On the use of multimodal cues for the prediction of degrees of involvement in spontaneous conversation

Catharine Oertel [1], *Stefan Scherer*[2], *Nick Campbell*[1]

[1]Speech Communication Lab, Trinity College Dublin, Dublin, Ireland
[2] Institute of Neural Information Processing, Ulm University, Ulm, Germany
oertelgc@tcd.ie, stefan.scherer@uni-ulm.de, nick@tcd.ie

Abstract

Quantifying the degree of involvement of a group of participants in a conversation is a task which humans accomplish every day, but it is something that, as of yet, machines are unable to do. In this study we first investigate the correlation between visual cues (gaze and blinking rate) and involvement. We then test the suitability of prosodic cues (acoustic model) as well as gaze and blinking (visual model) for the prediction of the degree of involvement by using a support vector machine (SVM). We also test whether the fusion of the acoustic and the visual model improves the prediction. We show that we are able to predict three classes of involvement with an reduction of error rate of 0.30 (accuracy =0.68).

Index Terms: involvement, multimodality, spontaneous speech, blinking, gaze

1. Introduction

Spontaneous conversations are characterised by various degrees of the participants' involvement [1] and emotional engagement [2], sometimes leading to what has been called conversational hotspots [3]. For this study, we define involvement as a group variable which is calculated as the average of the degree to which individual people in a group are engaged in spontaneous, non-task-directed conversations. Listerners are able to detect these degrees of involvement as integral parts of their daily interactions, basing their perception on multimodal cues. We use multimodal technology when we "skype" with our family and friends, when we record movies on our smartphones, and when we communicate with our business partners via video conferencing. In fact, we have so much multimodal data available that it has become rather difficult to keep track of. Wrede and Shriberg argue that if we were able to automatically detect these conversational hotspots, we could use them as a means to query huge multimodal databases in a time-efficient manner. A further possible application as suggested by Yu et al [2] could be found in the domain of telephony. Information about high involvement could be used to increase the richness of the communication system by adding an additional duplex channel or even video channel [2], to better facilitate interaction.

In order to approach the automatic detection of this intuitive, impressionistic but important feature the following steps need to be achieved. (1) Access to multimodal databases is necessary, (2) development of an annotation scheme to manually label the multimodal databases is required, and (3) acoustic and visual cues which are correlated with involvement need to be identified. A multitude of multimodal corpora, such as the AMI [4] or the CID [5] corpus, has been made available to the public in recent years. They either focus on dyadic conversation or are scenario specific. As a lot of communication captured by video technology is, however, not restricted to dyadic or task-specific conversations. Therefore, for this study, we use the D64 corpus, featuring 5 participants in non-task-directed, spontaneous conversation for this study [6].

While Wrede and Shriberg label involvement as a binary phenomenon, Dillon [7] uses a slider to let participants indicate their level of involvement. Following Dillon, we prefer to consider involvement as a scalar phenomenon rather than a binary one. In recent work [1], we confirmed that the acoustic cues, f_0-median, f_0-range and rms-intensity are strongly correlated with involvement and our results show that involvement is indeed a scalar phenomenon rather then a binary one.

However, it can be assumed that involvement is not only conveyed by means of prosodic cues but that other modalities are used as well. Two phenomena that are reported in the literature to be relevant in social interaction are gaze[8] [9] [10] and blinking [11]. Gaze, for example, has been found to be important for modulating and directing attention [10] and blink rate has been found "to vary systematically with specific behavior such as reading, conversing, watching film" [11]. Given their importance in social interaction it can be hypothesised that gaze and blinking rate are key factors for conveying involvement.

In this study we first investigate the correlation between visual cues (gaze and blinking rate) and involvement. We then test the suitability of prosodic cues (acoustic model) as well as gaze and blinking (visual model) for the prediction of the degree of involvement by using a support vector machine (SVM). We also test whether the fusion of the acoustic and visual model improves the prediction.

2. Data

The D64 corpus [6] was recorded over two successive days in a rented apartment, resulting in a total of eight hours of multimodal recordings subdivided into 3 sessions; session I, session II, and session III. For our analysis, two 30 minutes long intervals from sessions I and II were extracted. Sessions I and II differ, as in session I the conversation was mainly socially orientated whereas session II was mainly work oriented. There were 5 people (3 male; 2 female) present in session I but only 4 (3 male; 1 female) in session II. They were colleagues and/or friends, ranging in age from their early twenties to their early sixties. The conversation was not directed and ranged widely across topics both trivial and technical. While in session I, all participants contributed to the discussion, in session II the conversation was mainly dominated by two speakers (speaker F and speaker C).

