

VACE Multimodal Meeting Corpus

Lei Chen¹, R. Travis Rose², Fey Parrill³, Xu Han⁴, Jilin Tu⁴, Zhongqiang Huang¹, Mary Harper¹, Francis Quek², David McNeill³, Ronald Tuttle⁵, and Thomas Huang⁴

¹ School of Electrical Engineering, Purdue University, West Lafayette IN,
chenl@ecn.purdue.edu

² CHCI, Department of Computer Science, Virginia Tech, Blacksburg, VA

³ Department of Psychology, University of Chicago, Chicago, IL

⁴ Beckman Institute, University of Illinois Urbana Champaign, Urbana, IL

⁵ Air Force Institute of Technology, Dayton, OH

Abstract. In this paper, we report on the infrastructure we have developed to support our research on multimodal cues for understanding meetings. With our focus on multimodality, we investigate the interaction among speech, gesture, posture, and gaze in meetings. For this purpose, a high quality multimodal corpus is being produced.

1 Introduction

Meetings are gatherings of people for the purpose of communicating with each other in order to plan, resolve conflicts, negotiate, or collaborate. Recently, there has been increasing interest in the analysis of meetings based on audio and video recordings. In a meeting, the participants employ a variety of multimodal channels, including speech, gaze, gesture, and posture, to communicate with each other. Because of the communicative impact of these information sources, they can be used, for example, to analyze the structure of a meeting or determine the social dynamics among the participants. To better understand human interactions in a meeting, it is important to jointly examine both verbal and non-verbal cues produced by meeting participants.

Meetings also challenge current audio and video processing technologies. For example, there is a higher percentage of cross-talk among audio channels in a six party meeting than in a two party dialog, and this could reduce the accuracy of current speech recognizers. In a meeting setting, there may not be the ideal video image size or angle when attempting to recognize a face. Therefore, recorded meetings can push forward multimodal signal processing technologies.

To investigate meetings, several corpora have already been collected, including the **ISL** audio corpus [1] from Interactive Systems Laboratory (ISL) of CMU, the **ICSI** audio corpus [2], the **NIST** audio-visual corpus [3], and the **MM4** audio-visual corpus [4] from Multimodal Meeting (MM4) project in Europe. Using these existing meeting data resources, a large body of research has already been conducted, including automatic transcription of meetings [5], emotion detection [6], attention state [7], action tracking [8, 4], speaker identification [9],

speech segmentation [10, 11], and disfluency detection [12]. Most of this research has focused on low level processing (e.g., voice activity detection, speech recognition) or on elementary events and states. Research on the structure of a meeting or the dynamic interplay among participants in a meeting is only beginning to emerge. McCowan et al. [4] used low level audio and visual information to segment a meeting into meeting actions using an HMM approach.

Our research focuses not only on low level multimodal signal processing, but also on high level meeting event interpretation. In particular, we use low-level multimodal cues to interpret the high-level events related to meeting structure. To carry out this work, we require high quality multimodal data to jointly support multimodal data processing, meeting analysis and coding, as well as automatic event detection algorithms. Hence, we have been collecting a new meeting corpus supported by the ARDA VACE-II program. This collection will be called the VACE meeting corpus in this paper.

In rest of this paper, we will discuss our meeting corpus collection approach and briefly discuss some of the research that this effort is enabling. Section 2 describes the ongoing multimodal meeting corpus collection effort. Section 3 describes audio and video data processing algorithms needed for corpus production. Section 4 describes research that this corpus enables.

2 Meeting Data Collection

Our multimodal meeting data collection effort is depicted schematically in Figure 1. We next discuss three important aspects of this meeting data collection effort: (1) meeting room setup, (2) elicitation experiment design, and (3) data processing.

