Analyzing Group Interactions in Conversations: a Review

Daniel Gatica-Perez

Abstract— **Multiparty face-to-face conversations in professional and social settings represent an emerging research domain for which automatic activity-based analysis is relevant for scientific and practical reasons. The activity patterns emerging from groups engaged in conversations are intrinsically multimodal and thus constitute interesting target problems for multistream and multisensor fusion techniques. In this paper, a summarized review of the literature on automatic analysis of group activities in face-to-face conversational settings is presented. A basic categorization of group activities is proposed based on their typical temporal scale, and existing works are then discussed for various types of activities and trends including addressing, turn taking, interest, and dominance.**

I. Introduction

Devising computational frameworks to automatically infer human activities from sensor data constitutes an open problem in many domains, including signal processing, computer vision, sensor fusion, human-computer interaction, and ubiquitous computing. Moving beyond the individual-centered paradigm [47], an emerging body of research has started to explore multiperson scenarios, where group interactions and activities - and not only activities performed by single people - are relevant [27], [39], [6], [35].

In this context, face-to-face conversations represent a fundamental case whose automatic activity-based analysis has value on their own for several social sciences [5], [36], and that would open doors to a number of relevant applications. In the workplace, examples of face-to-face settings include casual "stop-by" peer conversations, regular group discussions, formal meetings, and presentations [50], [38], [11]. In the personal sphere, face-to-face interactions are ubiquitous, and constitute by far - despite the increased use of computed-mediated communication - the most natural, enjoyable, and effective way to experience and fulfill our social needs [42]. Needless to say, the automatic analysis of face-to-face multiparty conversations poses a diversity of technical challenges, given the intrinsic complexity of the patterns emerging in real communication, and the difficulty to represent and infer the activities of multiple interacting people from multisensor data with tractable yet accurate computational models.

This paper represents an attempt to draw a map of the existing work in this field. The goal of the paper is to gather and briefly discuss works which, given the interdisciplinary nature of the domain, have appeared in the literature spread over a number of communities, including signal processing, computer vision, multimodal processing, machine learning, human-computer interaction, and ubiquitous computing. Given the rapid developments in the domain, the author does not claim to have been fully exhaustive in the review, but rather aims at introducing, in a comparative fashion, a number of works regarded as representative either by the addressed research problem or by the proposed solution, while providing up-to-date pointers to the literature to a non-expert reader. The emphasis of the review is on (1) conversational settings, thus not including many other multiperson activity scenarios (e.g. surveillance); (2) face-to-face communication, rather than remote, computer-mediated one; (3) multiparty conversations, i.e., cases involving more than two people; and (4) the use of multimodal perceptual data, rather than only speech. Whenever available, pointers to the social psychology literature, which can be seen both as a motivating factor for some of the research described here and as a source of knowledge to support the design of computational models, will be provided.

The paper is organized as follows. Section II discusses a categorization of groups and conversational group activities. Sections III, IV, V, and VI describe the four major activity categories considered in the paper, namely addressing, turn taking, interest, and dominance. Section VII provides some concluding remarks.

II. Categorizing conversational group activities

As documented by a significant amount of work in social psychology for over 50 years [5], [36], groups in conversations, both in professional or social settings, can be seen as proceeding through diverse communication phases in the course of their interaction, sharing information, engaging in discussions, making decisions, dominating outcomes, etc. Group activities involve multiple participants effectively constrained by each other through complex social rules. Group interaction is also multimodal [33]. On one hand, speech is clearly the principal modality in conversations. However, in natural meetings, speech is spontaneous and multiparty, containing disfluencies, no clear sentence boundaries, and significant overlapping, phenomena that constitute challenges for speech processing [46]. On the other hand, there exists a wealth of information in the visual modality in the form of gaze, gestures, and expressions [37], due to the fact that both individual and group activities are often defined by the joint occurrence of specific audio and visual patterns. The same applies to other types of sensor data (e.g. body signals) that can be used as cues for inferring activity.

