Who's the Boss? : From Social Psychology to Social Signal Processing

Hayley Hung [*]

[*] IET Write Around the World Competition 2009

1 The First Day at Work

Imagine it is your first day at work. You're the new guy or girl. You don't really know how things work but you want to make a good impression, fit in and not offend anyone on your first day. You go to the first meeting of the group you are working with. You haven't met anyone yet and your arrival causes some interest. On the one hand, you want to show competence and that you are good at the job. Demonstrating confidence and expertise can sometimes mean that we need to openly disagree with others, which can mean that someone becomes undermined. There is also the possibility that there will be someone in the room that we do not want to undermine under any circumstances. This sense that all of us have to know when to assert our opinion and when not to, specifies our natural understanding of the pecking order in a group. Failure to judge this at the very start of an encounter could lead to bad impressions and possibily colleagues who are more wary to make us feel welcome. It is said that each of us does not require much time at all to get an idea of these social dynamics [1]. I'm sure this is a common scenario for many of us but what drives us to behave the way we do during this first encounter? That natural order of social interactions comes easily to us. This social intelligence [24] regulates our civilised society and governs the way we behave to get ahead but also to get along.

2 Leaders, Managers and Dominators

Social psychology research has identified a vertical dimension to social groups which relate to power, dominance, status hierarchy and other related concepts which serve to separate members in a group in a vertical fashion [11]. Dominance can be thought of as a personality trait and the desire to control others; status is an ascribed or achieved quality which implies respect and privilege but does not necessarily include an ability to control others or their resources; and power is the given right of a person to control others or their resources but does not necessarily imply that they have prestige or respect. Out of the three constructs, dominance is the one which is the easiest to measure when we have no previous information about a group. We concentrate here, on discussing dominance as a behavioural trait.

Over several decades, social psychologists have tried to characterise dominant behaviour in face to face discussions. Dunbar and Burgoon argue that perceived dominance is a set of "expressive, relationally based communicative acts by which power is exerted and influence achieved" [8] (p208). They also suggested that that while power and status are properties that exist through a long-term establishment of hierarchy, dominance is viewed as "necessarily manifest. It refers to context and relationship-dependent interactional patterns in which one actors assertion of control is met by acquiescence from another" (p.208) [8]. This idea of assertion and acquiescence was suggested previously by Rogers-Millar and Millar [23] who defined domineeringness and dominance as two separate control variables where domineeringness was the proportion of 'one-up' manoeuvres a person performs during a conversational interaction while dominance is the ratio of 'one-up' to 'one-down' manoeuvres. This idea of assertion and acquiescence was also suggested by Dovidio and Ellyson who defined a visual dominance ratio [6] to infer the level of dominance of two individuals. This was based on the ratio of the proportion of time someone spent addressing the other person divided by the time they spent looking and listening to the other.

Dunbar and Burgoon [8] quantified the effect of different non-verbal cues on a person's perceived dominance levels. These cues were categorised as vocalic and kinesic features, referring to speech (e.g. speaking time, loudness or energy, speaking rate, pitch vocal control or interruptions [25]) and gesture based cues (e.g. body movement, posture and elevation, facial expressions, gestures or eye gaze [6]) respectively. One that was consistently useful for inferring dominance was the total speaking time. This may seem surprising but is backed by many years of research across different groups worldwide. Marianne Schmid-Mast conducted a meta-analysis of 40 articles containing 45 studies in social psychology performed over 5 decades, concluding that dominance could be inferred through speaking time [17]. She found that dominance was expressed through speaking time more in role-based dominance scenarios (e.g. manager/employee or teacher/student) than where dominant personality traits were observed.

3 Socially Intelligent Computational Modelling

In recent years, the science and engineering research community has started to address whether social intelligence can be trained in machines [24, 19]. Historically, this has been of interest in the domain of affective computing where there is a strong emphasis on making interactions between humans and computers more "human-friendly". However building socially intelligent machines can have much more far reaching applications such as surveillance and crime prevention where one may want to predict if a fight is about to happen, for data-mining tasks or even automatic video editing of home movies. In the future, we may not be holding camcorders but allowing robots to film us and then edit the videos for us according to the high emotion points or the important people in our lives. Thus begins the birth of a new domain of research known as social signal processing[24].

