Towards the Automatic Detection of Involvement in Conversation

Catharine Oertel[1], Céline De Looze[1], Stefan Scherer[2], Andreas Windmann[3], Petra Wagner[3], and Nick Campbell[1]

[1]Speech Communication Laboratory, Trinity College Dublin, Ireland
[2]University of Ulm, Germany
[3]Bielefeld University, Germany

Abstract. Although an increasing amount of research has been carried out into human-machine interaction in the last century, even today we are not able to fully understand the dynamic changes in human interaction. Only when we achieve this, will we be able to go beyond a one-to-one mapping between text and speech and be able to add social information to speech technologies. Social information is expressed to a high degree through prosodic cues and movement of the body and the face. The aim of this paper is to use those cues to make one aspect of social information more tangible; namely participants' degree of involvement in a conversation. Our results for voice span and intensity, and our preliminary results on the movement of the body and face suggest that these cues are reliable cues for the detection of distinct levels of participants involvement in conversation. This will allow for the development of a statistical model which is able to classify these stages of involvement. Our data indicate that involvement may be a scalar phenomenon.

Key words: social involvement, multi-modal corpora, discourse prosody

1 Introduction

Language and speech, and later, writing systems, have evolved to serve human communication. In today's society human-machine interaction is becoming more and more ubiquitous. However, despite more than half a century of research in speech technology, neither computer scientists, linguists nor phoneticians have yet reached a full understanding of how the variations in speech function as a means of human communication and social interaction. A one-to-one mapping between text and speech is not sufficient to treat the social information exchanged in human interaction.

What makes a conversation a naturally interactive dialogue are the dynamic changes involved in spoken interaction. We propose that these changes might be explained by the concept of involvement. Following Antil [1] we define involvement as "the level of perceived personal importance and/or interest evoked by a stimulus (or stimuli) within a specific situation" [1].

Moreover, we consider involvement in our study to be a scalar phenomenon. Contrary to Wrede & Shriberg [2] who define involvement as a binary phenomenon, we agree with Antil in that "involvement must be conceptualized and operationalized as a continuous variable, not as a dichotomous variable" [1]. Similar to Dillon [3], who uses a slider to let participants indicate their level of emotional engagement, we used a scale from 1-10 in our annotation scheme to indicate distinct levels of involvement.

Studies on involvement [2], [4], or related concepts such as emotional engagement [5] [6], interest [7], or interactional rapport [8] reported that these phenomena are conveyed by specific prosodic cues. For example, Wrede and Shriberg [2], in their study on involvement found that there was an increase in mean and range of the fundamental frequency (F0) in more activated speech as well as tense voice quality. Moreover, Crystal and Davy [9] reported that, in live cricket commentaries, the more the commentator is involved in reporting the action (i.e. at the action peak), the quicker the speech rate.

2 Main Objectives & Hypotheses

In our study we looked at how prosodic parameters as well as visual cues may be used to indicate levels of involvement. A statistical model based on these cues would enable the automatisation of involvement detection. Automatic involvement detection allows for a time efficient search through multimodal corpora, and may be used for interactive speech synthesis.

The prosodic parameters (i.e. F0, duration and intensity) include level and span of the voice, articulation rate (i.e. excluding pauses) and intensity of the voice. The visual parameter includes the participants' amount of change in movement of the body and face.

Based on studies [2–9] our hypotheses are: the higher the degree of involvement, (1) the higher the level and (2) the wider the span of the voice, (3) the quicker the articulation rate, (4) the higher the intensity and (5) the higher the amount of movement in the face and body of the participants.

3 Experiment

3.1 Data Collection: The D64 corpus

We used the D64 corpus [10] for this study. It was recorded over two successive days in a rented apartment, resulting in a total of eight hours of multimodal recordings. Five participants took part on the first day and four on the second. Three of the participants were male and two female. They were colleagues and/or friends (with the exception of one naive participant), ranging in age from early twenties to early sixties. They were able to move freely around as well as to eat and drink refreshments as in normal daily life. The conversation was not directed and ranged widely over topics both trivial and technical.

3.2 Data selection

For our analysis, all 5 speakers were included. Data was chosen from two different recording sessions; Session 1 and Session 2 (a total of 1 hour of recording). For session 1, there was no predefined topic, and the conversation was allowed to meander freely. For session 2, the first author's Master's research was amongst the topics of discussion. Speaking time per speaker varies between 1 to 15 minutes (mean=9 min; sd=5,15).

3.3 Data Annotation

We developed an annotation scheme based on hearer independent, intuitive impressions [11] and annotated approximately 1 hour of video recordings for levels of involvement. The annotation scheme was validated perceptively and was combined with acoustic analysis and movement data. Our measure of involvement comprises the joint involvement of the entire group.