The author is with the IDIAP Research Institute, Martigny, and the Ecole Polytechnique Federale de Lausanne (EPFL), Switzerland. email: gatica@idiap.ch

For purposes of organization of the existing approaches, a simple categorization of group activities in conversations is presented here based on two axis, the first one representing the temporal scale spanned by the activities, and the second one describing the group size. In the first axis, group communication patterns vary from the short term to the very long term, ranging from *addressing* (i.e., who speaks to whom at every time step), to *turn taking patterns* of longer temporal support (e.g. floor control, and discussions vs. monologues), to group trends, like *interest level*, which require longer intervals for their definition (e.g. segments where a group was highly engaged in a discussion), to trends that emerge in a group from the regular interaction of its members over time, like *dominance*. The temporal scale of the described group activities can thus span single or multiple conversations. In the second axis, conversations can span dyadic cases, small, and large groups. It is well known in social psychology that the size of a group has a definite influence in the dynamics of a conversation [23]. In the remainder of the paper, we will describe existing work focusing on the four activity categories listed above. Additionally, research on pair-wise conversations will be discussed if they relate to works whose ultimate goal is the modeling of groups. Finally, it is important to clarify that the review will discuss techniques that rely on transcribed speech only briefly, whenever there is a clear relation with multimodal techniques.

III. Addressing

In a conversation, an addressee is the person at whom the speech is directed [12]. In social psychology, it is known that the addressing phenomenon occurs through different communication channels, including speech, gaze, and gesture, e.g. listeners manifest attention by orienting their gaze to speakers, who in turn use gaze to indicate whom they address, and to ensure visual attention from addressees to hold the floor [25]. It is also known that meeting participants, interacting and exchanging roles as speakers, addressees, and side participants, contribute to the emergence of conversational events that characterize the flow of a meeting (monologues, group discussions, side conversations).

Although there is an increasing body of work on automatic analysis of head pose (as a surrogate for gaze) and visual focus of attention (VFOA) in multiparty interaction [49], [3], [4], there are relatively few studies on automatic identification of addressees in multiparty conversations. In brief, the goals of the existing works are to identify what participant(s) in a conversation the current speaker is talking to, and to explore the connections between addressee modeling and other conversational activities, like the ones described in Section IV. Katzenmeier et al. [31] presented a study on identifying addressees between two people and a simulated robot, with the further goal of discriminating person-person interaction from human-robot interaction. Three cases were studied: audio-only using speech-derived features, visual-only based on head pose, and audio-visual combining the single modality cues. A Bayesian classification technique was used, in which neural networks were used to learn head pose and audio representations. In this three-agent scenario, it was found that head pose is indeed a strong cue for addressee identification, and that the best performance was obtained with the multimodal approach, despite a relatively low performance obtained with the audio modality. In other work, Jovanovic et al. [29] presented an initial scheme of verbal, non-verbal, and contextual features for addressee identification, but no experiments were conducted to validate the proposed scheme. In subsequent work, Jovanovic et al. [30] collected and annotated a five-hour corpus recorded in a multisensor meeting room (cameras and microphones) for studying addressee behavior, consisting of twelve 25-minute real meetings. The corpus was annotated with respect to discrete VFOA for each participant, addressee information, and dialogue acts (DA: units that include backchannels, floor grabbers, questions, and statements), and so it is relevant for studying the problem using a variety of cues. The annotation of addressees used dialog acts as basic units, assigning one of four possible tags to each DA, to indicate whether the speaker addresses a single person, a subgroup, the whole audience, or if the addressee is unknown. The detailed discussion about the reliability of the manual annotation process (inter-annotator agreement) in [30] indicates that the annotation ranges in quality from acceptable to good for those DAs whose boundaries are agreed upon by manual annotators, that the reliability is higher on those meeting segments where the speaker addresses a single person, and that annotators had problems distinguishing between subgroup and group addressing. All these findings should be useful to assess the type of performance that can be expected with automatic processing. In other related work, Otsuka et al. [41] recently presented a Dynamic Bayesian Network (DBN) approach which jointly infers the gaze pattern for multiple participants and the conversational gaze regime responsible for generating specific speaking activity and gaze patterns (e.g. all participants converging onto one person, or two people looking at each other). The work relied on clean observations extracted from magnetic head trackers attached to each participant and from a manual speaking-turn segmentation. Overall, it can be said that the area of automatic addressee modeling is still emerging, and that as of today, the performance of systems relying on fully automatic features remains unknown.