Social intelligence can be viewed largely as a multi-modal problem. That is to say that signals in both speech and body language of people are likely to be relevant in some way to the social signals given and therefore computational modelling must take this into account. Figure 1 a socially intelligent system might work. Indeed this model is nothing new and can be applied to many other tasks other than social signal processing. However, what this shows is that while social signal processing may deal with less objective classification tasks, the model for testing, building and evaluating such a system remains the same as any other type of computational modelling. However, while estimating the position of a plane uses radar and therefore the laws of physics to determine where the plane is, for social signal processing, we are governed by the laws of social interactions and the largely stable systems that they inhabit.

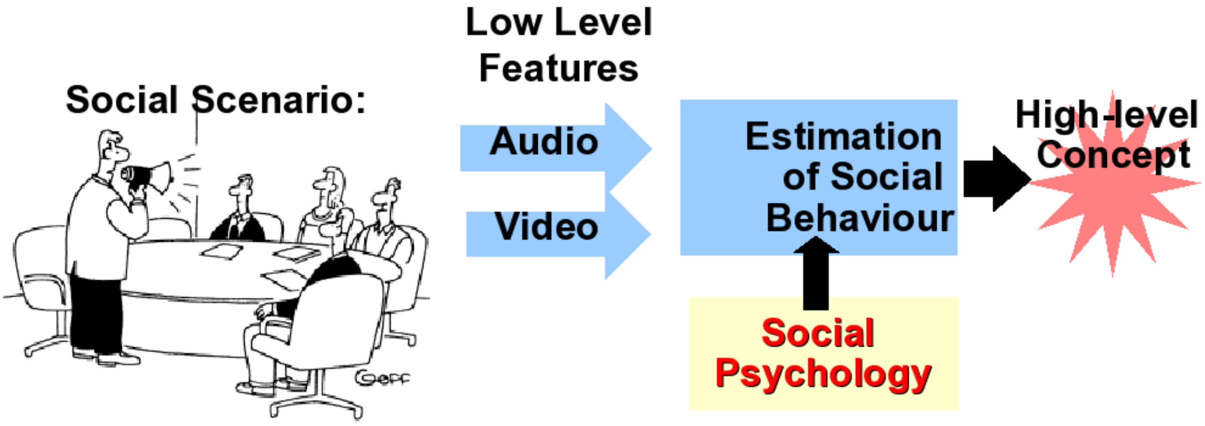


Figure 1: Overview of how computational modelling of social signals can work.

4 Automatic Detection and Modelling of Dominance

If you were to ask anyone on the street what dominance was, it is likely that there would be many answers from different people. Some would say it is to do with how authoritative the person is, their tone of voice, or whether they are leaders or followers. What is clear is that everyone has their own definition and sometimes these may contradict each other. Under such circumstances, one may conclude that trying to modelling something which even people cannot agree on is an ill-posed problem. However, the findings from recent research into computational models of dominance indicate otherwise and show that computational models could provide more efficient and consistent ways of estimating human behaviour types.

Typically, work on modelling dominance and influence has used data of meeting scenarios or discussions in a seated environment. One of the first works that was relevant in this area was on analysing influence in small group debates of 5 people [2]. Figure 2 shows an example of the resultant influence graph between participants in one of their debates where we see that Tammy and Anne were the dominating speakers. They modelled turn exchanges to represent the transitions between different states in the conversation such as who is most likely to speak after who. Their analysis was on just two different group debates and used a combination of cues extracted from automatically and manually both audio and video sources.

