Identifying Dominant People in Meetings from Audio-Visual Sensors

Anonymous FG2008 submission

Abstract

This paper provides an review of the relatively new area of automated dominance estimation in group meetings. We describe both research in social pyschology and use this to explain the motivations behind suggested automated systems. We use an overview of our own to motivate discussions for future work in this area.

1. Introduction

Human group behaviour is a complex and highly dynamic, time-varying process which defines our role and identity in a group through social interactions. An initial face-to-face encounter between unacquainted individuals commences immediately with an establishment of hierachy between the interactants [15]. These initial encounters can take the form of subtle non-verbal communication through eye-gaze with other participants. Such findings show that establishing hierachy occurs only during participation in social interactions and is an innate part of human behaviour.

Understanding the group dynamics in conversational settings allows us to analyse the effectiveness of teams or as a search query for browsing meeting data. There is a real need for automated systems for estimating dominance and this has led to an emergence of research which tries to cross the divide between social pyschology, and machine learning and artificial intelligence. This paper summarises and organises this interesting and challenging area of research and studies automated some solutions. A study of the major problems and solutions in this field are discussed in more detail through an overview of our own work in this area. More specifically, we highlight different solutions using multi-modal cues and provide preliminary results when more semantically meaningful labels which are more challengin to extract of human behaviour are considered.

In the remainder of this paper, Section 2 provides a summary of investigations in social pyschology on defining and understanding dominant behaviour; Section 3 describes work in the area of automated dominance analysis; Section 4 describes an overview of our contributions to automated dominance modelling using audio, visual and audio-visual measures of activity; Section 5 discusses some remaining challenges through the overview and preliminary study we provided and we conclude in Section 6.

2. What Is Social Dominance?

Dominance has been studied in social pyschology for many years where psychologists have tried to define dominance or find indications of it. Dominance can be viewed as a personality characteristic or a persons status within a group or the power they have within a group [9]. However, [5] suggested that power and dominance were not the same. They suggest that power is the “capacity to produce intended effects, and in particular, the ability to influence the behavior of another person...Because power is an ability...it is not always exercised...its magnitude may not be fully evident unless it is pitted gainst a counterforce of appropriate strength” (p 208). On the other hand, “dominance is 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). This definition of dominance was defined by Rogers-Millar and Millar [14] who defined domineeringness and dominance as two separate control variables where domineeringness was the proportion of ‘one-up’ maneouvers a person performs during a conversational interaction or the ratio of ‘one-up’ to ‘one-down’ maneouvers.

Similar findings have been made by Dovidio and Ellyson who defined a visual dominance ratio [4] to decode dominance. 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.

More recently, studies have been conducted to quantify the effect on different facets of non-verbal on perceived dominance levels of an individual based on their activity. Schmid Mast found through a meta-analysis of literature spanning several decades, that dominance could be inferred and expressed through speaking time [9]. Later, [5] conducted a study into decoding dominance through non-verbal cues which they categorised as vocalic and kinesic features which referred to speech (e.g. speaking time, loudness or energy, speaking rate, pitch vocal control [5] or interruptions [17]) and gesture based cues (e.g. body movement, posture and evelation, factial expressions, gestures [5] or eye gaze [4]) respectively.

In terms of decoding dominance, social pyschology literature has shown that it is possible to do this either as an participant or observer of the interaction, though there may be differences in perception [5]. For example Dovideo et al. [4] that people could decode vidual dominance displayed by others. This particularly relevant to automated dominance analysis for manual (first or third party) annotations are required for ground truth generation. An interesting comment by Dunbar and Burgoon was that "Perhaps coders' perception of dominance correspond more closely with objective measures of verbal and non-verbal dominance than those of participants themselves... However, the coders' observations are limited to the behaviors in a particular interaction, whereas participants are privy to the ongoing interaction that is part of a continuing relationship." [5] (pp. 228). More details on understanding dominance from a social pyschology perspective can be found in [5, 2].