3. Methods

3.1. Annotations

3.1.1. Annotation of Involvement

Involvement is defined as a group variable. It is assumed to capture the average degree to which individual people in a group are engaged in spontaneous, typically non-task-directed conversations. Involvement was annotated on a scale from 0-10 (0 being the lowest degree of involvement; 10 the highest) across the conversation. A perception test was conducted and the Inter-rater reliability was found to be $\kappa = 0.56$ for 30 raters [12].

3.1.2. Annotation of Gaze and Blinking

Blinking and gaze were annotated according to the annotation scheme proposed by Cummins [11]. For gaze, a distinction is made between "g" and"x". "g" is the abbreviation for gaze and is used in the case when speaker a is looking at speaker b. "x" is used when speaker x is not looking at speaker y. Gaze annotations were carried for two participants. Blinks are treated as single points in time and are annotated at "the first moment in which the visible part of the cornea [is]substantially occluded " [11]. In our annotations, we distinguished between ten subcategories of blinks based on duration and direction of gaze (for a full details see [13]), however, we conflated the blink categories to the sum of all blinks for a given interval for the use in this study.

3.1.3. Calculation of Mutual Gaze and Blinking

The annotation of mutual gaze and blinking are based on session II alone. Mutual Gaze is calculated as the proportion of the duration in which speaker F and speaker C simultaneously look at each other. Blinking rate is calculated as the sum of all blinks over time (we are only reporting on the blinking rate of speaker F, not speaker C).

3.2. Automatic Prediction

In order to determine the extent to which these features can be used on unseen corpus data we employ support vector machines (SVM) with radial basis function kernels for the automatic prediction [14, 15]. Further, we test early (feature level) and late (decision level multiplication fusion) fusion approaches, making use of the two modalities provided, for the prediction of involvement within session II [16]. Early level fusion combines the extracted unimodal features to a multimodal representation of the observations before classification. In this case the alignment of the observations in the different modalities is crucial for the training of a single multimodal classifier. In contrast to early fusion, late fusion, or decision level fusion, combines the decisions of multiple unimodal classifiers. Typical combination schemes comprise majority vote, or multiplication fusion [17]. The two fusion schemes investigated , early and late fusion, differ with respect to the time at which the information of the different modalities is combined. Since no annotated video data was available for session I and only audio features are common to both sessions, we compare both strategies in this study in order to be able to compare the results. At this stage of the work we are not considering sequential implications of these features that might be better modelled by statistical models such as hidden Markov models, and we consider an svm sufficient for the prediction.

4. Results

4.1. Acoustic Features

As reported in [1], we showed that the prosodic parameters f_0-median, and f_0-range, and rms-intensity are correlated with involvement. The higher the involvement the higher the f_0-median, the f_0-range and rms-intensity. Our results also suggest that involvement is a scalar rather than binary phenomenon (see Table 1).

Table 1: ANOVA analysis confirming the relationships between three acoustic categories and four levels of involvement.

	inv. 5-6	**inv.6-7**	**inv.7-8.**
f_0**-median**	p=0.006370	n.s.	p=0.0106
f_0**-range**	n.s.	p<0.001	n.s.
intensity	p<0.001	p<0.001	n.s.

4.2. Mutual Gaze and Involvement

The visual cue mutual gaze as can be seen in Figure 1 illustrates an increase in the proportion of mutual gaze the higher the involvement. Beyond point 6.8 on the x-axis we see indications of a bimodal distribution. This can be explained as in one set speaker C takes notes on her laptop Figure 3, whereas she does not in the other set Figure 2. For both individual sets, a high correlation (R=0.93 and R=0.96) between the proportion of mutual gaze and involvement was found.

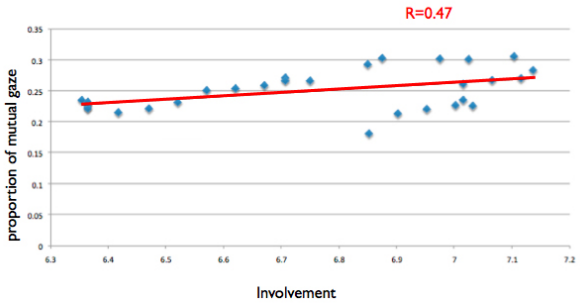

Figure 1: Mutual Gaze for the whole interaction and involvement for 20 second intervals with a moving window of 10 seconds.