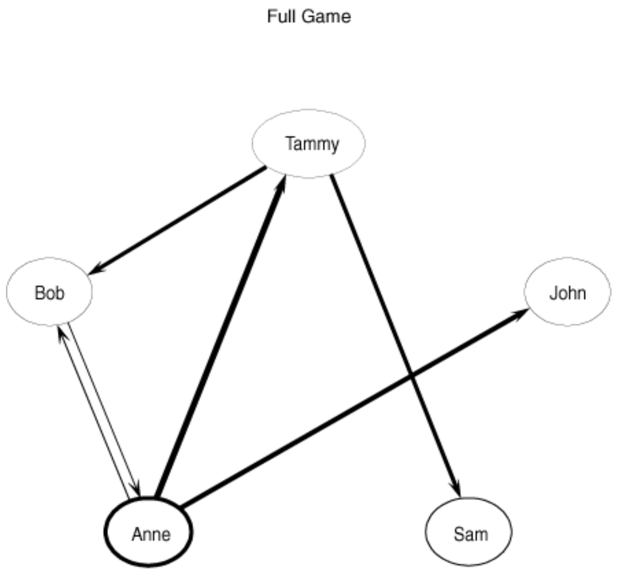


Figure 2: Influence Graph for full game showing strong links for the dominating speakers Tammy and Anne. Anne is circled in bold as the most dominating speaker and the thickness of the arrows shows how frequently each combination of dyads communicated with each other directly. Reproduced from [2]

Rienks et al. provided more quantitative analysis of the performance of different speech-based features on estimating influence as well as different methods of fusing the features together [22]. In addition to testing models that could model the temporal changes in group behaviour, they also tried cues that gave a static or single feature vector for each meeting that was considered. All the features they used were extracted manually from meetings and included cues such as the number of speaker turns, number of successful interruptions, number of times a person is interrupted, or the number of topics initialised by a particular person in the meeting. They estimated influence levels in terms of 'high', 'medium' and 'low' levels with good performance.

Otsuka et al. [18] was the first to use a fully automatic system for analysing influence in meetings. They concentrate on using non-verbal cues based on automatically extracted gaze patterns, to explain pair-wise influence in group discussions. The gaze of a person can be determined, to some extent by the orientation of their head. There is still a significant field of view from a single head orientation because we can also move our eyes while our head stays still. Depending on the types of conversations or conversational context that was occurring e.g. if there was a dialogue or a monologue. Cognitively speaking, people tend to look towards the speaker during a group discussion but if there is more than one person speaking, they tend to position their head relative to both speakers of interest. Why is gaze of interest in estimating influence? It is known that people with high status tend to receive more visual attention from

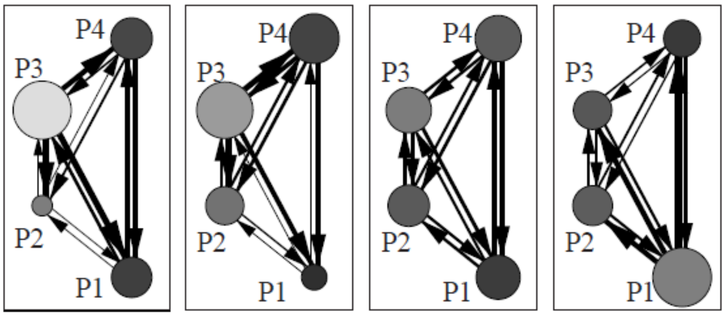


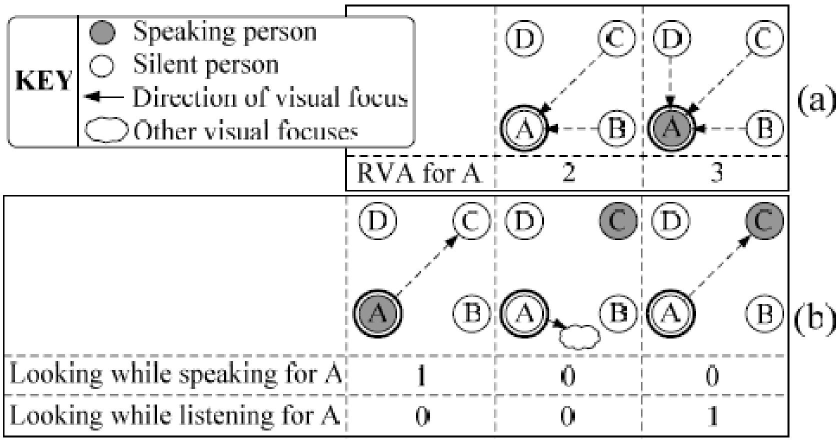
Figure 3: Influence Graphs for 4 discussions analysed by Otsuka et al.. Each circle represents one of the participants and the size of the circle indicates the amount of outgoing influence. The the thickness of each arrow indicates the direction and difference between the outgoing and incoming influence of each person. Reproduced from [18]

others [6]. Otsuka et al. used their model of gaze behaviour to estimate who would be looking at who during a conversation. By combining these cues together, they devised measures of incoming and outgoing influence for each person in the discussion so that some analysis of the meeting interactions could be observed. Similar to the work of Basu et al., they devised figures (see Figure 3) relating the influence between participants in the discussion The main assumption is that if a person is speaking and others are looking at him, this person is influencing them.