IV. Turn-taking patterns

Viewed as a whole, a group in a conversation proceeds through diverse communication phases in single meetings as well as during the course of long-term collaborative work. A model based on this observation would then assume a discrete set of group activities and view a group conversation as a sequence of such activities. In a formal meeting scenario, where people discuss around a table and use a whiteboard and a projector screen, McCowan et al. [34], [35] first investigated this approach and targeted the joint segmentation and recognition of meetings into a number of group activities that correspond to location-based turn-taking patterns, including monologues, discussions, presentations, note-taking, etc. The approach relied on supervised learning techniques, namely Hidden Markov Models (HMMs) [44], and used a number of simple audio and visual features automatically derived from multiple cameras and microphones, including close-talk and microphone arrays. The problem was studied as a multistream system, where data streams can correspond either to the features extracted from each person or to each perceptual modality (audio or visual). A number of variations of HMM models were tested, including multistream HMMs [17], coupled HMMs [8], and asynchronous HMMs [7]. Results, measured in terms of Action Error Rate (AER), were encouraging and showed the benefits of audio-visual fusion. The approach, however, has two limitations. First, HMMs can be challenged by a large number of parameters, and by the risk of overfitting when learned from limited data [40]. This situation might occur in the case of multimodal group activity recognition where, in the simplest case, large vectors of audio-visual features from each participant are concatenated to define the observation space. Second, the framework does not explicitly model activities at different semantic levels, despite the known fact that models in social psychology describe meetings as comprising both individual actions and interactions [36].

Zhang et al. [53], [57] addressed the above limitations with a two-layer HMM framework [40]. In the first layer, individual actions performed by each person, such as *writing* and *speaking*, are recognized (i.e., estimating posterior probabilities of the individual actions) from raw audio-visual observations. Then, the second layer recognizes the group activities using as input the results of the first-layer recognizers for all meeting participants and a set of group features, directly extracted from the raw streams and not associated to any person. Compared with single-layer HMMs, layered HMMs have several advantages, including the use of much smaller observation spaces, the fact that the low-layer HMMs can in practice be better estimated as much more data (arising from multiple people) is available, the reduction in sensitivity for group activity recognition as the observations for the high-layer are posterior-based features, and the possibility of exploring different HMM options for each layer. The experiments in [53], [57] led to three findings. First, the two-layer HMM approach outperformed the single-layer one. Second, the use of audio-visual features outperformed the use of single modalities for both single-layer and two-layer HMMs, supporting the hypothesis that the target group activities are inherently multimodal. Third, the best low-layer model was the asynchronous HMM (a model that explicitly accounts for variations of alignment between two data streams), which suggested that some asynchrony might exist for the group activities, and that such asynchrony was reasonably captured by the model.

Other works have targeted the recognition of the same group activities with hierarchical representations. Dielmann et al. [15] proposed two approaches for meeting structuring from audio-only features using multilevel Dynamic Bayesian Networks (DBNs). The first DBN decomposed the group activities as sequences of sub-actions with no explicit meaning. The second DBN processed features of different nature independently, and integrated them at a higher level. In this work, the sub-actions have no obvious interpretation, and their number is a model parameter learned during training or set by hand, which makes the structure of the models more difficult to interpret. An initial comparison of various recognition models on the same task, including the layered HMM, the multilevel DBN, and other approaches, was presented by Al-Hames et al. [1].

Two variations of the problem have been explored by Zhang et al. [54], [55]. These approaches look at the problem from a practical perspective: the manual labeling of group activities for training purposes is both difficult (in terms of the annotation agreement that might be obtained from multiple annotators) and expensive. The use of unsupervised or partially supervised approaches could thus be attractive alternatives. The approach in [54] proposed a two-layer framework where the low-layer is identical to the one presented in [53], and the high-layer is a fully unsupervised HMM that discovers group activity patterns using the output from the low-layer as features. The method in [55] uses model adaptation techniques, where instead of directly training one model for each group activity (as done in [53]), a general activity model is first estimated using unlabeled data, and then adapted to each group activity model using its own labeled data using Bayesian adaptation. Both methods define a tradeoff between performance and availability of labeled data. In the author's opinion, the investigation of models that rely on unsupervised or lightly supervised approaches remains as a research area of practical relevance, given the increasing availability of unlabeled data and the annotation costs required by supervised methods.