Involvement annotations are based on the following criteria: Involvement level 1 is reserved for cases in which virtually no interaction is taking place and in which interlocutors are not taking notice of each other at all and are engaged in completely different pursuits. Involvement level 2 is a less extreme variant of involvement level 1. Involvement level 3 is annotated when subgroups emerge. For example, in a conversation with four participants, this would mean that two subgroups of two interlocutors each would be talking about different subjects and ignore the respective other subgroup. Involvement level 4 is annotated when only one conversation is taking place while for involvement level 5 interlocutors also need to show mild interest in the conversation. Involvement level 6 is annotated when conditions for involvement level 5 are fulfilled and interlocutors encourage the turnholder to carry on. Involvement level 7 is annotated when interlocutors show increased interest and actively contribute to the conversation. For involvement level 8, interlocutors must fulfil the conditions for involvement level 7 and contribute even more actively to the conversation. They might for example jointly, wholeheartedly laugh or totally freeze following a remark of one of the participants. Involvement level 9 is annotated when interlocutors show absolute, undivided interest in the conversation and each other and vehemently emphasise the points they want to make. Participants signal that they either strongly agree or disagree with the turn-holder. Involvement level 10 is an extreme variant of involvement level 9.

A ten point scale was chosen for annotation but only values 4-9 were actually used in the annotations.This fact might be explained by the calm and friendly nature of the conversation. The numbers of times in which involvement level 4 and 9 were annotated were statistically not sufficient and were thus excluded from further analysis.

3.4 Measurements & Statistical Analyses

Acoustic measurements were obtained using the software Praat [12]. The level and span of the voice were measured by calculating the F0 median (the mean

being too sensitive to erroneous values) and the $log_2(F0max - F0min)$ respectively. The F0-level is given on a linear scale (i.e. Hertz) while F0-span is given on a logarithmic scale (i.e. octave). In order to avoid possible pitch tracking errors, pitch floor and pitch ceiling were set to the values $q_{15} \cdot 0.83$ (where 'q' stands for percentile) and $q_{65} \cdot 1.92$ (De Looze [13]). Articulation rate was calculated in terms of number of syllables per second. Syllables were detected automatically using a prominence detection tool developed by Tamburini [14]. In order to neutralise speaker differences in voice level and span, articulation rate and intensity, data were normalised by a z-score transformation.

For the movement extraction an algorithm was chosen which is not restricted to calculating movement changes for the whole picture but rather for individual people (note that movement measurements were only calculated in this study for two speakers). From the video data coordinates of the faces and bodies at each frame composed by the exact spot of the top left corner and the bottom right corner of the face are extracted as in Scherer et al [15] by utilising the standard Viola Jones algorithm [16]. Normalisation is carried out as these coordinates are highly dependent on the distance of the person to the camera in order to obtain relative movement over the size of the detected face and body. Only in the case where a face is recognised a moving average is calculated. ANOVA analyses were carried out for the above mentioned cues.

3.5 Results

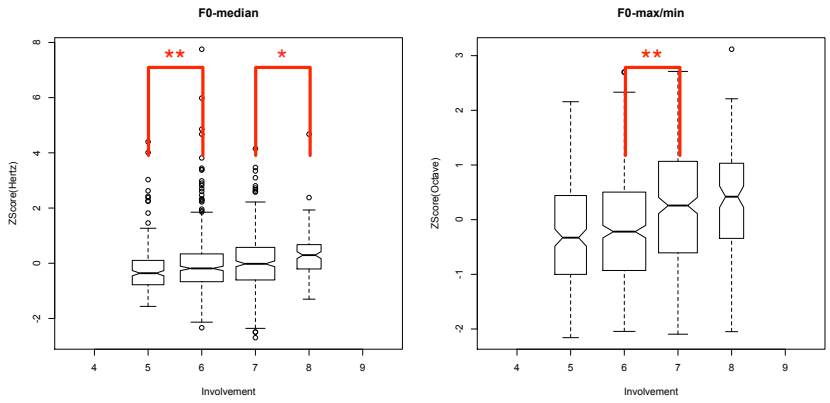

Fig. 1. Boxplots of F0-median and F0 max/min according to four levels of involvement.

Level and Span of the voice As illustrated in Figure 1, involvement level 6 is significantly higher than involvement level 5 ($F(3,1041)=8.843$; $p=0.006370$) and involvement level 8 is significantly higher than involvement level 7

(F(1,440)=6.58; p=0.0106). Involvement level 7 is however not significantly higher than 6 (F(2,830)=4.899; p=0.35040).