It was not until later that there was a real drive to create both real working fully automatic systems that could estimate dominance and also evaluate the effectiveness of different cues in meetings. Jayagopi et al. [14] created a detailed study to observe how audio and video cues and also audio-visual cue fusion could help in the estimation of dominance. The hypotheses used were related to dominant people tending to gesture more when speaking, talking more, interrupting more [7, 3]. Both the audio and video cues that were extracted were extremely simple and included features such as the speaking length, number of interruptions, total number of speaking turns and a histogram of the turn durations for each participant. Equivalent features were extracted from video activity features. The estimation tasks that were tested included estimating the most dominant person in a meeting and the least dominant person. Overall, they found that audio cues has superior performance compared to video cues alone, with speaking length giving superior performance as a single cue. Feature fusion was performed using support vector machines where it was found that the estimations of dominance were significantly higher. Interestingly, fusion of audio-only cues led to comparable results to fusing the audio and video modalities together. Estimating the least dominant person was shown to be more difficult because submissive people can tend to speak and move a lot less. They also found that when human judgements were completely unanimous about who the most or least dominant person was, the computational models were also more accurate than when only majority agreement among annotators was available.

The disadvantage of the work of Jayagopi et al. was that while all the features were extracted automatically, the audio sensors came from headset microphones which are uncomfortable to wear and not practical to use. In an ideal situation, we would probably want to take a single microphone, put it in the centre of meeting room and then just press 'start'. Hung et al. [12] investigated how the dominant person could be estimated from a single microphone, showing that despite a low signal to noise ratio from the single audio sensor, estimating the most dominant person in a meeting. One thing that ought to be mentioned about this scenario is also that since there is only one microphone in the room and many speakers, part of the problem is to estimate who is speaking and when. This problem is known as 'speaker diarization' in the speech research community [21].

Most recently, Hung et al. [13] tried to address the problem of why the video features in the work of Jayagopi et al. did not perform well for estimating dominance in meetings. One possible answer to this is that the features that were used were too simple and using features closer to Otsuka's gaze model might give more robust estimates. The work directly addressed Dunbar and Burgoon's [6] measure of dominance as an audio-visual phenomena in terms of the visual dominance ratio (Figure 4). Another measure of visual dominance which depends only on visual features would be to accumulate the total amount of received visual attention that each person in a discussion receives. This was defined as the ratio of the looking while speaking to looking while listening time in two-person conversations. For the case of four-person meetings, the scenario changes and so the looking while listening measure applies only when looking to the speaker. Using automatic estimates of visual focus of attention and speaking status, it was possible to estimate the most dominant person, but unfortunately not the same level of performance as the audio features explored by Jayagopi et al. [14].



RVA for A	2	3

Looking while speaking for A	1	0	0
Looking while listening for A	0	0	1

Figure 4: Example scorings of the total received visual attention (RVA) and visual dominance ratio for person A (highlighted node), at time t: (a) two examples of *RVA*; (b) three example scenarios for looking-while-speaking and looking-while-listening. Reproduced from [13]

5 Towards a complete model of social intelligence

Of course, dominance is only one of many behavioural constructs. In terms of social signal processing, work has been carried out to estimate a person's personality [20], role [5, 26, 10], status [15], and even if they are telling the truth [9], to name but a few. Other constructs that have not been approached so much in computational modelling relates to leadership and how leaders are formed.

In terms of social signal processin in general, perhaps it is important to highlight two points. The first is that the future of socially intelligent machines goes beyond human computing in the sense of having human friendly interactions with machines. As our world becomes more cluttered with increasing amounts of recorded audio and video data, automatic means of sorting them according to our own emotional or social responses to situations may help us to find data that is important to us more efficiently than the usual keyword-based methods which are more readily available to us today.