2.1 Multimodal Meeting Room Setup

Under this research effort, Air Force Institute of Technology (AFIT) modified a lecture room to collect multimodal, time-synchronized audio, video, and motion data. In the middle of the room, up to 8 participants can sit around a small, rectangular conference table. An overhead rail system permits the data acquisition technician to position the 10 Canon GL2 camcorders in any configuration required to capture all participants by at least two of the 10 camcorders. Using S-video transfer, 10 Panasonic AG-DV2500 recorders capture video data from the camcorders. The rail system also supports the 9 Vicon MCam2 near-IR cameras and are driven by the Vicon V8i Data Station. The Vicon system records temporal position data. For audio recording, we utilized a setup similar to the ICSI and NIST meeting room. In particular, participants wear Countryman ISOMAX Earset wireless microphones to record their individual sound tracks. Table-mounted wired microphones are used to record the audio of all participants (two to six XLR-3M connector microphones configured for the number of participants and scenario, including two cardioid Shure MX412 D/C microphones and several types of low-profile boundary microphones (two hemispherical polar

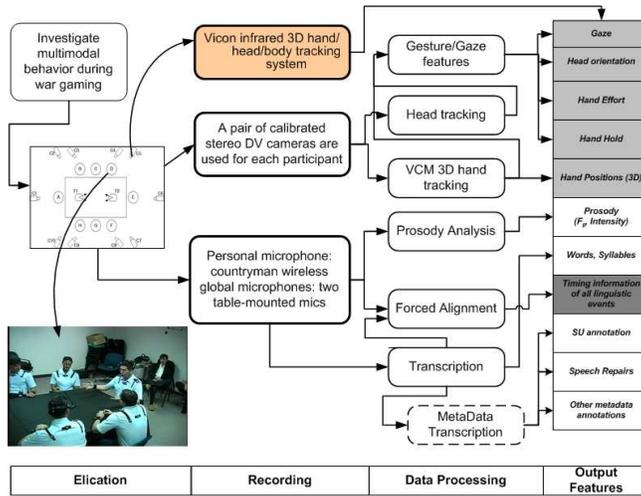


Fig. 1. VACE Corpus Production

pattern Crown PZM-6D, one omni-directional Audio Technica AT841a, and one four-channel cardioid Audio Technica AT854R). A TASCAM MX-2424 records the sound tracks from both the wireless and wired microphones.

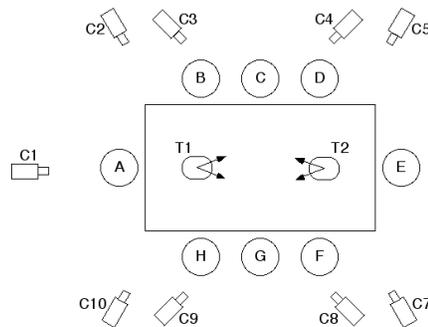


Fig. 2. The camera configuration used for the VACE meeting corpus

There are some significant differences between our video recording setup and those used by previous efforts. For example, in the NIST and MM4 collections, because stereo camera views are not used to record each participant, only 2D tracking results can be obtained. For the VACE corpus, each participant is recorded with a stereo calibrated camera pair. Ten video cameras are placed facing different participants seated around the table as shown in Figure 2.

To obtain the 3D tracking of 6 participants, 12 stereo camera pairs are setup to ensure that each participant is recorded by at least 2 cameras. This is important because we wish to accurately track head, torso and hand positions in 3D. We also utilize the Vicon system to obtain more accurate tracking results to inform subsequent coding efforts, while also providing ground truth for our video-tracking algorithms.

2.2 Elicitation Experiment

The VACE meeting corpus involves meetings based on wargame scenarios and military exercises. We have selected this domain because the planning activity spans multiple military functional disciplines, the mission objectives are defined, the hierarchical relationships are known, and there is an underpinning doctrine associated with the planning activity. Doctrine-based planning meetings draw upon tremendous expertise in scenario construction and implementation. Wargames and military exercises provide real-world scenarios requiring input from all functionals for plan construction and decision making. This elicits rich multimodal behavior from participants, while permitting us to produce high-confidence coding of the meeting behavior. Examples of scenarios include planning a Delta II rocket launch, humanitarian assistance, foreign material exploitation, and scholarship assignment.

2.3 Multimodal Data Processing

After obtaining the audio and video recordings, we must process the data to obtain features to assist researchers with their coding efforts or to train and evaluate automatic event detection algorithms. The computer vision researchers on our team from University of Illinois and Virginia Tech focus on video-based tracking, in particular, body torso, head, and hand tracking. The VCM tracking approach is to obtain 3D positions of the hands, which is important for obtaining a wide variety of gesture features, such as gesture hold and velocity. The change in position of the head, torso, and hands provide important cues for the analysis of the meetings. Researchers on our team from Purdue handle the audio processing of the meetings. More detail on video and audio processing appear in the next section.