In other related work, Chen et al. [9] described initial efforts to combine gaze, gesture, and speech for floor control modeling, using meeting data collected with multiple cameras, microphones, and magnetic sensors. As a conversation progresses, the active speaker holds the floor, while other conversants participate either cooperating or competing to share the floor and advance towards completing the goals of a conversation. Floor control is a phenomenon studied in psychology and linguistics for many years [18] and has been observed that multimodal cues (including gaze exchanges between the floor holder and the interlocutors, and discourse markers) are related to floor control changes. The study about the use of audio-visual cues for floor control has been recently extended in [10]. The work includes the definition of a scheme for floor control annotation, and the use of a small labeled corpus to identify multimodal cues correlated with floor changes. The analysis of the corpus suggests that discourse markers occur frequently at the beginning of a floor, that mutual gaze between the current floor holder and the next one offer occurs during floor transitions, and that gestures related to floor capturing also occur. No attempt to perform automatic processing was reported

Finally, works related to recognition of speaker turn categories that rely on transcribed speech have also appeared in the speech community. A number of existing works has focused on recognizing DAs automatically. Examples include the approaches for joint DA segmentation and classification presented in [2], [28], [58].

V. Group interest

Group interest-level, i.e., the degree of engagement that the members of a group collectively display during their interaction, is an important trend to extract from formal meetings and other social settings [42]. Segments of conversations where participants are highly engaged (e.g. in a discussion) are likely to be of interest to other observers too. In this view, group interest-level helps define a form of relevance around which conversations can be indexed or summarized.

Modeling interest-level and other related concepts is an emerging problem in social computing that has been explored in multiperson conversational settings [51], [52], [32], [19], [24], [42], [43]. However, with a few exceptions which have explored the use of multimodal cues [24], [43], all existing work has only analyzed the relation between interest and the speech modality. With speech utterances as the basic units, work by Wrede et al. introduced the concept of hot-spots [51], defining them in terms of participants highly involved in a discussion, and relating it to the concept of activation in emotion modeling [14]. The work in [51] studied the relation between prosodic cues and human-annotated hot-spots. This work was later extended to study the relation between hot-spots and dialog acts [52], using both contextual features (such as speaker identity or type of the meeting) and lexical features (such as utterance length and perplexity). In a related line of work, Kennedy et al. defined emphasis for speech utterances [32], acknowledging that this concept and emotional involvement might be acoustically and perceptually similar. Other works in the speech community are also related to detection of high-interest segments. For instance, Hillard et al. [26] proposed to recognize a specific kind of interaction pattern in meetings (agreement vs. disagreement) that is related to high group interest. The work used both word-based features (such as the total number of words, and the number of "positive" and "negative" keywords), as well as prosodic cues (such as pause, frequency and duration), in a learning approach that made use of unlabeled data.

A number of wearable computing systems have also dealt with the interest-level problem, either introducing it manually as in the work by Eagle et al. [19], or estimating it automatically from acoustic features as proposed by Pentland et al. [43]. In the latter case, audio-based features of activity, engagement, stress, and mirroring, and body motion features from accelerometers were automatically extracted. The conversational settings varied from dyadic conversations (including same-sex conversations with random topics and speed-dates) to multiparty meetings (e.g. conference attendees where participants are likely to exchange business e-cards at some point if they are interested in each other).

Gatica-Perez et al. [24] presented a preliminary investigation of the performance of audio-visual cues on discriminating high vs. neutral group interest-level segments in multiparty meetings in a fully supervised approach, simultaneously deriving a segmentation of a meeting and the binary classification of the segments into high or neutral interest-level classes. Two classic HMM recognition strategies were investigated: early integration, where all desired streams (audio, visual, or audio-visual) are synchronized and concatenated to form the observation vectors, and multistream HMMs, used for audio-visual fusion. The fully supervised approach called for human annotation of group interest-level. The investigated features included audio features derived from microphone arrays and lapel microphones, and visual features extracted from skin-color blobs from each participant. Various combinations of models and features (audio-only, video-only, audio-video) were investigated. The analysis of the results suggested that the automatic detection of group interest-level is promising, and that, while the audio modality turned out to be dominant, audio-visual fusion improves performance and is thus beneficial. The investigation visual features better correlated with communicative tasks (e.g. visual focus) remains as an open issue.