Secondly, it must not be ignored that machines are different from humans but this does not make them less able to make judgements about the physical world. Human performance is certainly a good benchmark but could still be surpassed. A good example of this is work by Littlewort et al. [16], who investigated whether it was possible to detect automatically if people's facial expressions showed real or fake pain (Figure 5. Their experiments showed that using their computational models, they were able to significantly improve on human judgements of whether the expressions were fake or not. This particular example asks a fundamental question about socially aware computing since it becomes increasingly plausible that machines can out-perform humans.

Ethically speaking, do we always want to know the truth? When someone praises us, do we want to know whether it is genuine? For example, if you invite a friend over for dinner and you know you burnt the food but they tell you the food was delicious. So your friend has lied to you but only to spare your feelings so we could say this lie is necessary to keep the friendship. According to Professor Judith Donath [4] this is certainly true and states that all social signals need to have a level of ambiguity to them since if we were all honest with each other, civil society would break down. Perhaps the key to socially aware computing in future is the skill to know what the truth is in order to bend it...

References

[1] N. Ambady and H. M. Gray. On being sad and mistaken: Mood effects on the accuracy of thin-slice judgments. *Journal of Personality & Social Psychology*, 83(4):947–961, 2002.

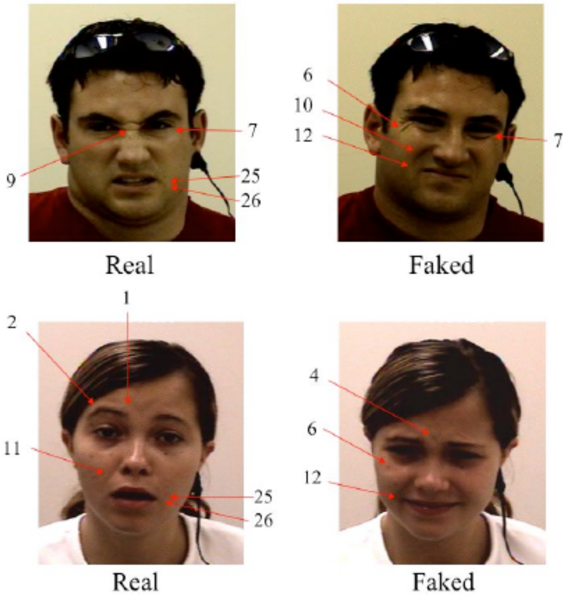


Figure 5: Examples of real and fake pain when subjects were asked to keep one hand in either water that was very cold or at room temperature. Reproduced from [16]

[2] S. Basu, T. Choudhury, B. Clarkson, and A. Pentland. Towards measuring human interactions in conversational settings. In *Proc. IEEE CVPR Int. Workshop on Cues in Communication (CVPR-CUES)*, Kauai, Dec. 2001.

[3] J. K. Burgoon and N. E. Dunbar. Nonverbal expressions of dominance and power in human relationships. In V. Manusov and M. Patterson, editors, *The Sage Handbook of Nonverbal Communication*. Sage, 2006.

[4] Judith Donath. *Signals, Truth and Design*. Forthcoming,MIT Press.

[5] Wen Dong, Bruno Lepri, Alessandro Cappelletti, Alex Pentland, Fabio Pianesi, and Massimo Zancanaro. Using the influence model to recognize functional roles in meetings. In *ICMI '07: Proceedings of the 9th international conference on Multimodal interfaces*, November 2007.

[6] J. F. Dovidio and S. L. Ellyson. Decoding visual dominance: Attributions of power based on relative percentages of looking while speaking and looking while listening. *Social Psychology Quarterly*, 45(2):106–113, June 1982.

[7] N. E. Dunbar and J. K. Burgoon. Measuring nonverbal dominance. In V. Manusov, editor, *The sourcebook of nonverbal measures: Going beyond words*. Erlbaum, 2005.

[8] N. E. Dunbar and J. K. Burgoon. Perceptions of power and interactional dominance in interpersonal relationships. *Journal of Social and Personal Relationships*, 22(2):207–233, 2005.

