Modeling interest in face-to-face conversations from multimodal nonverbal behavior

Daniel Gatica-Perez

June 15, 2009

0.1 Introduction

Many readers can likely recall having seen young children literally jumping off their seat when they meet somebody they specially like. Many readers might also have observed the same children being mesmerized, almost still, when somebody or something catches their full attention. These are examples of interest, a fundamental internal state related to many human processes - including imagination, creativity, and learning - that is known to be revealed by nonverbal behavior expressed through voice, gestures, and facial expressions [23, 10], and that has recently been added to the research agenda on multimodal signal processing for human computing.

Dictionaries define interest as "a state of curiosity or concern about or attention to something: an interest in sports; something, such as a quality, subject, or activity, that evokes this mental state" (The American Heritage Dictionary of the English Language) or as "a feeling that accompanies or causes special attention to an object or class of objects; something that arouses such attention" (Merriam-Webster). In this chapter, which is focused on face-to-face conversations, the term interest is used to designate people's internal states related to the degree of engagement displayed, consciously or not, during social interaction. Such engagement can be the result of many factors, ranging from interest in the theme of a conversation, attraction to the interlocutor, and social rapport. Displays of social interest through nonverbal cues have been widely studied in social psychology and include mimicry [7, 8] (an imitation phenomenon displayed through vocal cues but also via body postures and mannerisms, and facial expressions), elevated displays of speaking and kinesic activity, and higher conversational dynamics. In a conversation, interest can be expressed both as a speaker and as a listener. As a speaker, an interested person often increases both voice and body activity. The case of attraction could also involve persisting gaze. As a listener, an interested person would often show attention, expressed e.g. via intense gaze, diminished body motion, and backchannels. Mimicry would appear while playing both roles. The degree of interest that the members of a dyad or a group collectively display during their interaction could be used to extract important information. This could include inferring whether a brief interaction has been interesting to the participants and segmenting a long interaction (e.g. a group meeting at work) into periods of high and low interest. Interest level categories could therefore be used to index and browse conversations involving oneself, and in some contexts involving others (e.g.

at work) where segments of meetings in which participants of a team were highly engaged could be of interest to other team members who had not had the chance to attend the meeting.

This chapter briefly reviews the existing work on automatic modeling of interest in face-to-face interactions, discussing research involving both dyads and groups, and focuses on discussing the multimodal cues and machine learning models that have been used for detection and recognition of interest and related concepts. The domain is relatively new, and therefore poses a considerable number of research challenges in multimodal signal processing. From a larger perspective, interest is one of many aspects that are currently studied in social computing, the computational analysis of social behavior from sensor data [36, 37, 14].

The rest of the chapter is organized as follows. Section 0.2 summarizes the various computational perspectives related to interest modeling that have been addressed in the literature. Section 0.3 reviews work on conversational interest modeling from audio nonverbal cues. Section 0.4 reviews the emerging work on conversational interest modeling from audiovisual cues. Section 0.5 discusses other research investigating problems related to interest modeling. Finally, Section 0.6 offers some concluding remarks. Parts of the material presented in this chapter have been adapted from [14].

0.2 Perspectives on interest modeling

While other authors have advocated for a distinction between interest and several other related concepts like engagement or attraction [24], given the relatively small number of existing works in this domain, a presentation under a generic term was chosen to facilitate a comparative discussion. The literature on computational modeling of interest in face-to-face conversations can be categorized according to different perspectives (see also Figure 1):

1. *Interaction type.* Dyads, small groups, and large groups have all been analyzed in the literature.

2. *Processing units.* Existing works have considered the units of analysis to be (1) speech utterances by individuals; (2) interaction segments (not necessarily aligned with speech utterances); and (3) whole interactions.

3. *Target tasks.* Depending on the processing units, the target tasks have included (1) classification of pre-segmented speech utterances into a

small set of interest-level classes (e.g. high or low interest); (2) automatic segmentation and classification of interaction segments into interest-level classes; and (3) prediction of concrete, interest-related behavioral outcomes (e.g. mutually interested people exchanging business cards after discussing at a conference), which often results in binary classification tasks. Cases 1 and 2 require manual annotation of interest-level classes, which is commonly derived from first or third-party human judgments. Case 3, on the other hand, can use the interaction outcomes themselves as annotation. In most cases, the occurrence of high interest might be an infrequent situation, which results in imbalanced data sets for learning statistical models.