VI. Dominance and Influence

Some people seem particularly capable of driving a conversation and often have the largest influence on a meeting, shifting its focus when they speak. Dominance and influence are important research problems in social psychology, and a solid body of knowledge about the multimodal nature of these phenomena exists [21]. However, the problem of automatically estimating them has only begun to be studied in the contexts of social and wearable computing [6], [11], [45], [56]

The perception of dominance is a multimodal task, in which visual gaze and speaking activity are involved. In social psychology, early research by Efran showed that high-status persons receive more visual attention than low-status people [20], and work by Cook et al. showed that people who very rarely look at others in conversations are perceived as weak [13]. Further studies have shown that the joint occurrence of visual attention and speaking activity patterns are correlated with dominance. For instance, Exline et al. showed that high-power people exhibit a relatively high ratio of looking-while-speaking to looking-while-listening periods [22]. Importantly, Dovidio et al. showed that people can systematically decode patterns of visual dominance displayed by others [16], which provides support for both the expectation of producing reliable human annotations and the hope of designing methods for automatic analysis. This is in fact what the initial work in this domain has suggested [45], [56].

Basu et al. [6] described an approach for automatic discovery of influence in a lounge room equipped with cameras and microphones where people played interactive debating games. The influence model, a DBN which models the members of a group as a set of Markov chains, each of which influences the others' state transitions, was applied to automatically determine how much influence a person has on each of the others on a pair-wise basis. Although the influence model (and other related models, e.g. [11]) is a tractable and thus attractive alternative to model group interactions, it has the limitation that it only models the interactions between individual players on a pair-wise basis, i.e., the influence of one player on another player, and does not explicitly model the group as such.

As an alterative, Zhang et al. recently proposed a two-level influence model [56], which is a DBN with a two-level structure: the player level and the team level. The player level represents the actions of individual players. The team level represents group-level actions. The team state at the current time step influences the players' states at the next time step. In turn, the team state at the current time step is also influenced by all the players' states at the current time step. The explicit hierarchy in the model allows for the estimation of the influence of each of the players on the team state, and the distribution of player-to-team influence is automatically learned from data in an unsupervised fashion. Regarding features, audio and speech features were extracted from multiparty meetings from speaker turns, using close-talk microphones, microphone arrays, and manual speech transcripts. Using ground truth obtained by manually annotating influence by multiple annotators, the team-player influence model was found to outperform a method that used each participant's speaking length (i.e., the proportion of time during which each participant speaks) as an estimate of their overall influence in the meeting.

Rienks et al. [45] recently proposed a supervised learning approach to detect dominance in meetings. Their method was based on the formulation of the problem as a three-class classification task in which, through manually annotated data, meeting participants were labeled as having high, normal, or low dominance. A number of features related to speaker-turns and their content were extracted for each participant from speaker-turn segmentations, speech transcriptions, and addressing labels, all of which were manually produced. These features included a person's speaking time, her number of taken turns, the number of times the person grabbed the floor, the number of times the person was privately addressed, etc. Using a small corpus of meetings and a Support Vector Machine (SVM) classifier, the authors obtained a performance of 75% correct classification rate for the best feature combination (number of floorgrabs and number of taken turns).

Overall, the automatic estimation of dominance and influence is also a research problem for which many issues, both theoretical and empirical, remain open, including the validation of cues from the social psychology literature for automatic analysis, a clear understanding of the benefits of audio-visual fusion, the evaluation of fully automatic features, and the design of models to estimate variations of these trends over time.