4.3. Blinking and Involvement

The visual cue blinking rate (blinks per second) only shows one significant change between the various involvement states. While there is no significant change between involvement level 5 and involvement level 6 (F(4,511)=7.214); p=0.08763) there is a significant change between involvement level 6 and involvement level 7 (F(3,478)=7.594; 0.01279). The blinking rate in involvement level 7 is significantly lower than in involvement level 6. However, the change from involvement level 7 to involvement level 8 is not significant (F(2,322)=8.378; p= 0.310814). There is not a sufficient number of blinks for involvement level 4 and involvement level 9 for any sufficient analysis.

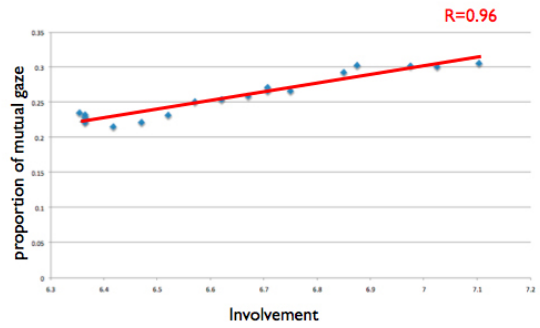

Figure 2: Mutual Gaze and involvement without laptop interference for 20 second intervals with a moving window of 10 seconds

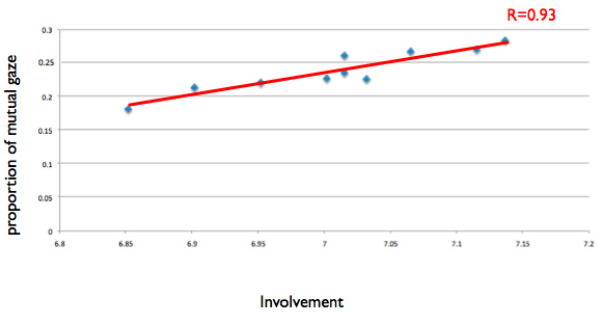

Figure 3: Mutual Gaze and involvement with laptop interference for 20 second intervals with a moving window of 10 seconds

4.4. Prediction

A list of the features used in the acoustic and visual model can be found in Table 2.

Table 2: List of features comprised in the acoustic and visual model.

acoustic model	visual model
f_0-median	(mutual) gaze
f_0-range	blinking rate
f_0-max	
f_0-sd	
f_0-min	
intensity	

In [1] we showed that the proposed cues can be used to distinguish between different levels of involvement. To show that involvement is scalar rather than binary we compared two different models for the prediction of involvement. One based on a two class model (Model I) and the other on a three class model (model II). In Model I, the first class contained data for low involvement (level 4,5 and 6), and the second class contained data of high involvement (level 7,8 and 9). Model II contained a class of low (level 4, 5 and 6) and class of high involvement (level 8 and 9). Moreover, we introduced an intermediate class (level 7) of involvement due to high proportion of annotations obtained for involvement level 7.

Table 3: Prediction results for involvement.

	Model I (two classes)		**Model II (three classes)**	
	mean acc.	**ERR**	**mean acc.**	**ERR**
Early fusion	0.7440	**0.11**	0.6820	**0.30**
Late fusion	0.7420	0.11	0.6420	0.26
Audio only	0.6940	0.06	0.5060	0.12
Video only	0.6640	0.03	0.6060	0.22

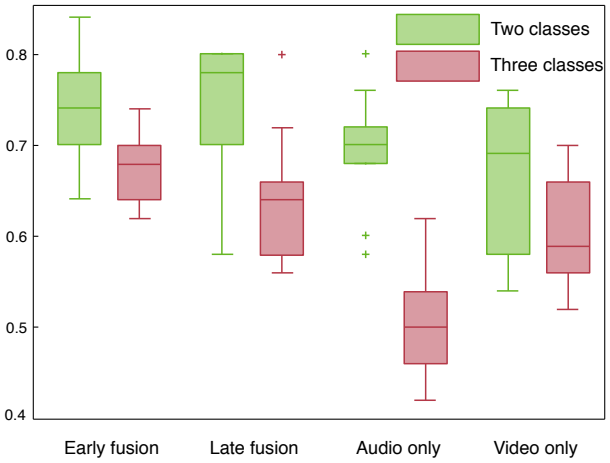

Figure 4: Plot of the accuracies of the 10 fold cross validation experiments using the data of session II, comparing early and late fusion as well as the single modalities in both two and three class cases. The boxes denote 50% of the data and the median value is shown as the middle line of the plot. Whiskers include 1.5 times the standard deviation of the data and outliers marked as crosses are further away from the median.