4. *Single vs. multimodal cues.* Speech (captured by close-talk or distant microphones) is the predominant modality in conversations and has been the most commonly investigated. A few works, however, have studied the possibility of integrating other modalities: vision from cameras, or motion from wearable sensors.

The research reviewed in this chapter is summarized in Table 1. Examples of some of the data used in the discussed methods appear in Fig. 2. The next two sections review the existing work based on the use of single and multiple perceptual modalities, respectively.

0.3 Computing interest from audio cues

Most existing work on automatic interest modeling has focused on the relations between interest (or related concepts) and the speech modality, using both verbal and nonverbal cues. Wrede and Shriberg [43, 44] introduced the notion of hot spots in group meetings, defining it in terms of participants highly involved in a discussion, and relating it to the concept of activation in emotion modeling, i.e., the "strength of a person's disposition to take action" [11]. The authors used data from the International Computer Science Institute (ICSI) Meeting Recording (MR) corpus [21] containing 4- to 8-person conversations, close-talk microphones, and speech utterances as the basic units. In [43], defining a hot spot utterance as one in which a speaker sounded "especially interested, surprised or enthusiastic about what is being said, or he or she could express strong disagreement, amusement, or stress" [44], the authors first developed an annotation scheme that included three

categories of involvement (*amused*, *disagreeing*, and *other*), one *not specially involved* category, and one *I don't know* category, which human annotators used to label utterances based as much as possible on the acoustic information (rather than the content) of each utterance. This study found that human annotators could reliably perceive involvement at the utterance level (a Kappa inter-annotator analysis produced a value of 0.59 in discriminating between involved and non-involved utterances, and lower values for the multi-category case). This work also studied a number of prosodic cues related to the energy of voiced segments and the fundamental frequency (F0) aggregated over speech utterances, computed from individual close-talk microphones. Based on a relatively small number of speech utterances, the authors found that a number of these features (mainly those derived from F0) appear to be discriminating of involved vs. non-involved utterances. No experiments for automatic hot-spot classification from single or multiple features were reported.

In subsequent work [44], the same authors extended their study to analyze the relation between hot spots and dialog acts (DAs), that indicate the function of an utterance (question, statement, backchannels, jokes, acknowledgements, etc.). The study used 32 meetings where the annotation of involvement was done by one annotator continuously listening to a meeting and using the same categories as in [43] (*amused*, *disagreeing*,*other*, and *non-involved*). In this larger corpus, the authors found that a rather small proportion of utterances (about 2%) corresponded to involved utterances, and also found a number of trends between DA categories and involvement categories (e.g., jokes DAs occur often for amused involvement, and backchannels do so for non-involvement).

In a related line of work, Kennedy and Ellis [22] addressed the problem of detecting emphasis or excitement of speech utterances in meetings from prosodic cues, acknowledging that this concept and emotional involvement might be acoustically similar. The authors first asked human annotators to label utterances as emphasized or neutral as they listened to 22 minutes of a 5-person meeting, and found that people could reliably identify emphasized utterances (full agreement across five annotators in 62% of the data, and 4-out-of-5 agreement in 84%), but also that the number of emphasized frames is low (about 15%). The authors later used a very simple approach to measure emphasis based on the assumption that heightened pitch corresponds to emphasis, and using pitch and its aperiodicity computed with the Yin pitch estimator as cues [9], from signals coming from individual close-

talk microphones. A basic pitch model was estimated for each speaker, to take into account each person's pitch distribution, and a threshold-based rule was established to distinguish higher pitch for frames and utterances. After eliminating very short noisy speech segments, the method produced a performance of 24% precision, 73% recall, and 92% accuracy for utterances with high agreement in human judgement of emphasis.

Other existing works can also be related to the detection of high-interest segments of conversations. As one example, Yu et al. [45] also attempted to detect conversational engagement, but used telephone, rather than face-to-face, dyadic conversations for experiments. As another example, Hillard et al. [20] proposed a method to recognize a specific kind of interaction in meetings (agreement vs. disagreement) that is likely related to high interest. Using 7 meetings from the ISCI corpus, the work used speech "spurts" (speech intervals with no pauses greater than 0.5 sec) as processing units, that are to be classified as *positive* or *backchannel* (corresponding to the agreement class), *negative* (the disagreement class), and *other*. On a subset of the data, about 15% of the spurts corresponded to either agreement or disagreement. For classification, both prosodic cues (including pause duration, fundamental frequency, and vowel duration) and word-based features (including the total number of words, and the number of "positive" and "negative" keywords) were used in a learning approach that made use of unlabeled data. In the three-way classification task, the authors found that clean speech transcripts performed the best (which is not surprising given that the manually annotation of spurts took their content into account), and that prosody produced promising performance (with classification accuracy similar to the option of using keywords and noisy ASR transcripts), although fusing ASR transcripts and prosody did not improve performance.

