. Material

As part of our ongoing work, we are collecting a corpus of human-human spontaneous causal face to face dyadic conversation among English native speakers. Conversation length ranges from 8 to 22 minutes. The recording was carried out in a quiet room, avoiding any background noise. Participants sat face to face. Each participant wore a lapel microphone. Two HD video recorders were placed between the two speakers to record a frontal view of each face, while a 360 degree camera and a low resolution birds-eye view camera were also used to record the overall recording procedure. The raw recorded video file is 60 frames per second, and the audio file is 48kHz with 16 bits. For our current study, we used data from 8 conversations involving 15 native English speakers (one speaker was repeated), a total of 257 minutes of data. The data was annotated by two annotators trained on engagement concepts, and generated a Cohen's Kappa coefficient of 0.87.

We also annotated the Cardiff natural conversations(CCDb) [24] which contain 6 subjects, eight 5-minute dyadic conversations, but found very few segments of disengagement and no strong disengagement. In the data we collected, strong disengagement was very rare; only one male participant appeared unwilling to continue the conversation with his female partner and finally ended the conversation by just sitting with no mutual gaze interaction. In Yu's work on human-machine interaction, strong engagement and strong disengagement were very rare [14]. However, in contrast to Yu's human-machine results, we found strong engagement quite often in the human-human datasets we have analysed - both our own corpus and the CCDb corpus.

The proposed methods were evaluated on recognition of engagement levels scaling from 0 to 3 as shown in table 1. A 10-fold cross-validation procedure was used to evaluate generalization performance. LibSVM [25] with one-to-one approach for multi-class classification was used. The Torch deep learning framework [26] was used for ConvNet.

5-level Engagement Annotation		
End of the previous segment		
Engagement Initialization		
Maintain	0. Strong 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. Nature	Neither show interest or lack of interest
	3. Disengaged	Less interest in the conversation
	4. Strong Disengaged	No interest to continue the conversation at all, want to leave the conversation
End Connection		

Table 1: Engagement annotation scheme

3.2. Evaluation: LBP + PCA

8000 cropped face images were randomly selected from the whole dataset for LBP feature extraction - each class contained 2000 images. Each image had 5120 LBP operators as inputs to PCA. The first 200 principal components (PCs) were kept and accounted for 96.14% of total principal component variance. Figure 8 shows a 3D plot of the first three principal components. The radial basis function (RBF) kernel type and grid search approach for best parameters were used. 10-fold cross-validation was carried out for the 4-class recognition task. Prediction accuracy was 85.75% (Precision: 0.857, Recall: 0.858, F-sore: 0.856), compared to an appropriate baseline of 25% in our case

(each engagement level has 2000 instances, thus 2000/8000 = 25% can be used as the baseline).

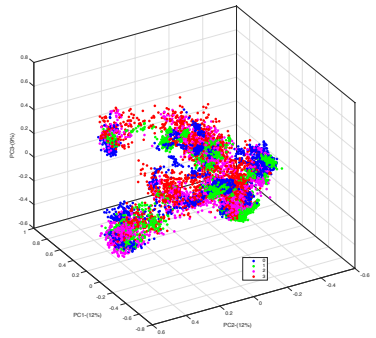

Figure 8: *PCA visualization with the first three PCs*

3.3. Evaluation: Head pose + Landmark

A feature set comprising raw head pose with yaw/pitch/roll resulted in recognition accuracy of 46.49% (Precision: 0.470, Recall: 0.465, F-Measure: 0.464), a 21% absolute improvement on baseline (25%). Each engagement level used 2827 frames. For the facial landmark feature set, seven facial component distances and three angles were computed based on the landmark locations shown in Figure 4. This feature set achieved accuracy of 65.39 % (Precision: 0.653, Recall: 0.654, F score: 0.653). Combining head pose and facial component features led to higher accuracy – 88.77 % (Precision: 0.895, Recall: 0.888, F-Measure: 0.889).

3.4. Evaluation: Acoustic features

After applying 0.5 second windows as described in section 2.2.1, the middle level features were directly used for classification. This evaluation was prone to overfitting as some annotated segments of nature and disengagement contained silent parts. Silence does not seem to carry any information for low level acoustic features, while it represents a significant amount information in the engagement case (e.g. silence is observed when participants are thinking, listening etc.). A crude method was firstly implemented by cropping out all the silent parts. A second method was used based on work in [7], [27], where length of silence, and number of turns were computed within every 10 seconds windows.