3. Automated Dominance Estimation

Basu et al. [1] were the first to investigate influence in group a discussion scenario. Their approach treated each relation between individuals on a dyadic basis and modelled all group interactions in terms of Markov chains where the transitions were affected by the influence that one participant could exert on another. In each discussion, two out of five participants were requested to debate on a pre-specified topic for one minute before the floor was open to everyone. In this work, a combination of manually and automatically extracted audio-visual features were extracted such as speaking status, turns, and visual activity patterns from skin-colour blob-tracking. They quantified dyadic influence of person A on B as the number of times that B responds to A and used this for qualitative evaluation of the Influence Model (IM) on the same conversational data.

Later in 2002, Ohsawa et al. [10] presented a method of describing influence in terms of an influence diffusion model in text-based communication. The assumption was that influence could be inferred by the diffusion of key words through communication chains. It was not until 2005 that the idea of influence modelling was addressed again, starting with the team-player influence model (TPIM) proposed by Zhang et al. [18]. Here, only audio cues were used which included using manually annotated transcripts of the meetings for automatic topic analysis. Automatically extracted speaking activity features from a microphone array and headset microphones were also used. Their meeting corpus had 150 minutes of conversations where there were pre-defined discussion topics and an action agenda, to encourage discussions and monologues. The TPIM represents explicitly the states of the group and its influence on the state transitions of individuals using a two-layer dynamic bayesian network (DBN) so that an influence parameter could be estimated for each participant. This was evaluated qualitatively by comparison with ground truth annotations. The annotations involved distributing a proportionate level dominance to each participant such that the total summed to 1. However, there was no systematic evaluation of the annotations nor quantitative evaluation.

Concurrently, Rienks et al. [12] used audio cues to estimate dominance. They used a larger and more varied corpus consisting of a selection of meetings from the MultiModal Meeting Manager (M4) corpus that was used by [18] and also the AMI corpus [3], totalling 1.5 hours of audio-visual data. They used supervised support vector machines (SVMs) to estimate the dominance of participants in the meetings. Here the dominance of each participant was ranked manually according to their perceived dominance by 10 annotators who were organised such that there were 5 annotations for each of the 8 meetings. Due to the high level of variance between the absolute rankings for all the participants in the meeting, the rankings were distributed into three different bins which represented high, normal, and low perceived dominance. All the features were labelled manually and included non-verbal (e.g. speaking turns, speaking length and floor grabs) and verbal cues (e.g. number of words spoken, number of questions asked).

Soon after, Rienks et al. [13] conducted a comparative study of both their approach presented in [12] and the TPIM of Zhang et al. [18]. They used the same three-point dominance scale created from absolute rankings but this time, the annotations were provided by the meeting participants themselves. The same audio features that were extracted in [12] were used again and they found that the SVM model outperformed the TPIM. The absolute dominance rankings did not provide very clear consensus and since the meetings lasted from 5 to 35 minutes, it is likely that longer meetings would be more difficult to annotate, leading to lower annotator confidence. Also, the idea of splitting the rankings into a three-point dominance scale was rather arbitrary the benefits of this method were not explored more concretely.

More recently, Otsuka et al. [11] used complex nonverbal cues based on automatically extracted gaze patterns to explain pair-wise influence in group discussions. They used 10 minutes of conversational data collected from two 4-participant groups where pre-defined discussion topics. The participants were required to agree on a conclusion to each topic after 5 minutes. Unfortunately, there was no quantative evaluation of their method.

The discussions above highlights three issues;non-verbal feature extraction, the nature of the data, and the annotation and evaluation procedure. A summary of the differences between these different papers are shown in Table 1.