VII. Conclusions

This paper has presented a concise overview of some of the many facets of research on automatic recognition and discovery of group activities in conversational settings from multiple sensors, with the intention of providing links to recent literature on a number of relevant communicative tasks. As the discussion has tried to highlight, the domain is very challenging and is still emerging. Research resources, including data, annotations, and performance evaluation measures are emerging too. However, it is expected that work in this domain will soon address, at least initially, some of the many open issues, finding principled ways of integrating the diverse knowledge brought by the various communities working in this domain.

Acknowledgements. This work was supported by the Swiss NCCR on Interactive Multimodal Information Management (IM2), and the EC project AMI (Augmented Multi-Party Interaction, publication AMI-189). The author thanks Dong Zhang and Samy Bengio (IDIAP), and Iain McCowan (eHealth Research Center, Brisbane, Australia), for their contribution to our research on conversational activity modeling, Jean-Marc Odobez (IDIAP) for discussions about VFOA, and Rutger Rienks (University of Twente, the Netherlands) for pointers to the social psychology literature on dominance and influence.

References

[1] M. Al-Hames, A. Dielmann, D. Gatica-Perez, S. Reiter, S. Renals, G. Rigoll, and D. Zhang, "Multimodal integration for meeting group action segmentation and recognition," in *Proc. Workshop on Machine Learning for Multimodal Interaction (MLMI)*, Edinburgh.

[2] J. Ang, Y. Liu, and E. Shriberg, "Automatic dialog act segmentation and classification in multiparty meetings," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Philadelphia, Mar. 2005.

[3] S. O. Ba and J.-M. Odobez, "A probabilistic framework for joint head tracking and pose estimation," in *Proc. Int. Conf. on Pattern Recognition (ICPR)*, Cambridge, Aug. 2004.

[4] S. O. Ba and J.-M. Odobez, "A study on visual focus of attention modeling using head pose," in *Proc. Workshop on Machine Learning for Multimodal Interaction (MLMI)*, Washington, DC, May 2006.

[5] R.F. Bales, *Interaction Process Analysis: a method for the study of small groups*, Addison-Wesley, 1951.

[6] 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.

[7] S. Bengio, "An asynchronous Hidden Markov Model for audio-visual speech recognition," in *Proc. Conf. on Advances in Neural Information Processing Systems, NIPS 15*, Vancouver, Dec. 2002.

[8] M. Brand, "Coupled Hidden Markov Model for modeling interacting processes," *MIT Media Lab, Vision and Modeling, Technical Report 405*, Nov. 2002.

[9] L. Chen, T. R. Rose, F. Parrill, X. Han, J. Tu, Z. Huang, M. Harper, F. Quek, D. McNeill, R. Tuttle, and T. Huang, "VACE multimodal meeting corpus," in *Proc. Workshop on Machine Learning for Multimodal Interaction (MLMI)*, Edinburgh, Jul. 2005.

[10] L. Chen, M. Harper, A. Franklin, T. R. Rose, I. Kimbara, Z. Huang, and F. Quek, "A multimodal analysis of floor control in meetings," in *Proc. Workshop on Machine Learning for Multimodal Interaction (MLMI)*, Washington, DC, May 2006.

[11] T. Choudhury and S. Basu, "Modeling conversational dynamics as a mixed memory markov process," in *Proc. NIPS*, Dec. 2004.

[12] H. H. Clark and T. B. Carlson, "Hearers and speech acts," *Language*, vol. 58, no. 2, pp. 332–373, Jun. 1982.

[13] M. Cook and J. M. C. Smith, "The role of gaze in impression formation," *British Journal of Social and Clinical Psychology*, 1975.

[14] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. Taylor, "Emotion Recognition in Human-Computer Interaction," *IEEE Signal Processing Magazine*, 2001.

[15] A. Dielmann and S. Renals, "Dynamic Bayesian networks for meeting structuring," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Montreal, May 2004.

[16] J. F. Dovidio and S. L. Ellyson, "Decoding visual dominance: atributions of power based on relative percentages of looking while speaking and looking while listening," *Social Psychology Quarterly*, vol. 45, no. 2, pp. 106–113, Jun. 1982.

[17] S. Dupont and J. Luettin, "Audio-visual speech modeling for continuous speech recognition," *IEEE Transactions on Multimedia*, vol. 2, no. 3, pp. 141–151, September 2000.