[9] F. Enos, E. Shriberg, M. Graciarena, J. Hirschberg, and A. Stolcke. Detecting deception using critical segments. Interspeech, 2007.

[10] Sarah Favre, Hugues Salamin, John Dines, and Alessandro Vinciarelli. Role recognition in multiparty recordings using social affiliation networks and discrete distributions. In *IMCI '08: Proceedings of the 10th international conference on Multimodal interfaces*, pages 29–36, New York, NY, USA, 2008. ACM.

[11] J.A. Hall, E.J. Coats, and L.S. LeBeau. Nonverbal Behavior and the Vertical Dimension of Social Relations: A Meta-Analysis. *Psychological bulletin*, 131(6):27, 2005.

[12] Hayley Hung, Yan Huang, Gerald Friedland, and Daniel Gatica-Perez. Estimating the dominant person in multi-party conversations using speaker diarization strategies. In *International Conference on Acoustics, Speech and Signal Processing*, 2008.

[13] Hayley Hung, Dinesh Babu Jayagopi, Silèye O. Ba, Jean-Marc Odobez, and Daniel Gatica-Perez. Investigating automatic dominance estimation in groups from visual attention and speaking activity. In *International Conference on Multi-modal Interfaces*, 2008.

[14] D. Jayagopi, H. Hung, C. Yeo, and D. GaticaPerez. Modeling dominance in group conversations from non-verbal activity cues. *Special issue on Multimedia in IEEE Transactions on Audio, Speech and Language Processing*.

[15] Dinesh Babu Jayagopi, Sileye Ba, Jean-Marc Odobez, and Daniel Gatica-Perez. Predicting two facets of social verticality in meetings from five-minute time slices and nonverbal cues. In *IMCI '08: Proceedings of the 10th international conference on Multimodal interfaces*, pages 45–52, New York, NY, USA, 2008. ACM.

[16] Gwen C. Littlewort, Marian Stewart Bartlett, and Kang Lee. Faces of pain: automated measurement of spontaneousallfacial expressions of genuine and posed pain. In *ICMI '07: Proceedings of the 9th international conference on Multimodal interfaces*, pages 15–21, New York, NY, USA, 2007. ACM.

[17] Marianne Schmid Mast. Dominance as expressed and inferred through speaking time. *Human Communication Research*, (3):420–450, July 2002.

[18] K. Otsuka, J. Yamato, Y. Takemae, and H. Murase. Quantifying interpersonal influence in face-to-face conversations based on visual attention patterns. In *Proc. ACM CHI Extended Abstract*, Montreal, Apr. 2006.

[19] A.S. Pentland. *Honest signals: how they shape our world.* The MIT Press, 2008.

[20] Fabio Pianesi, Nadia Mana, Alessandro Cappelletti, Bruno Lepri, and Massimo Zancanaro. Multimodal recognition of personality traits in social interactions. In *IMCI '08: Proceedings of the 10th international conference on Multimodal interfaces*, pages 53–60, New York, NY, USA, 2008.

[21] D. A. Reynolds and P. Torres-Carrasquillo. Approaches and applications of audio diarization. In *Proc. of International Conference on Audio and Speech Signal Processing*, 2005.

[22] Rutger Rienks, Dong Zhang, Daniel Gatica-Perez, and Wilfried Post. Detection and application of influence rankings in small group meetings. In *Proceedings of the 8th International Conference on Multimodal interfaces*. ACM Press, 2006.

[23] E. Rogers-Millar and F. Millar III. Domineeringness and dominance: A transactional view. *Human Communication Research*, 5(3):238–246, 1979.

[24] A. Vinciarelli, M. Pantic, and H. Bourlard. Social signal processing: Survey of an emerging domain. *Image and Vision Computing*, December 2008.

[25] C. West and D. H. Zimmerman. *Language, Gender, and Society*, chapter Small Insults: A study of interruptions in cross-sex conversations between unaquainted persons, pages 103–117. Newbury House, 1983.

[26] M. Zancanaro, B. Lepri, and F. Pianesi. Automatic detection of group functional roles in face to face interactions. In *in Proc. Int. Conf. on Multimodal Interfaces (ICMI)*, Banff, Nov. 2006.