In terms of audio-visual feature extraction, some past work has relied on manually annotated features, which may not be simple to extract automatically, particularly in the case of automatic transcription which would be difficult to extract robustly. While many different audio and visual fea-

Reference	Data	Features	Manual/Automatic	Dominance models	Task
[1]	Debating games (2 hrs)	Audio, Visual	Automatic+Manual	Influence Model	Predict influence
[18]	scripted (M4) meetings (2.5 hrs)	Audio	Automatic	Team-Player Influence Model	Predict influence
[12]	M4 and AMI meetings (1.5 hrs)	Audio	Manual	Static	Dominance (high, normal, low)
[13]	M4 and AMI meetings (1.5 hrs)	Audio	Manual	Static,Dynamic	Dominance (high, normal, low)
[11]	scripted meetings (10 mins)	Audio	Manual	Static	Dominance
[7, 6, 8]	AMI meetings (3-5 hrs)	Audio, visual, Audio-Visual	Automatic	Static	Most and Least Dominant

Table 1. Summary of literature in audomatic dominance estimation.

ture extraction methods were used, there was no systematic study of the benefits of each feature for dominance.

The variety of corpora indicates that dominance can be decoded, to some in both conversational and meeting environments. However, analysing meetings in task-driven scenarios which involve more than just debate leads to more challenging but more realistic data where people are able to move freely within the room, and conduct different activity types that can involve standing and sitting during the meeting. In addition, the length of the meeting can greatly affect the behaviour of participants so that it is likely that shorter time limits on discussions lead to higher levels of engagement and observable behaviour. In real meetings, participants may not maintain such high levels of interest, leading to more subtle underlying group dynamics.

In terms of the annotation, we can observe that analysing perceived or self-reported dominance levels is not straightforward due to the variability of human judgements. However, a full analysis of annotator variability would be interesting may highlight ways in which we can solve automated dominance analysis better. Our own work has tried to address these three issues and the next section provides an overview of the work.

4. Overview of our work

In this section, we summarise our work so far and highlight some interesting results. More details about this work can be found in [7, 6, 8]. We took a subset of the publicly available AMI meeting corpus [3] which contained audio and visual data capturing 5 different teams of 4 participants who met on several occassions to complete a task. 12 meetings sessions from this corpus were selected for our experiments such that a set of 59 non-overlaping 5-minute meeting segments were used for annotation. A total of 21 annotators were used to label the data who were grouped such that the same 3 individuals annotated common meeting segments to enable a majority consensus.

For each meeting, annotators were asked to rank the participants in order of dominance from 1 (most) to 4 (least). It is important to note that the annotators were not provided with a definition for dominance but were asked to provide their own in free-form on completion of the annotations. Using these annotations, we were able to find 34 meetings where all 3 annotators agreed on the most dominant person and 31 meetings where there was full agreement about the least dominant person. Other dominance tasks were also defined in terms of the degree of annotation variability for each meeting. More details can be found in [8].

For each meeting, the annotators reported their level of confidence about the annotations on a 7-point scale where 1 represented high confidence. The average annotator confidence was 1.74 and 2.11 for the labelling of the most and least dominant person respectively, which highlights the increased difficulty of labelling the least rather than most dominant person.

4.1. Audio Features from Individual Microphones

Audio activity features were generated by extracting speech from individual headset microphones for each participant. From this signal, a binary and real-valued speech signal was generated by firstly extracting the energy from the signal and then thresholding this to form a binary signal that represented speaking status as 1 and no speaking as 0. We used the total of the energy (TSE) and the speaking status (TSL) to represent audio activity for each participant. Derived audio features were also used to represent speech activity such as total speaker turns (TST) and total turns without short turns (TSTwoBC). We also used the total number of successful interruptions (TSI) for each participant to indicate their level of dominance. In addition, a histogram was created to characterise the distribution of the turn durations of each person in the meeting (SDHist) to capture so that the frequency of longer and shorter turns.