The work by Pentland and collaborators has also dealt with the estimation of interest and related quantities [12, 38, 25, 42, 37], in both dyadic and group cases. One key feature of this line of work is that it has often studied social situations with concrete behavioral outcomes (e.g. people declaring common attraction in a speed dating situation, or people exchanging business cards at a conference as a sign of mutual interest) which substantially reduces the need for third-party annotation of interest. Madan et al. studied a speed-dating dyadic scenario for prediction of attraction (that is, romantic or friendly interest) between different-gender strangers in five-minute encounters [24, 25]. In this scenario, participants interact with several randomly assigned "dates" and introduce each other for a short pe-

riod of time, and privately decide whether they are interested in seeing this person again (labeling their interaction partner as a 'yes' or 'no' for three cases: *romantically attracted*, *interested in friendship*, or *interested in business*). Matches are then found by a third person at the end of the session, when two interaction partners agree on their mutual interest in exchanging contact information. The authors recorded 60 5-minute speed dates with audio-only sensors (directional microphones). Four nonverbal audio cues, dubbed activity level, engagement, stress, and mirroring were extracted [36]. The activity level is the z-scored percentage of speaking time computed over speaking voiced segments. Engagement is the z-scored influence a person has on the turn taking patterns of the others (influence itself is computed with an HMM model). Stress is the z-scored sums of the standard deviations of the mean-scaled energy, fundamental frequency, and spectral entropy of voiced segments. Finally, mirroring is the z-scored frequency of short utterance (less than 1-sec long) exchanges. For the *attracted* category, the authors observed that women's nonverbal cues were correlated to both female and male attraction (*yes*) responses (activity level being the most predictive cue), while men's nonverbal cues had no significant correlation with attraction responses. Other results also showed some other cues to be correlated with the *friendship* or *business* responses. An additional analysis of the results, along with pointers for implementation of the used nonverbal cues, can be found in [42]. Madan et al. also used these cues in different combinations as input to standard classifiers like linear classifiers or Support Vector Machines (SVM), and obtained promising performance (70-80% classification accuracy).

In another dyadic case, Madan and Pentland targeted the prediction of interest-level (*high* vs. *low*) in three-minute conversations between same-gender people discussion about random topics [24, 38]. 20 participants of both genders were first paired with same-gender partners. Each pair participated in 10 consecutive 3-minute conversations, and ranked their interest on a 10-point scale after each encounter. In [24], using the same set of features as for the speed dating case and a linear SVM classifier, the best features could be correctly classify binary interest levels with 74% accuracy for males, whereas different behavior was observed for females and no results were reported for automatic classification.

Pentland et al. have also investigated multi-party scenarios. In early work, Eagle and Pentland investigated the group conversation case, where the interest level in the ongoing conversation was manually introduced by users in a mobile device [12], from which a group interest level could be

inferred via averaging. While the device was designed so that the annotation process would not be over distracting, there is still a cognitive load cost associated to this interactive task.

0.4 Computing interest from multimodal cues

Even though it is known that interest in conversations is displayed through vocal and kinesic nonverbal behavior, few works up to date have studied the use of multiple modalities for interest estimation, using joint data captured by microphones, cameras, or other sensors.