Our meeting room corpus contains time synchronized audio and video recordings, features derived by the visual trackers and Vicon tracker, audio features such as pitch tracking and duration of words, and coding markups. Details on the data types appear in an organized fashion in Table 1.

3 Multimodal Data Processing

3.1 Visual Tracking

Torso Tracking Vision-based human body tracking is a very difficult problem. Given that joint-angle is a natural and complete way to describe human

Table 1. Composition of the VACE corpus

Video	MPEG-4 video from 10 cameras
Audio	AIFF audio from all microphones
Vicon	3D positions of head, shoulders, torsos, and hands
Visual-Tracking	head pose, torso configuration, and hand position
Audio Processing	speech segments, transcripts, alignments
Prosody	pitch, word and phone duration, energy, etc.
Gaze	gaze target estimation
Gesture	gesture phase/phrase semiotic gesture coding, e.g., <i>deictic</i> , <i>iconics</i>
MetaData	language metadata, e.g., sentence boundaries, speech repairs, floor control change

body motion and human joint-angles are not independent, we have been investigating an approach to learn these latent constraints and then use them for articulated body tracking. After learning the constraints as potential functions, belief propagation is used to find the MAP of the body configuration on the Markov Random Field (MRF) to achieve globally optimal tracking. When tested on the VACE meeting data, we have obtained satisfactory tracking results. See Figure 3(a) for an example of torso tracking.

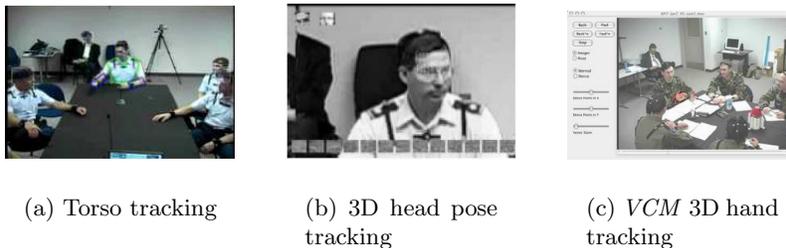


Fig. 3. Visual tracking of torso, head, and hands

Head Pose Tracking For the video analysis of human interactions, the head pose of the person being analyzed is very important for determining gaze direction and the person being spoken to. In our meeting scenario, the resolution of a face is usually low. We therefore have developed a hybrid 2D/3D head pose tracking framework. In this framework, a 2D head position tracking algorithm [13] tracks the head location and determines a coarse head orientation (such as the

frontal view of the face, the side view and the rear view of the head) based on the appearance at that time. Once the frontal view of a face is detected, a 3D head pose tracking algorithm [14] is activated to track the 3D head orientation. For 2D head tracking, we have developed a meanshift tracking algorithm with an online updating appearance generative mixture model. When doing meanshift tracking, our algorithm updates the appearance histogram online based on some key features acquired before the tracking, allowing it to be more accurate and robust. The coarse head pose (such as frontal face) can be inferred by simply checking the generative appearance model parameters. The 3D head pose tracking algorithm acquires the facial texture from the video based on the 3D face model. The appearance likelihood is modelled by an incremental PCA subspace. And the 3D head pose is inferred using an annealed particle filtering technique. An example of a tracking result can be found in Figure 3(b).

Hand Tracking In order to interpret gestures used by participants, exact 3D hand positions are obtained using hand tracking algorithms developed by researchers in the Vislab [15, 16]. See [17] for details on the Vector Coherence Mapping (VCM) approach that is being used. The algorithm is currently being ported to the Apple G5 platform with parallel processors in order to address the challenge of tracking multiple participants in meetings. An example of a tracking result can be found in Figure 3(c).

3.2 Audio Processing

A meeting involves multiple time synchronized audio channels, which increases the workload for transcribers [18]. Our goal is to produce good quality transcriptions that are time aligned with the audio and visual channels, so that the words can be synchronized with video features to support coding efforts and to extract prosodic and visual features needed by our automatic meeting event detection algorithms. For transcriptions, we utilize the Quick Transcription (QTR) methodology developed by LDC for the 2004 NIST Meeting Recognition Evaluations.