[18] S. Duncan, "Some signals and rules for taking speaker turns in conversations", *Journal of Personality and Social Psychology*, vol. 23, no. 2, pp. 283–292, 1972.

[19] N. Eagle and A. Pentland, "Social network computing," in *Proc. Int. Conf. on Ubiquituous Computing (UBICOMP)*, Seattle, Oct. 2003.

[20] J. S. Efran, "Looking for approval: effects of visual behavior of approbation from persons differing in importance," *Journal of Personality and Social Psychology*, vol. 10, no. 1, pp. 21–25, Sep. 1968.

[21] S. L. Ellyson and J. F. Dovidio, Eds., *Power, Dominance, and Nonverbal Behavior*, Springer-Verlag., 1985.

[22] R. V. Exline, S. L. Ellyson, and B. Long, "Visual behavior as an aspect of power role relationships," *Advances in the study of communication and affect*, 1975.

[23] N. Fay, S. Garod, and J. Carletta, "Group discussion as interactive dialogue or serial monologue: the influence of group size," *Psychological Science*, vol. 11, no. 6, pp. 487–492, 2000.

[24] D. Gatica-Perez, I. McCowan, D. Zhang, and S. Bengio, "Detecting group interest-level in meetings," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Philadelphia, Mar. 2005.

[25] C. Goodwin, *Conversational Organization: Interaction Between Speakers and Hearers*, vol. 11, Academic Press, New York, NY, 1981.

[26] D. Hillard, M. Ostendorf, and E. Shriberg, "Detection of agreement vs. disagreement in meetings: Training with unlabeled data," in *Proc. HLT-NAACL Conference*, Edmonton, May 2003.

[27] S. Hongeng and R. Nevatia, "Multi-agent event recognition," in *Proc. IEEE Int. Conference on Computer Vision*, Vancouver, July 2001.

[28] G. Ji and J. Bilmes, "Dialog act tagging using graphical models," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Philadelphia, Mar. 2005.

[29] N Jovanovic and R. op den Akker, "Towards automatic addressee identification in multi-party dialogues," in *Proc. SIGDial Workshop on Discourse and Dialogue*, Boston, Apr. 2004.

[30] N Jovanovic, R. op den Akker, and A. Nijholt, "A corpus for studying addressing behavior in multi-party dialogues," in *Proc. SIGDial Workshop on Discourse and Dialogue*, Lisbon, Sep. 2005.

[31] M. Katzenmeier, R. Stiefelhagen, and T. Schultz, "Identifying the addressee in human-human-robot interactions based on head pose and speech," in *Proc. Int. Conf. on Multimodal Interfaces (ICMI)*, State College, PA, Oct. 2004.

[32] L. Kennedy and D. Ellis, "Pitch-based emphasis detection for characterization of meeting recordings," in *Proc. ASRU*, Virgin Islands, Dec. 2003.

[33] R. Krauss, C. Garlock, P. Bricker, and L. McMahon, "The role of audible and visible back-channel responses in interpersonal communication," *Journal of Personality and Social Psychology*, vol. 35, no. 7, pp. 523–529, 1977.

[34] I. McCowan, S. Bengio, D. Gatica-Perez, G. Lathoud, F. Monay, D. Moore, P. Wellner, and H. Bourlard, "Modeling human interactions in meetings," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Hong-Kong, Apr. 2003.

[35] I. McCowan, D. Gatica-Perez, S. Bengio, G. Lathoud, M. Barnard, and D. Zhang, "Automatic analysis of multimodal group actions in meetings," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 27, no. 3, pp. 305–317, Mar. 2005.

[36] J.E. McGrath, *Groups: Interaction and Performance*, Prentice-Hall, 1984.

[37] D. McNeill, Ed., *Language and gesture*, Cambridge University Press, 2000.

[38] N. Morgan, D. Baron, J. Edwards, D. Ellis, D. Gelbart, A. Janin, T. Pfau, E. Shriberg, and A. Stolcke, "The meeting project at ICSI," in *Proc. Human Language Technology Conf. (HLT)*, San Diego, CA, March 2001.