In the context of small group meetings, Gatica-Perez et al. presented in [13] an investigation of the performance of audio-visual cues on discriminating high vs. neutral group interest-level segments, i.e., on estimating single labels for meeting segments, much like hot-spots, using a supervised learning approach that simultaneously produces a temporal segmentation of the meeting and the binary classification of the segments into high or neutral interest-level classes. Experiments were conducted on a subset of the MultiModal Meeting Manager (M4) data corpus [27], consisting of 50 five-minute four-person conversations recorded with three cameras and 12 microphones (including 4 lapels and one 8-microphone array). These meetings were recorded based on turn-taking scripts, but otherwise the participants behavior was reasonably natural with respect to emotional engagement. Regarding human annotation of interest, unlike other works discussed in this chapter [43, 44, 22], which used speech utterances to produce the ground-truth, the work in [13] used interval coding [4], and relied on multiple annotators that continuously watched the meeting and labeled 15-second intervals in a 5-point scale. The ground truth (meeting segments labeled either *neutral interest* or *high-interest*) was produced after an analysis of inter-annotator agreement which showed reasonable agreement, and later used for training and evaluation purposes (about 80% of the frames were labeled as neutral). The investigated nonverbal cues included audio cues derived from lapel microphones (pitch, energy, speaking rate) and from microphone arrays (speech activity estimated by the steered power response phase transform (SRP-PHAT)). Visual nonverbal cues were also extracted for each participant's by computing skin-color blobs motion and location, as a rough proxy for head and body motion and pose. Two Hidden Markov Model (HMM) recognition strategies were investigated [39]: early integration, where all cues were synchronized

and concatenated to form the observation vectors; and multistream HMMs, in which the audio and the visual modalities are used separately to train a single-model HMM, and then both models are fused at the state level to do inference (decoding). Various combinations of audio, visual, and multimodal cues and HMM models were investigated. The performance of automatic segmentation and segment labeling was evaluated at the frame-level based on a convex combination of precision and recall (instead of using a more standard measure similar to the Word Error Rate in speech recognition that might not be meaningful when recognizing binary sequences). Overall, the results were promising (some of the best reported precision/recall combinations were 63/85 and 77/60), and indicated that combining multiple audio cues outperformed the use of individual cues; that audio-only cues outperformed visual-only cues; and that audio-visual fusion brought benefits in some precision/recall conditions, outperforming audio-only cues, but not in others.

In a different scenario, Gips and Pentland investigated the conference case, where large groups of attendees participate and multiple brief conversational exchanges occur [38, 15]. A sensor badge worn by the attendees recorded audio, motion from accelerometers, and proximity to other badges via IR. Additionally, people could bookmark other attendees they had interacted with by pressing a button, in the understanding that the contact details of bookmarked people would be automatically made available after the conference. In this case, the task was to predict for what encounters people bookmark their conversation partner. Two data sets were collected, one involving 113 people in a sponsor conference, and another involving 84 participants recorded six months later. Using a set of 15 basic features derived from the accelerometer and microphone (mean and standard deviation of the amplitude and difference of the raw signals computed over time windows), the authors found that both audio and motion cues were significantly correlated with bookmarks (specially with the standard deviation cues). Using a quadratic classifier and the subset of the six most correlated cues resulted in 82.9% and 74.6 % encounter classification accuracy (*bookmarked* vs. *non-bookmarked*) for each of the two data sets.

0.5 Other concepts related to interest

As discussed in the introduction, there is a clear relation between conversational interest and attention [37]. The automatic estimation of attention could thus be important as a cue for interest modeling. It is known that listeners manifest attention by orienting their gaze to speakers, who also use gaze to indicate whom they address and are interested in interacting with [17]. As pointed out by Knapp and Hall, people "gaze more at people and things perceived as rewarding" and "at those whom they are interpersonally involved" [23] (p. 349 and 351), and this, in the context of conversations, includes people of interest. As two examples of the above, increased gaze often occurs in cases of physical attraction [23], and mutual liking has been reported to be related to gaze in dyadic discussions about controversial topics [5].

Estimating eye gaze in arbitrary conversational situations is a difficult problem given the difficulty in using eye trackers due to sensor setting and image resolution. While some solutions using wearable cameras have started to appear [28], the problem of estimating gaze in conversations has been more often tackled by using head pose as a gaze surrogate. This has generated an increasing body of work [41, 40, 1, 2] that is not reviewed here for space reasons. However, one of the most interesting aspects of current research for attention modeling is the integration of audio-visual information to estimate visual attention in conversations. In the context of group conversations, the works by Otsuka et al. [32, 33, 34] and Ba and Odobez [3] stand out as examples of models of the interplay between speaking activity and visual attention. Otsuka et al. proposed a Dynamic Bayesian Network (DBN) approach which jointly infers the gaze pattern for multiple participants and the conversational gaze regime responsible for specific speaking activity and gaze patterns (e.g. all participants converging onto one person, or two people looking at each other) [32]. Gaze was approximated by head pose, observed either through magnetic head trackers attached to each participant [32], or automatically estimated from video [33]. Otsuka et al. later extended their model in an attempt to respond to the 'who responds to whom, when, and how' questions in a joint manner [34]. With somewhat similar hypotheses, Ba and Odobez proposed a DBN model for the estimation of the joint attention of group participants by using people's speaking activity as a contextual cue, defining a prior distribution on the potential visual attention targets of each participant [3]. This observation resulted in improved visual attention recog-