The acoustic cue F0-max/min as illustrated in Figure 1 increases with involvement. While involvement level 7 is significantly higher than involvement level 6 (F(2,831)=22.82; p=7.96e-08), involvement level 6 is not significantly higher than involvement level 5 (F(3,1041)=18.31; p=0.6325) and involvement level 8 is not significantly higher than involvement level 7 (F(1,440)=21.2; p=0.274).

Articulation Rate The acoustic cue articulation rate does not illustrate any significant changes. The articulation rate of the individual speakers stays approximately the same over the various involvement levels.

Intensity The acoustic cue intensity illustrates an increasing slope as can be seen in Figure 2. While involvement level 6 is significantly higher than involvement level 5 (F(3,1130)= 139.5 ; p=1.62e-05) and involvement level 7 is significantly higher than involvement level 6 (F(2,889)= 121; p=<2e-16), involvement level 8 is not significantly different from involvement level 7 (F(1,453)=0.223 ; p=0.637).

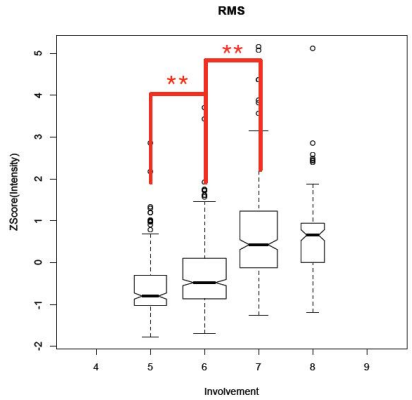

Fig. 2. Boxplots of intensity according to four levels of involvement.

Movement For movement of body and face, it can be seen in Figure 3 that for speaker F there is an increasing slope. Involvement level 6 is significantly higher than involvement level 5 (F(3,611)=11.67; p=0.00484). However, involvement level 7 is not significantly different from involvement level 6 (F(2,464)=5.617; p=0.04029) and involvement level 8 is neither significantly different from involvement level 7(F(1,240)=2.152; p=0.144).

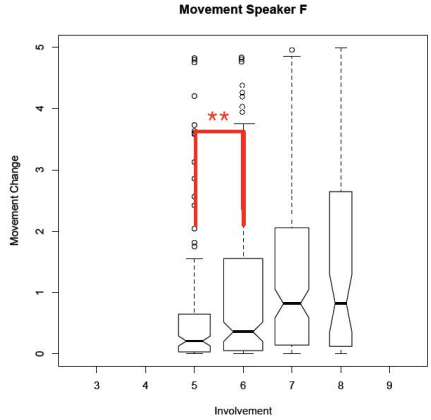

Fig. 3. Boxplots of Movement for speaker F according to four levels of involvement.

For Movement of Body and Face for speaker C we present the results separately for session 1 and 2 since the trends of both sessions are significantly different here. It can be seen in Figure 4 that there is an increasing slope for session 1. Except for the increase from involvement level 6 to involvement level 7 ($F(2,192)=3.913$; $p=0.00753$), this increase is however not significant for other levels. It can be seen in Figure 4 that there is a decreasing slope for session 2 in involvement level 6, 7 and 8 where the decrease from involvement level 6 to 7 is significant ($F(2,269)=7.497$; $p=0.000256$).

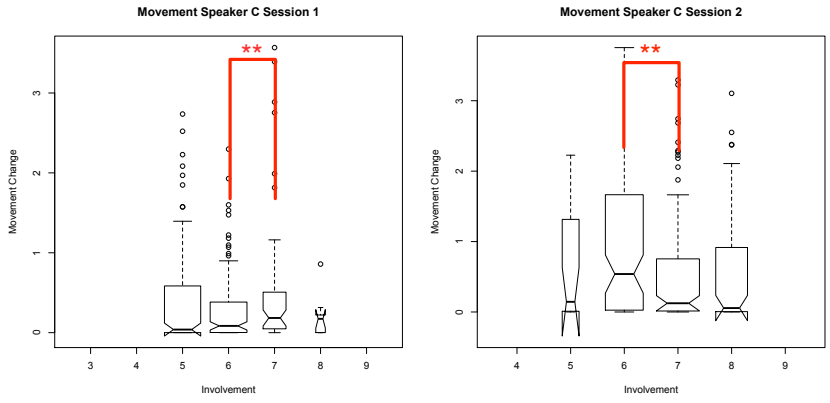

Fig. 4. Boxplots of Movement for speaker C in Session 1 and 2 according to four levels of involvement.

4 Discussion

In this study, we examined the prosodic parameters correlated with the dynamic changes that characterise social conversation. We looked at the level and span of the voice, intensity and articulation rate. We confirmed the findings of Wrede and Shriberg (2003) that the level and span of the voice, as well as the intensity, increase in more activated speech. Contrary to their binary distinction however, our data suggests that involvement seems to be a scalar rather than binary phenomenon.