We only present 2-level engagement (1 - low engagement, 0 - strong engagement) prediction results here using the cropped data set with no silence interruptions with the combined MFCCs and pitch level features. The crude method resulted in 57.28% accuracy (close to the baseline 50%). Results for the second method did not show much improvement over this, at 62.29%, they were only 12% better than random guessing. Other kernel types of linear, sigmoid and polynomial were tried, without an improvement in results. These results may be due to the particular data set we used, or the features themselves may not be suitable for recognition of engagement.

3.5. Evaluation: Auditory shape features

The auditory loudness raw amplitude features of two classes: strong and low engagement were used for prediction, obtaining 68.049% prediction accuracy. (771 instances / 1542 instances = 0.5 as the baseline). Peak and minimum values of subset input signal vectors were found. Distance and angle were computed from the selected two smallest plus two largest values as shown in Figure 6. Each sub-basement contains 6 distances and 4 angles are used as feature vectors. Auditory shape features resulted in slightly higher accuracy of 76.78% (Precision: 0.788, Recall: 0.768, F-Measure: 0.731) compared to just using loudness amplitude, and were 26% better than the baseline of 50%.

3.6. Evaluation: CNN

12528 cropped face images were selected to train the CNN, each level had 3132 images. 80% of total images (10022) were used for the training set and the remaining 20% (2506) for a test dataset. Data argument can extend the possible training set to a much larger amount by applying several transforms, random scales/crops or image jitter and flip, while increasing computational consumption. As a first attempt for the 4-classes engagement level recognition, the original dataset was used. The average rows correct was 91.689%, average rowUcol correct (VOC measure) was 84.72%, and the global correct score was 91.660%.

	Disengagement	Nature	Engagement	Strong Engagement	Row Precision
Disengagement	603	23	11	1	94.514%
Nature	5	583	18	5	95.417%
Engagement	12	26	552	48	86.52%
Strong Engagement	3	19	38	559	90.307%

Table 2: CNN Confusion Matrix

4. Conclusions

The use of facial detail components features as in Hernandez's work [20] obtain higher accuracy in our implementation, even though the method was originally proposed for analysis of states of TV viewers rather than interlocutors. Results from circular LBP features were slightly lower (3%) than facial components with head pose. We believe using other LBP methods, such as manifold LBP, may get a better results. Both texture and geometry features are useful for engagement recognition. Low level acoustic cues, pitch level and MFCCs, did not perform well in our work. However, loudness was significant when analysing engagement using auditory features, with improved performance when dynamic characteristics were considered. Further data collection may be needed to explore engagement in phone conversations or longer face to face conversations. In addition, this data set was collected in a experimental environment, so there is plenty of scope for work to explore engagement models in the wild. Fusion of visual and auditory cues is especially needed for engagement research as the occurrence of engagement during silent periods requires visual cues, and all recognition is enhanced by the use of both modalities. The CNN method obtained the best results, and thus we plan to further investigate deep learning methods and multimodal feature fusion for engagement research in the futures.

5. Acknowledgements

This research is supported by Science Foundation Ireland through the CNGL Programme (Grant 12/CE/I2267) in the ADAPT Centre (www.adaptcentre.ie) at Trinity College Dublin, and by the European Coordinated Research on Long-term Challenges in Information and Communication Sciences Technologies ERA-NET (CHIS-TERA) JOKER project, JOKe and Empathy of a Robot/ECA: Towards social and affective relations with a robot.

6. References

[1] C. L. Sidner, C. Lee, C. D. Kidd, N. Lesh, and C. Rich, "Explorations in engagement for humans and robots," *Artificial Intelligence*, vol. 166, no. 12, pp. 140 – 164, 2005. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0004370205000512

[2] C. L. Sidner and M. Dzikovska, "Human-robot interaction: engagement between humans and robots for hosting activities," in *Multimodal Interfaces, 2002. Proceedings. Fourth IEEE International Conference on*, 2002, pp. 123–128.

[3] D. Gatica-Perez, "Modeling interest in face-to-face conversations from multimodal nonverbal behavior," in *In J.-P. Thiran, H. Bourlard, and F. Marques, Multimodal Signal Processing, Academic Press*. Academic Press, 0 2009.

[4] P. Ekman and W. V. Friesen, "The repertoire of nonverbal behavior: Categories, origins, usage, and coding," *Semiotica*, vol. 1, no. 1, pp. 49–98, 1969.