4.2. Audio Features from a Single Source

In addition to extracting audio features from individual headset microphones, we also experimented with different single-source scenarios where speaker diarization was applied to the signal to discover who the most dominant person was [6]. The task of speaker diarization is to identify speakers and when they spoke from a single source. The diarization engine that we used involved applying an agglomerative clustering method which iteratively merged clusters according to a pair-wise Bayesian information criterion (BIC) score. Calculating the BIC score for each potential cluster pair is a time consuming process and through some faster pre-selection steps to prune the hypothesis space, the computation time could be decreased without serious degradation in performance. With these speed-based improvements, we extracted speaker diarization outputs using increasingly faster versions of the algorithm. We also performed robustness testing by simulating different distant microphone sources with decreasing signal to noise ratio.

Figure 1. Example screenshots from the close-view cameras.

4.3. Video Features from Individual Cameras

Computationally efficient visual activity features were extracted by taking advantage of the features that are already computed for video compression. In our case, we were able to extract visual activity features taken from motion vectors and the residual coding bitrate from MPEG-4 encoded video of each person taken using close-up cameras in the meeting, as shown in Figure 1. Then, 3 different visual activity features were generated that represented the average motion vector magnitude (Vector), residual coding bitrate (Residue) of the visual activity that could be not be attributed to the motion vectors, and the average of both these features (Combo). Again, a real-valued and thresholded binary visual activity signal was extracted for each partipant and the totals of each value were used to indicate dominance. Derived visual activity features were also created based on the same principles as those of the audio features. More details about how these compressed domain video features were extracted can be found in [7].

4.4. Unsupervised Method

Our initial experiments on dominance estimation were based on the hypothesis that dominant people move and talk more [5]. The person with the highest or lowest total feature value for each meeting was selected to be the most or least dominant person respectively. A summary of the feature types and their acronyms are provided in Table 2.

Glossary of Feature Acronyms	
Total Speaking Energy	TSE
Total Speaking Length	TSL
Total Speaker Turns	TST
Total Speaker Turns without short turns	TSTwoBC
Total Speaker Interruptions	TSI
Turn Duration Histogram	SDHist
Total Motion Length	TVL
Total Motion Turns	TVT
Total Motion Interruptions	TVI
Motion Turn Duration Histogram	VDHist

Table 2. Glossary of feature abbreviations

4.4.1 Audio Activity Cues

Using our audio cues we found the highest total value of each feature to indicated the most and least dominant person well. Table 3 shows a summary of the results for the most and least dominant person estimation task. The best performing category for each dominance task is highlighted in bold. It was interesting to observe that both the total speaking length (TSL) and total speaker turns without short turns (TWTwoBC) performed the best for both dominance tasks. There was a slight drop in performance for the least dominant person task, which could be an indication of the difficulty of identifying passive people. This is observed further in the difference in performance between TSE and TST, which could indicate that noise levels in the energy signal is much higher for the passive participants compared to the more active ones. This difficulty in finding the least dominant person was also reflected in the self-reported annotator confidence for annotating the least dominant person. Another interesting observation was the marked improvement in performance of both dominance tasks when the shorter turns were removed from TST to form TSTwoBC indicating that the shorter turns are less correlated with dominance.

Features	Most Dom. Class. Acc.(%)	Least Dom. Class. Acc.(%)
TSL	**85.3**	**83.9**
TSE	82.4	67.7
TST	61.8	71.0
TSTwoBC	**85.3**	**83.9**

Table 3. Performance of **audio** cues for both dominance tasks using the **unsupervised** model.

Dominance estimation from a single audio source In addition to estimating the most dominant person from speech activity levels extracted from individual headset microphones, we performed some experiments based on the assumption that there was only a single audio source in the meeting [6]. The single sources were taken from a single microphone located on either the table or ceiling of the meeting room. In addition synthesised audio signals were created from performing a delay-sum on the individual microphones of which there were two types; headset and lapel. While the diarization error rate increased with a lower signal-to-noise ratio (SNR), there was not a clear decrease in performance for the dominance estimation task. This was also observed for the different diarization strategies that were used to decrease computation time.