nition from head orientation automatically estimated from a single camera on a subset of the Augmented Multiparty Interaction (AMI) meeting corpus, a publicly available meeting collection with audio, video, slides, whiteboard, and handwritten note recordings [6] (also see Fig. 2.)

Listening is a second conversational construct clearly related to attention. Listening is in principle a multimodal phenomenon, and some works have started to investigate potential computational models. Heylen et al. [19] presented research towards building a Sensitive Artificial Listener, based on the manual annotation of basic nonverbal behaviors displayed by listeners in group conversations, including gaze, head movements, and facial expressions extracted from the AMI corpus.

Finally, there is recent body of work by Pentland et al. that is beginning to investigate the recognition of longer-term phenomena in real-life organizational settings involving large groups. These organizational behavior phenomena, although clearly distinct from the concept of interest as discussed here, are nevertheless related to the aggregation of states of human interest over time. More specifically, this research has investigated the correlation between automatically extracted nonverbal cues and concepts like workload, job satisfaction, and productivity in banks [16, 30] and hospitals [31], as well as team performance and individual networking performance in professional gatherings [29]. In all cases, sensing is done through sociometers, i.e., wearable devices capable of measuring a number of nonverbal cues including physical proximity, actual face-to-face interaction, body motion, and audio. Overall, this is an example of the complex sociotechnical research that will continue to appear in the future regarding social behavior analysis, and that might make use of interest or similar concepts as mid-level representations for higher social inference.

0.6 Concluding remarks

This chapter has presented a concise review of representative work related to interest modeling in face-to-face conversations from multimodal nonverbal behavior. The discussion in the previous sections highlights the facts that this domain is still emerging, and that many opportunities lie ahead regarding the study of other automatic nonverbal cues that are better correlated with displays of interest (importantly, from the visual modality), the design of new multimodal integration strategies, and the application of cues and models to

other social scenarios. The improvement of the technological means to record real interaction, both in multi-sensor spaces and with wearable devices, are opening the possibility to analyze multiple social situations where interest emerges and correlates with concrete social outcomes, and also to develop new applications related to self-assessment and group awareness. Given the increasing attention in signal processing and machine learning with respect to social interaction analysis, there is much to look forward to in the future regarding advances on computational modeling of social interest and related concepts.

0.7 Acknowledgements

The author thanks the support of the Swiss National Center of Competence in Research on Interactive Multimodal Information Management (IM2), and of the EC project Augmented Multi-Party Interaction with Distant Access (AMIDA). He also thanks Nelson Morgan (ICSI) and Sandy Pentland (MIT) for giving permission to reproduce some of the pictures presented in this Chapter.

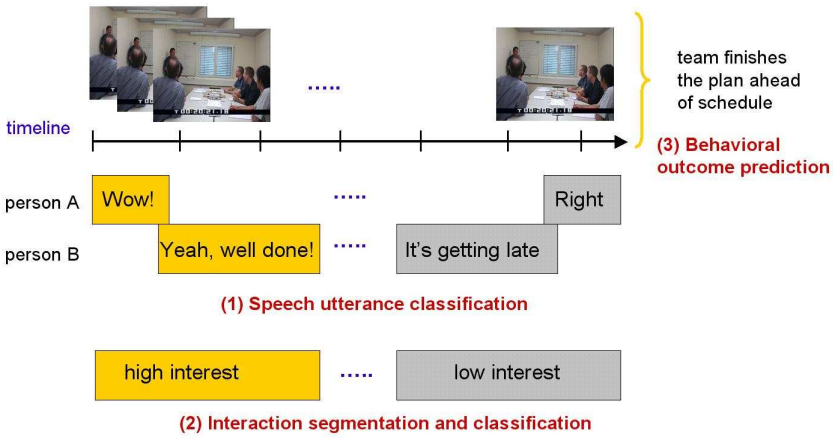
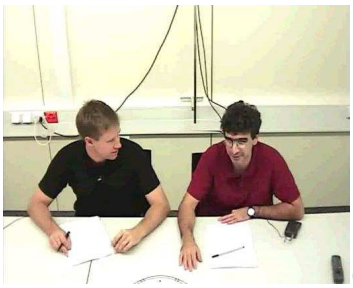