[5] E. Delaherche, M. Chetouani, A. Mahdhaoui, C. Saint-Georges, S. Viaux, and D. Cohen, "Interpersonal synchrony: A survey of evaluation methods across disciplines," *IEEE Transactions on Affective Computing*, vol. 3, no. 3, pp. 349–365, July 2012.

[6] C. Yu, P. M. Aoki, and A. Woodruff, "Detecting user engagement in everyday conversations," *arXiv preprint cs/0410027*, 2004.

[7] J. C.-y. Hsiao, W.-r. Jih, and J. Y.-j. Hsu, "Recognizing continuous social engagement level in dyadic conversation by using turn-taking and speech emotion patterns," in *Workshops at the Twenty-Sixth AAAI Conference on Artificial Intelligence*, 2012.

[8] C. Oertel, F. Cummins, N. Campbell, J. Edlund, and P. Wagner, "D64: A corpus of richly recorded conversational interaction," in *Proceedings of LREC 2010, Workshop on Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality*, 2010.

[9] C. OERTEL GEN BIERBACH, "On the use of multimodal cues for the prediction of involvement in spontaneous conversation," *INTERSPEECH-2011, Interspeech*, pp. 1541–1544, 2011.

[10] H. Salam and M. Chetouani, "Engagement detection based on mutli-party cues for human robot interaction," in *Affective Computing and Intelligent Interaction (ACII), 2015 International Conference on*. IEEE, 2015, pp. 341–347.

[11] D. Bohus and E. Horvitz, "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*. Association for Computational Linguistics, 2009, pp. 244–252.

[12] S. M. Anzalone, S. Boucenna, S. Ivaldi, and M. Chetouani, "Evaluating the engagement with social robots," *International Journal of Social Robotics*, vol. 7, no. 4, pp. 465–478, 2015.

[13] J. Hall, T. Tritton, A. Rowe, A. Pipe, C. Melhuish, and U. Leonards, "Perception of own and robot engagement in humanrobot interactions and their dependence on robotics knowledge," *Robotics and Autonomous Systems*, vol. 62, no. 3, pp. 392 – 399, 2014, advances in Autonomous Robotics Selected extended papers of the joint 2012 {TAROS} Conference and the {FIRA} RoboWorld Congress, Bristol, {UK}.

[14] Z. Yu, A. Papangelis, and A. Rudnicky, "Ticktock: A non-goal-oriented multimodal dialog system with engagement awareness," in *2015 AAAI Spring Symposium Series*, 2015.

[15] T. Ojala, M. Pietikäinen, and D. Harwood, "A comparative study of texture measures with classification based on featured distributions," *Pattern recognition*, vol. 29, no. 1, pp. 51–59, 1996.

[16] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 971–987, Jul 2002.

[17] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, vol. 1. IEEE, 2001, pp. I–511.

[18] R. Lienhart and J. Maydt, "An extended set of haar-like features for rapid object detection," in *Image Processing. 2002. Proceedings. 2002 International Conference on*, vol. 1. IEEE, 2002, pp. I–900.

[19] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic, "Incremental face alignment in the wild," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1859–1866.

[20] J. Hernandez, Z. Liu, G. Hulten, D. DeBarr, K. Krum, and Z. Zhang, "Measuring the engagement level of tv viewers," in *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*, April 2013, pp. 1–7.

[21] Z. Yu, "Attention and engagement aware multimodal conversational systems," in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, ser. ICMI '15. New York, NY, USA: ACM, 2015, pp. 593–597. [Online]. Available: http://doi.acm.org/10.1145/2818346.2823309

[22] R. Voigt, R. J. Podesva, and D. Jurafsky, "Speaker movement correlates with prosodic indicators of engagement," in *Speech Prosody*, vol. 7, 2014.

[23] J. Hernandez, D. J. McDuff, and R. W. Picard, "Bioinsights: Extracting personal data from still wearable motion sensors," in *Wearable and Implantable Body Sensor Networks (BSN), 2015 IEEE 12th International Conference on*. IEEE, 2015, pp. 1–6.

[24] A. J. Aubrey, D. Marshall, P. L. Rosin, J. Vandeventer, D. W. Cunningham, and C. Wallraven, "Cardiff conversation database (ccdb): A database of natural dyadic conversations," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2013 IEEE Conference on*, June 2013, pp. 277–282.

[25] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011, software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.

[26] "Torch | Scientific computing for LuaJIT." [Online]. Available: http://torch.ch/

[27] M. Bruijnes *et al.*, "Computational models of social and emotional turn-taking for embodied conversational agents: a review," *Centre for Telematics and Information Technology, University of Twente*, 2012.