4.4.2 Visual Activity Cues

The visual activity features were less effective for decoding dominance but performed surprisingly well. Similarly, total visual activity length (TVL) and the visual activity turns (TVT), were the most effective single features for decoding dominance. A summary of the results are shown in Table 4 where the total visual activity turns only include the turns which are greater than $4s$ long. Similar to the audio activity features, there was a decrease in performance between the most and least dominant tasks but for the visual activity features, this decrease was more pronounced, and highlights that these features are less well correlated with non-dominant behaviour. Also, the TVT feature performed better than TVL alone for the least dominant person task, which highlights that the shorter turns are not discrminative. Overall, the single audio features TSL and TSTwoBC performed the best and reflects the findings in [9].

Features	Most Dom. Class. Acc.(%)	Least Dom. Class. Acc.(%)
TVL(Residue)	**76.5**	45.2
TVT(Combo)	**76.5**	**64.5**

Table 4. Performance of **visual** cues for both dominance tasks with **unsupervised** model.

4.4.3 Feature Fusion for Dominance Estimation

We conducted experiments to observe how both the audio and video modalities affected the performance of the dominance estimation task using supervised SVMs.

Audio Feature Fusion Our results from fusing audio-activity features only, found that there was some complementary nature to the features which led to a 6% absolute increase in performance for the most-dominant person classification task as shown in Tables 5 and 3. Also, while the total speaker interruptions (TSI) did not perform so well as an individual feature, it appeared often as a good complementary feature for other audio cues. This is supported by [17] who stated that interruptions could be “a device for exercising power and control in conversation” and also by [16] since interruptions do not always correspond to dominant behaviour but to an individual’s level of engagement. For the least dominant person task, we did not observe any increase in performance when audio features were combined.

Features	Class. Acc.(%)
TSL, TSE, TST	88.2
SDHist, TSE, TST, TSI	**91.2**
Random Guess	25.0

Table 5. Performance of **audio** cues for **most**-dominant person with a supervised model.

Video Feature Fusion We conducted similar experiments to measure the complementary nature of the visual activity features. Results for estimating the most dominant person in the meetings are shown in Table 6. After combining the features, approximately 3% in improvement was possible with the feature combination (VDHist, TVL, TVT (Residue)). For the least dominant person task, fusing video features led to a reduction in performance.

Features	Class. Acc.(%)
VDHist, TVL (Residue)	73.5
VDHist, TVT (Residue)	76.5
VDHist, TVL, TVT (Residue)	**79.4**

Table 6. Performance of **visual** cues for **most**-dominant person task using supervised model.

Audio-Visual feature fusion SVMs were applied to fusing the speech and visual activity features and some results for the estimation of the most dominant person are shown in Table 7. Overall, the audio-visual feature combinations did not perform better than the audio-only combinations for the most dominant person task. Also, for the least dominant person task, the best performing (80.7%) audio-visual combination (SDHist, TSE, TST, TSI, VDHist) or (SDHist, TSE, TST, TSI, TVL) performed worse than the best unsupervised case (SDHist, TVL) (83.9%).

Feature	Class. acc. (%)
SDHist, TSE, TST, TSI, TVL	**91.2**
SDHist, TSE, TST, TSI, VDHist	**91.2**
SDHist, TSE, TST, TSI, VDHist, TVL	82.4

Table 7. Performance of **audio-visual** cues with **most**-dominant person task.

A summary of the best performance of the different dominance tasks using audio, visual and audio-visual cues is shown in Figure 2. The visual activity features performed worse than the audio features and also worst for each dominance task. Also the audio-visual activity features could not out-perform the audio-only features. Estimating the least dominant person was more difficult, resulting in lower performance in all cases which highlights the increase in noise of the features when we are trying to identify people who have a low activity levels. This is also reflected in the self-reported annotator confidence values.