Figure 1: Interest modeling tasks for an interacting group: (1) classification of pre-segmented individual speech utterances as corresponding to high interest (in orange) or low interest (in gray); (2) segmentation and classification of meeting segments as high or low interest (orange or gray, resp.) ; (3) prediction of behavioral outcomes that relate to interest level (orange bracket in the example).


(a)

(b)

(c)

(d)


(e)

Figure 2: Scenarios and data for estimation of interest in face-to-face conversations: (a) ICSI Meeting Recording corpus [43]. (b) MIT speed dating corpus [25]. (c) M4 (MultiModal Meeting Manager) corpus [13]. (d) MIT conference corpus [15]. (e) AMI (Augmented Multi-Party Interaction) corpus [6]. All pictures are reproduced with permission.

Ref.	Scenario and Task	Data	NVB
[43]	4 to 8-person meetings; relation between prosodic cues and hot spots for utterances	ICSI MR corpus; 88 speech utterances from 13 meetings	A
[44]	5 to 8-person meetings; relation between hot spots and dialog acts for utterances	ICSI MR corpus; 32 meetings; approx. 32 h	A
[22]	5-person meeting; classification of utterances as emphasized/neutral	ICSI MR corpus; 1 meeting; 22 min; 861 utterances	A
[20]	5 to 8-person meetings; classification of speech "spurts" as agreement/disagreement	ICSI MR corpus; 7 meetings	A
[25]	dyadic speed dates; prediction of matches of mutually interested people	MIT data; 60 5-minute meetings	A
[24]	dyadic interaction; classification of short conversations as high/low interest	MIT data; 100 3-minute conversations	A
[12]	9-person meetings; manual annotation of individual interest level	MIT data; 1 one-hour meeting	A
[13]	4-person meetings; segmentation+classification of high/neutral group interest	M4 corpus; 50 5-min meetings;	A,V
[15]	113 and 84 conference attendees; bookmarking of dyadic encounters (high interest)	MIT data; 1 day (approx. 8 hours) in each case	A,M

Table 1: Research on automatic modeling of conversational interest. The investigated nonverbal behavior includes audio (A), video (V), and body motion (M) cues.

Bibliography

[1] 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.

[2] 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.

[3] S.O. Ba and J.M. Odobez, "Multi-party Focus of Attention Recognition in Meetings from Head Pose and Multimodal Contextual Cues," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Las Vegas, Mar. 2008.

[4] R. Bakeman and J. M. Gottman, *Observing Interaction: An Introduction to Sequential Analysis*, Cambridge University Press, 1997.

[5] F. J. Bernieri, J. S. Gills, J.M Davis, and J.E. Grahe, "Dyad rapport and the accuracy of its judgment across situations: a lens model analysis," *Journal of Personality and Social Psychology*, Vol. 71, pp. 110-129, 1996.

[6] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec,V. Karaiskos, W. Kraaij, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, I. McCowan, W. Post, D. Reidsma, and P. Wellner, "The AMI meeting corpus: A pre-announcement," in *Proc. Workshop on Machine Learning for Multimodal Interaction (MLMI)*, Edinburgh, Jul. 2005.

[7] T. L. Chartrand and J. A. Bargh, "The chameleon effect: the perception-behavior link and social interaction," *Journal of Personality and Social Psychology*, Vol. 76, No. 6, pp. 893-910, Jun. 1999.

[8] T.L Chartrand, W. Maddux, and J. Lakin, "Beyond the perception-behavior link: The ubiquitous utility and motivational moderators of nonconscious mimicry," in R. Hassin, J. Uleman, and J.A. Bargh (Eds.), *The New Unconscious*, Oxford Univ. Press, 2005.

[9] A. de Cheveigne and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *Journal of the Acoustic Society of America*, 2001.

[10] Y. S. Choi, H. M. Gray, and N. Ambady, "The glimpsed world: unintended communication and unintended perception," In R.H. Hassin, J.S. Uleman, J.A. Bargh (eds.), *The new unconscious*, Oxford University Press, 2005.