Table 3 and Figure 4 illustrate the results of the experiments. Error rate reduction (ERR) is calculated as an improvement in accuracy rate from a hypothesised classifier relying on the a priori probability of the most likely class (for Model I that is class 2 with 0.63; for Model II that is class 2 with 0.38). The results are based on a standard 10 fold cross validation with a 90/10 split of the available data.

Concerning Model I the best performance is achieved for an early fusion of both audio and video data (ERR = 0.11 ; accuracy = 0.7440). The late fusion of the audio and video data has a similar ERR of 0.11 and an accuracy of 0.7420. Video only produced lower accuracy and only a small reduction in error rate (ERR = 0.03; accuracy = 0.66). The single modality approaches are both significantly outperformed by both of the fusion approaches in paired t-tests over the ten fold cross validation (late fusion vs. audio only $p = .011$; late fusion vs. video only $p = .023$; early fusion vs. audio only $p = .008$; early fusion vs. video only $p = .002$).

The best performance overall in terms of ERR is achieved for Model II using an early fusion of both acoustic and visual data (ERR = 0.30; accuracy = 0.6820). The late fusion of the acoustic and video data has a ERR of 0.26 and an accuracy of

0.64. The least accurate results are achieved for audio data only (ERR = 0.12 ; accuracy = 0.50). The early fusion again outperforms the single modalities significantly (early fusion vs. audio only $p < .001$; early fusion vs. video only p = .002), but late fusion only outperforms the audio only approach ($p < .001$). Further, video only outperforms audio only in the paired t-test with a p-value $< .001$.

In order to test how well the model generalises we trained it on session II and tested it on session I and achieved a prediction accuracy of 0.5830 (ERR = 0.20) for Model II, which shows a good generalisation performance.

5. Discussion

We found a very high correlation between mutual gaze and involvement. The fact that speaker C, however, at a certain point started to take notes on her laptop resulted in a decrease of mutual gaze. This finding is not only obvious, it is also in line with the findings of Argyle & Graham [9] who found that an object relevant to the conversation will reduce the amount to which the conversants look at each other. This finding confirms that (mutual) gaze is a very good indicator of involvement.

Speaker C's use of her laptop did not only influence mutual gaze but also the average movement of body and face. As reported in [1] we found that the movement of body and face of speaker C was strongly influenced by the laptop she carried on her lap. We are planning to solve this problem by applying session-based normalisations and will report on this in future work.

We showed that using the features from the audiovisual channels enables the classifier to successfully predict the involvement state. For three classes of involvement, the video model alone achieved better results than the audio model alone. In contrast, for two classes of involvement, the audio model alone performs slightly better than the video model alone. Furthermore, the fusion approaches, and especially the early feature level fusion used, outperforms the individual modalities significantly. This is in line with results of [18] who found the same improvement in a similar classification task. Additionally, we were able to show that the approach we utilised scales over different sessions and allows for a good generalisation.

6. Conclusion & Future Work

Our model is able to predict three classes of involvement with an reduction in error rate of 0.30 (accuracy = 0.68). Further, it is possible to generalise over unseen sessions using the unimodal SVM approach. Further analysis will be carried out towards measuring involvement between individual participants (inter-participant-involvement or social distance) in a group. This work will include further annotation of visual cues as well as research into the identification of further multimodal cues suitable for the prediction of involvement. Additionally, we plan to account for the human variability in the annotation of involvement by using a fuzzy SVM architecture [19].

7. Acknowledgements

Catharine Oertel is supported by the Irish Research Council for Science, Engineering & Technology Embark Initiative. This work was undertaken as part of the FASTNET project - Focus on Action in Social Talk: Network Enabling Technology funded by Science Foundation Ireland (SFI) 09/IN.1/I2631. Stefan Scherer wishes to acknowledge the Transregional Collaborative Research Centre SFB/TRR 62 "Companion-Technology for Cognitive Technical Systems" funded by the German Research Foundation (DFG). The authors would like thank Dr. Fred Cummins, Prof. Petra Wagner, Dr. Celine de Looze and Dr. Brian Vaughan for their comments and help.