We found a clear linear relationship between our perceptual measure of involvement and the level and span of the voice as well as intensity. Wrede and Shriberg make no mention of articulation rate. We looked at articulation rate and found no relationship.

In a pilot study we added a multimodal aspect to our analysis through consideration of automatically extracted measures of body and head movement. We found this parameter to be well correlated with our perceptual measures of involvement. Our movement analysis (based on two speakers) for speaker F indicates that the more involvement the more the amount of movement. For speaker C however, movement in session 2 does not fit this pattern. This may be explained perhaps by the fact that speaker C held a laptop on her lap in session 2, hiding her hands for part of the session.

Our analyses based on level and span of the voice and intensity suggest that involvement is a scalar phenomenon. Furthermore, the preliminary measures of movement appear to correlate strongly with the acoustic parameters and so it might be advantageous to merge them to give a more robust automatic measure of involvement. Further analysis will be carried out to confirm our preliminary results. Our current and future work involves building a statistical model, incorporating both sources of information in order that we may clarify the mutual information. The number of levels to quantify involvement is not clear at the moment, but we will continue to use a scale of one to ten.

5 Conclusion

Our study confirmed that social information is expressed to a high degree through prosodic cues and movement of the body and face. The aim of this paper was to use those cues to make one aspect of social information more tangible; namely participants' degree of involvement in a conversation. Our results for voice span and intensity, and our preliminary results on the movement of the body and face suggest that these cues are reliable cues for the detection of distinct levels of participants involvement in conversation. This will allow for the development of a statistical model which is able to classify these stages of involvement. This would have applications in automatic multimodal corpus search, automatic spoken dialog systems, robotics, games and other such technologies.

References

1. Antil, J.H.: Conceptualization and Operationalization of Involvement. Advances in Consumer Research. 11(1), 203–209 (1984).
2. Wrede, B., Shriberg, E.: Spotting Hot Spots in Meetings: Human Judgements and Prosodic Cues. In: Proceedings of Eurospeech 2003, pp.2805-2808. Geneva (2003).
3. Dillon, R.: In: Lecture Notes in Computer Science: A Possible Model for Predicting Listener's Emotional Engagement. Springer Press. Heidelberg (2006).
4. Selting, M.: Emphatic speech style: with special focus on the prosodic signalling of heightened emotive involvement in conversation. Journal of pragmatics. 22(3–4), 375–408 (1994).
5. Gustafson, J., Neiberg, D.: Prosodic cues to engagement in non- lexical response tokens in Swedish. In: DiSS-LPSS Joint Workshop 2010. Tokyo, Japan (2010).
6. Yu, C., Aoki, P.M., Woodruff, A.: Detecting user engagement in everyday conversations. In: 8th International Conference on Spoken Language Processing (ICSLP ' 04), pp.1329-1332. Jeju Island, Korea (2004).
7. Gatica-Perez,D.: Modeling Interest in Face-to-Face Conversations from Multimodal Nonverbal Behavior. In J.-P. Thiran, H. Bourlard, and F. Marques, (Eds.), Multimodal Signal Processing. pp.309–323, Academic Press. San Diego, USA (2009).
8. Duncan, S., Baldenebro, T., Lawandow, A., Levow, G.-A. : Multi-modal Analysis of Interactional Rapport in Three LanguageCultural Groups. In: Workshop on Modeling Human Communication Dynamics, pp.42–45 Vancouver, B.C., Canada (2010).
9. Crystal, D., Davy, D.: Investigating English Style. Longman Group. Ltd., London (1969).
10. Oertel, C., Cummins, F., Campbell, N., Edlund, J., Wagner, P.: D64: a corpus of richly recorded conversational interaction. In: Proceedings of LREC 2010; Workshop on multimodal corpora: advances in capturing, coding and analyzing multimodality, pp.27–30. Valetta (2010).
11. Oertel, C.: Identification of Cues for the Automatic Detection of Hotspots. Bielefeld University (unpublished). Bielefeld(2010).
12. Boersma, P., Weenink, D.: Praat: doing phonetics by computer.
13. De Looze, C., Hirst, D.J.: Integrating changes of register into automatic intonation analysis. In: Proceedings of the Speech Prosody 2010 Conferene, 4 pages. Chicago (2010).
14. Tamburini, F., Wagner, P.: On automatic prominence detection for german. In: Proceedings of Interspeech 2007 pp. 1809-1802. Antwerp (2007).
15. Scherer, S., Campbell, N.: Multimodal laughter detection in natural discourses. In: Proceedings of the 3rd international workshop on human- centered robotic systems (HCRS09), pp.111–121. (2009).
16. Viola, P., Jones, M.J.: Robust real-time face detection. International Journal of Computer Vision. 57(2), 137–154 (2004).