[39] N. Oliver, B. Rosario, and A. Pentland, "A Bayesian computer vision system for modeling human interactions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, Aug. 2000.

[40] N. Oliver, E. Horvitz, and A. Garg, "Layered representations for learning and inferring office activity from multiple sensory channels," in *Proceedings of the International Conference on Multimodal Interfaces (ICMI'02)*, October 2002.

[41] K. Otsuka, Y. Takemae, J. Yamato, and H. Murase, "Probabilistic inference of multiparty-conversation structure based on markov-switching models of gaze patterns, head directions, and utterances," in *Proc. Int. Conf. on Multimodal Interfaces (ICMI)*, Trento, Oct. 2005.

[42] A. Pentland, "Socially aware computation and communication," *IEEE Computer*, vol. 38, pp. 63–70, Mar. 2005.

[43] A. Pentland and A. Madan, "Perception of social interest," in *Proc. IEEE Int. Conf. on Computer Vision, Workshop on Modeling People and Human Interaction (ICCV-PHI)*, Beijing, Oct. 2005.

[44] L. R. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*, Prentice-Hall, 1993.

[45] R.J. 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.

[46] E. Shriberg, "Spontaneous speech: How people really talk and why engineers should care," in *Proc. European Conf. on Speech Communication and Technology (Eurospeech)*, Lisbon, Sep. 2005.

[47] T. Starner and A. Pentland, "Visual recognition of american sign language using HMMs," in *Proc. Int. Work. on Auto. Face and Gesture Recognition (AFGR)*, Zurich, 1995.

[48] R. Stiefelhagen, J. Yang, and A. Waibel, "Modeling focus of attention for meeting indexing based on multiple cues," *IEEE IEEE Trans. on Neural Networks*, vol. 13, no. 4, pp. 928–938, 2002.

[49] R. Stiefelhagen, "Tracking focus of attention in meetings," in *Int. Conf. on Multimodal Interfaces (ICMI)*, Pittsburgh, PA, 2002.

[50] A. Waibel, M. Bett, F. Metze, K. Ries, T. Schaaf, T. Schultz, H. Soltau, H. Yu, and K. Zechner, "Advances in automatic meeting record creation and access," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Salt Lake City, UT, May 2001.

[51] B. Wrede and E. Shriberg, "Spotting hotspots in meetings: Human judgments and prosodic cues," in *Proc. Eurospeech*, Geneva, Sep. 2003.

[52] B. Wrede and E. Shriberg, "The relationship between dialogue acts and hot spots in meetings," in *Proc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, Virgin Islands, Dec. 2003.

[53] D. Zhang, D. Gatica-Perez, S. Bengio, I. McCowan, and G. Lathoud, "Modeling individual and group actions in meetings: a two-layer HMM framework," in *Proc. IEEE CVPR Workshop on Event Mining*, Washington, DC, Jun. 2004.

[54] D. Zhang, D. Gatica-Perez, S. Bengio, I. McCowan, and G. Lathoud, "Multimodal group action clustering in meetings," in *Proc. ACM Int. Conf. on Multimedia, Workshop on Video Surveillance and Sensor Networks (ACM MM-VSSN)*, New York, Oct. 2004.

[55] D. Zhang, D. Gatica-Perez, and S. Bengio, "Semi-supervised meeting event recognition with adapted HMMs," in *Proc. IEEE Int. Conf. on Multimedia (ICME)*, Amsterdam, Jul. 2005.

[56] D. Zhang, D. Gatica-Perez, S. Bengio, and D. Roy, "Learning influence among interacting markov chains," in *Proc. Neural Information Processing Systems (NIPS)*, Vancouver, Dec. 2005.

[57] D. Zhang, D. Gatica-Perez, S. Bengio, and I. McCowan, "Modeling individual and group actions in meetings with layered HMMs," *IEEE Trans. on Multimedia*, vol. 8, no. 3. Jun. 2006

[58] M. Zimmermann, Y. Liu, E. Shriberg, and A. Stolcke, "Toward Joint Segmentation And Classification Of Dialog Acts In Multipary Meetings," in *Proc. Workshop on Machine Learning for Multimodal Interaction (MLMI)*, Edinburgh, Jun. 2005.