8. References

[1] C. Oertel, C. De Looze, S. Scherer, A. Windmann, and P. Wagner, "Towards the automatic detection of involvement in conversation," in *accepted at: The Proceedings of the SSPnet-COST 2102 PINK International Conference*. Springer, 2011.

[2] C. Yu, P. M. Aoki, and A. Woodruff, "Detecting user engagement in everyday conversations," in *8th International Conference on Spoken Language Processing (ICSLP ' 04)*, 2004, pp. 1329–1332.

[3] B. Wrede and E. Shriberg, "Spotting "Hot Spots" in Meetings: Human Judgements and Prosodic Cues." in *Proceedings of Eurospeech 2003*, Geneva, 2003, pp. 2805–2808.

[4] J. Carletta, "Unleashing the killer corpus: experiences in creating the multi-everything AMI Meeting Corpus," in *Language Resources and Evaluation*, 2007, pp. 181–190.

[5] R. Betrand, P. Blache, R. Espesser, G. Ferre, C. Meunier, B. Priego-Valverde, and S. Rauzy, "Le CID- Corpus of Interactional Data- Annotation et Exploitation Multimodale de Parole Conversationelle." *Phonétique et Phonologie*, vol. 49, no. 3, 2008.

[6] C. Oertel, F. Cummins, N. Campbell, J. Edlund, and P. Wagner, "D64: a corpus of richly recorded conversational interaction," in *Proceedings of Language Resources and Evaluation Conference (LREC'10)*, 2010, pp. 2992–2995.

[7] R. Dillon, *In: Lecture Notes in Computer Science: A Possible Model for Predicting Listener's Emotional Engagement*, R. Kronland-Martinet, T. Voinier, and S. Ystad, Eds. Heidelberg: Springer, 2006.

[8] A. Kendon, "Some functions of gaze-direction in social interaction," *Acta Psychologica*, vol. 26, pp. 22–63, 1967.

[9] M. Argyle and J. Graham, "The central Europe experiment: Looking at persons and looking at objects," *Environmental Psychology & Nonverbal Behavior*, vol. 1, no. 1, pp. 6–16, 1967.

[10] R. Vertegaal, R. Slagter, G. van Der Veer, and A. Nijholt, "Eye gaze patterns in conversations: there is more to conversational agents than meets the eyes," in *SIGCHI Conference on Human Factors in Computing Systems*, 2001, pp. 301–308.

[11] F. Cummins, "Gaze and Blinking in Dyadic Conversation: A study in Coordinated Behavior Among Individuals," *submitted*.

[12] C. Oertel, "Identification of Cues for the Automatic Detection of Hotspots," Master's Thesis, Bielefeld University, 2010.

[13] F. Cummins, "Annotating Blinks and Gaze." [Online]. Available: http://cspeech.ucd.ie/ fred/Blinking/

[14] K. P. Bennett and C. Campbell, "Support vector machines: hype or hallelujah?" *ACM SIGKDD Explorations Newsletter*, vol. 2, no. 2, pp. 1–13, 2000.

[15] B. Schölkopf and A. J. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. Cambridge, MA, USA: MIT Press, 2001.

[16] F. Schwenker, S. Scherer, M. Schmidt, M. Schels, and M. Glodek, "Multiple classifier systems for the recognition of human emotions," in *9th International Workshop on Multiple Classifier Systems (MCS 2010)*, N. El Gayar, J. Kittler, and F. Roli, Eds. Springer, 2010, pp. 315–324.

[17] L. Kuncheva, *Combining pattern classifiers: methods and algorithms*. Wiley, 2004.

[18] C.-C. Lee, S. Lee, and S. S. Narayanana, "An Analysis od Multimodal Cues of Interruption in Dyadic Spoken Interaction," in *Interspeech*, Brisbane, Australia, 2008, pp. 1678–1681.

[19] C. Thiel, S. Scherer, and F. Schwenker, "Fuzzy-input fuzzy-output one-against-all support vector machines," in *11th International Conference on Knowledge-Based and Intelligent Information and Engineering Systems (KES 2007)*, ser. Lecture Notes in Artificial Intelligence, vol. 3. Springer, 2007, pp. 156–165.