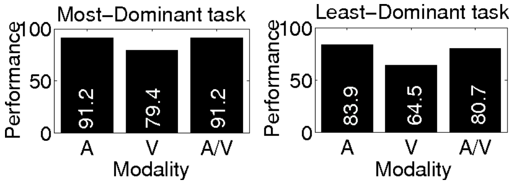

Figure 2. Comparison of the best performance values for (A)udio, (V)ideo, audio-visual (A/V) modalities, and each dominance task.

4.5. Beyond simple single-modality features

So far, we have described our work on investigating simple single-modality features where our hypothesis was that a person’s non-verbal behaviour could be characterised in terms of audio or visual activity levels. However, a brief observation of our annotators free-form definitions of dominance found that the dominant person tended to receive more visual attention from the others when they spoke. This is very similar to Dovidio and Ellyson’s idea of the visual dominance ratio (VDR) [4]. This ratio is defined as the proportion of time spent looking while speaking over looking while listening. They hypothesised that dominant people tend to address the other more while listening to them less.

To use the VDR to decode dominance in group meetings requires some modifications to the ratio since the VDR is designed for dyadic conversations. We redefined it for multi-party conversations so that the ratio quantifies the time each participant looks at others while speaking (TLWS) compared to the time they spend looking at other speakers (TLWL). We used human annotations of the visual focus of each partipant for these preliminary experiments and the results are summarised in the Table 8 below. It was interesting to observe that visual dominance ratio (VDR) performed worse because the total time spent looking while listening (TLWL) was not very decriminative. In contrast, the total time spent looking while speaking was a more descriminative feature. This higlights again the problem of detecting passive behaviour, and this case, listening.

Features	Class. Acc. (%)
VDR	70
TLWS	**76**
TLWL	44

Table 8. Summary of results for estimating dominance using the visual dominance ratio and other VFOA related features

5. Challenges

The discussions and overview have explored some of the challenges in automated dominance modeling and estimation but some open issues still remain. During our own studies, we found that there were ambiguous cases where the most dominant person was estimated inaccurately because there was ambiguity between the dominance of two of the participants. In these cases, it is possible that dominant cliques exist rather than just individuals. This can be viewed as a sort cooperative behaviour that would be interesting to infer from the group dynamics and addresses more subtle aspects of group hierachy.

Experiments using a single distant microphone to estimate the dominance of meeting participants showed that the performance was not sensitive to increased noise levels in the signal. Visual features from more distant cameras could also be used to extract visual focus of attention in meetings where people could look at others and objects in the environment. This could also be used to estimate when someone is addressing others or listening to someone. However, the problem of detecting passive behaviour such as listening remains to be challenging. Finally, while speaker interruptions have been addressed, the extraction method is crude and could be improved by analysing the quality of the interruption, as suggested by [16].

The bulk of the work presented here has dealt with static measures of dominance. For those which were dynamic, their performance tended to be worse. This could be viewed as counter-intuitive since Millar and Millar [14] already defined dominance in terms of 'one-up' and 'one-down' interactions. However, while this described dyadic interactions, group interactions may have a different dynamic where individuals can have influence on more than one person at a time. One possible explanation is that the proposed dynamic models either encoded dyadic interactions or group interactions but not both in combination. Finding a suitable way to identify dominance through dyadic relationships is difficult however, particularly as the number of participants increases and the number of combinations of dyadic or clique-based interactions becomes intractable.

Finally, using natural data where individuals are strongly driven to dominate others for their own goals is difficult to acheive. While the data that we used captured natural meetings, the participants volunteered to take part in the study but did not have a particular vested interested in the outcome. It is also important to note that real corporate meetings do not always involve just talking so the use of tools such as a whiteboard or slide screen or table makes extracting contextual features about the meeting activities for dominance analysis interesting but also challenging.

6. Conclusion

While much work in the area of automated dominance estimation has relied on using complex models, we have shown that using challenging meeting scenarios where the participants are able to behave naturally and much simpler methods, we are able to gain superior performance. In addition, our detailed studies of the limitations with working with single modalities as well as the benefits of fusion highlights what can already be acheived and where focus in this challenging area of research should be in the future.