[11] 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.

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

[13] 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.

[14] D. Gatica-Perez, "Automatic Nonverbal Analysis of Social Interaction in Small Groups: a Review," *Image and Vision Computing, Special Issue on Naturalistic Human Behavior*, in press.

[15] J. Gips and A. Pentland, "Mapping Human Networks," in *Proc. IEEE Int. Conf. on Pervasive Computing and Communications*, Pisa, Mar. 2006.

[16] P. A. Gloor, D. Oster, J. Putzke, K. Fischback, D. Schoder, K. Ara, T. J. Kim, R. Laubacher, A. Mohan, D. Olguin Olguin, A. Pentland, and B. N. Waber, "Studying Microscopic Peer-to-Peer Communication Patterns," in *Proc. Americas Conference on Information Systems*, Keystone, Aug. 2007.

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

[18] R.H. Hassin, J.S. Uleman, J.A. Bargh (eds.) *The new unconscious*, Oxford University Press, 2005.

[19] D. Heylen, A. Nijholt and M. Poel, "Generating Nonverbal Signals for a Sensitive Artificial Listener," in *Proc. COST 2102 Workshop on Verbal and Nonverbal Communication Behaviours*, Vietri sul Mare, Mar. 2007.

[20] 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.

[21] A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, and C. Wooters, "The ICSI meeting corpus," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Hong-Kong, Apr. 2003.

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

[23] M. L. Knapp and J. A. Hall, *Nonverbal Communication in Human Interaction,* 6th ed., Wadsworth Publishing, 2005.

[24] A. Madan, "Thin Slices of Interest," Master's Thesis, Massachusetts Institute of Technology, 2005.

[25] A. Madan, R. Caneel and A. Pentland, "Voices of Attraction," in *Proc. Int. Conf. on Augmented Cognition (AC-HCI)*, Las Vegas, Jul. 2005.

[26] V. Manusov and M. L. Patterson (eds.), *The SAGE Handbook of Nonverbal Communication*, Sage Publications, 2006.

[27] 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.

[28] B. Noris, K. Benmachiche and A. Billard, "Calibration-Free Eye Gaze Direction Detection with Gaussian Processes," in *Proc. Int. Conf. on Computer Vision Theory and Applications*, 2008.

[29] D. Olguin Olguin, and A. Pentland, "Social Sensors for Automatic Data Collection," in *Proc. Americas Conference on Information Systems*, Toronto, Aug. 2008.

[30] D. Olguin Olguin, B. Waber, T. Kim, A. Mohan, K. Ara, and A. Pentland, "Sensible Organizations: Technology and Methodology for Automatically Measuring Organizational Behavior," *IEEE Trans. on Systems, Man, and Cybernetics-Part B.* Vol. 39, No. 1, Feb. 2009.

[31] D. Olguin Olguin, P. A. Gloor and A. Pentland, "Capturing Individual and Group Behavior with Wearable Sensors," in *Proc. AAAI Spring Symposium on Human Behavior Modeling*, Stanford, Mar. 2009.

[32] 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.

[33] K. Otsuka, J. Yamato, Y. Takemae, and H. Murase, "Conversation scene analysis with dynamic Bayesian network based on visual head tracking," in *Proc. IEEE Int. Conf. on Multimedia (ICME)*, Toronto, Jul. 2006.

[34] K. Otsuka, J. Yamato, and H. Sawada, "Automatic inference of cross-modal nonverbal interactions in multiparty conversations," in *Proc. Int. Conf. on Multimodal Interfaces (ICMI)*, Nagoya, Nov. 2007.

[35] K. Otsuka, S. Araki, K. Ishizuka, M. Fujimoto, M. Heinrich, and J. Yamato, "A Realtime Multimodal System for Analyzing Group Meetings by Combining Face Pose Tracking and Speaker Diarization," in *Proc. Int. Conf. on Multimodal Interfaces (ICMI)*, Chania, Oct. 2008.

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

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

[38] 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.

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

[40] 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.

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

[42] W. T. Stoltzman, "Toward a Social Signaling Framework: Activity and Emphasis in Speech," Master's Thesis, Massachusetts Institute of Technology, Sep. 2006.

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

[44] 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.

[45] C. Yu, P. Aoki, and A. Woodruff, "Detecting User Engagement in Everyday Conversations," in *Proc. ICSLP*, 2004.