While our meeting data has provided natural interactive activity, we were not able to capture participants who were truly driven to attain their own or their team's goals during the group exchanges. Capturing data which captures the everyday dynamic of employees in a company for example, would provide a richer framework for analysing dominant behaviour in individuals and indeed cliques. However, it remains to be seen whether this data could be captured easily and accurately within an environment where individuals are aware that their behaviour, motivated by their own personal goals, is being recorded in detail.

Acknowledgements This research was partly funded by the US VACE program, the EU project AMIDA (pub. AMIDA-30), the Agency for Science, Technology and Research, Singapore (A*STAR), and the Swiss NCCR IM2. We thank Dinesh Babu Jayagopi, Chuohao Yeo, Yan Huang, Gerald Friedland, Sileye Ba, and Jean-Marc Odobez for their helpful discussions.

References

[1] S. Basu, T. Choudhury, B. Clarkson, and A. Pentland. Learning human interactions with the influence model. In *NIPS*, 2001. 2, 3

[2] 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. 2

[3] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraiij, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, M. McCowan, W. Post, D. Reidsma, and P. Wellner. The ami meeting corpus: A pre-announcement. In *Proc. MLMI*, 2005. 2, 3

[4] 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. 1, 2, 5

[5] N. E. Dunbar and J. K. Burgoon. Perceptions of power and interactional dominance in interpersonal relationships. *Jour-*

nal of Social and Personal Relationships, 22(2):207–233, 2005. 1, 2, 4

[6] H. Hung, Y. Huang, G. Friedland, and D. Gatica-Perez. Estimating the dominant person in multi-party conversations using speaker diarization strategies. In *International Conference on Speech and Signal Processing*, 2008. 3, 4

[7] H. Hung, D. Jayagopi, C. Yeo, G. Friedland, S. Ba, J.-M. Odobez, K. Ramchandran, N. Mirghafori, and D. Gatica-Perez. Using audio and video features to classify the most dominant person in a group meeting. In *ACM Multimedia*, 2007. 3, 4

[8] D. B. Jayagopi, H. Hung, C. Yeo, and D. Gatica-Perez. Modeling dominance in group conversations using non-verbal activity cues. Technical Report 78, IDIAP, Rue de Pres Beudin, December 2007. 3

[9] M. S. Mast. Dominance as expressed and inferred through speaking time. *Human Communication Research*, (3):420–450, July 2002. 1, 5

[10] Y. Ohsawa, N. Matsumura, and M. Ishizuka. Influence diffusion model in text-based communication. In *World Wide Web Conference*, Honalulu, Hawaii, May 2002. 2

[11] 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. 2, 3

[12] R. Rienks and D. Heylen. Automatic dominance detection in meetings using easily detectable features. In *Proc. Workshop on Machine Learning for Multimodal Interaction (MLMI)*, Edinburgh, Jul. 2005. 2, 3

[13] R. Rienks, D. Zhang, D. Gatica-Perez, and W. Post. Detection and application of influence rankings in small group meetings. In *ICMI '06: Proceedings of the 8th international conference on Multimodal interfaces*, pages 257–264. ACM Press, 2006. 2, 3

[14] E. Rogers-Millar and F. M. III. Domineeringness and dominance: A transactional view. *Human Communication Research*, 5(3):238–246, 1979. 1, 6

[15] E. Rosa and A. Mazur. Incipient status in small groups. *Social Forces*, 58(1):18–37, September 1979. 1

[16] D. Tannen. *Gender and Discourse*, chapter Interpreting Interruption in Conversation, pages 53–83. Oxford Univesrity Press, 1993. 5, 6

[17] 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. 1, 5

[18] D. Zhang, D. Gatica-Perez, S. Bengio, and D. Roy. Learning influence among interacting Markov chains. In *NIPS*, 2005. 2, 3