Automatic Pre-Segmentation Since in meetings typically one speaker is speaking at any given time, the resulting audio files contain significant portions of audio that do not require transcription. Hence, if each channel of audio is automatically segmented into transcribable and non-transcribable regions, the transcribers only need to focus on the smaller pre-identified regions of speech, lowering the cognitive burden significantly compared with handling a large undifferentiated stream of audio. We perform audio segmentation based on the close-talking audio recordings using a novel automatic multi-step segmentation [19]. The first step involves silence detection utilizing pitch and energy, followed by BIC-based Viterbi segmentation and energy based clustering. Information from each channel is employed to provide a rough preliminary segmentation. The second step makes use of the segment information obtained in the first step to train

a Gaussian mixture model for each speech activity category, followed by decoding to refine the segmentation. A final post-processing step is applied to remove short segments and pad silence to speech segments.

Meeting Transcription Tools Meeting audio includes multi-channel recordings with substantial cross-talk among the audio channels. These two aspects make the transcription process quite different from those previously developed to support the transcription of monologs and dialogues. For meetings, the transcribers must utilize many audio channels and often jump back and forth among the channels to support transcription and coding efforts [18]. There are many linguistic annotation tools currently available [20]; however, most of these tools were designed for monologs and dialogues. To support multi-channel audio annotation, researchers have attempted to either tailor currently available tools or design new tools specific for meeting transcription and annotation, e.g., the modification of Transcriber [21] for meeting transcription by ICSI (<http://www.icsi.berkeley.edu/Speech/mr/channeltrans.html>), iTranscriber by ICSI, and XTrans by LDC.

For our efforts, we have designed a Praat [22] extension package to support multi-channel audio annotation. We chose the Praat tool for the following reasons: 1) it is a widely used speech analysis and transcription tool available for almost any platform, 2) it is easy to use, 3) *long sound* supports the quick loading of multiple audio files, and 4) it has a built-in script language for implementation future extensions. We added two new menu options to the interface: the first supports batch loading of all the audio files associated with a meeting, and the second enables transcribers to switch easily among audio files.

Improved Forced Alignment Forced alignment is used to obtain the starting and ending time of the words and phones in the audio. Since such timing information is widely used for multimodal feature fusion, we have investigated factors for achieving accurate alignments. Based on a systematic study [23], we have found more accurate forced alignments can be obtained by having transcribers directly transcribe pre-identified segments of speech and by using sufficiently trained, genre matched speech recognizers to produce the alignments. For meeting room alignments, we utilize ISIP's ASR system with a triphone acoustic model trained from more than 60 hours long spontaneous speech data [24] to force align transcriptions provided for segments of speech identified in the pre-segmentation step. The alignment performance given this setup on VACE meeting data is quite satisfactory.

4 Meeting Interaction Analysis Research

The VACE meeting corpus enables the analysis of meeting interactions at a number of different levels. Using the visualization and annotation tool Macvissta, developed by researchers at Virginia Tech, the features extracted from the

recorded video and audio can be displayed to support psycholinguistic coding efforts at University of Chicago and Purdue. Some annotations of interest to our team include: F-formation [25], dominant speaker, structural events (sentence boundary, interruption point), and floor control challenges and change. Given the data and annotations in this corpus, we will then carry out measurement studies to investigate how visual and verbal cues combine to predict events such as sentence or topic boundaries, interruption points in a speech repair, or floor control changes. With the rich set of features and annotations, we will also develop data-driven models for meeting room event detection along the lines of our research on multimodal models for detecting sentence boundaries [26].

In the rest of this section, we will first describe the MacVissta tool that supports visualization and annotation of our multimodal corpus. Then we will provide more detail on the psycholinguistic annotations we are producing. Finally, we will discuss a new group research effort involving multimodal cues governing floor control.

4.1 Visualization Tool and Coding Efforts

Visualization of visual and verbal activities is an important first step for developing a better understanding of how these modalities interact in human communication. The ability to add annotations of important verbal and visual events further enriches this data. For example, annotation of gaze and gesture activities is important for developing a better understanding of those activities in floor control. Hence, the availability of a high quality, flexible multimodal visualization/annotation tool is quite important. To give a complete display of a meeting, we need to display the multimodal signals and annotations of all participants. The Vissta tool developed for multimodal dialog data has been recently ported to the Mac OS X while being adapted for meeting room data [27]. Currently the tool supports showing multiple angle view videos, as shown in Figure 4. This tool can display transcriptions and visual features, together with the spoken transcripts and a wide variety annotations. It has been widely used by our team and is continually refined based on their feedback.

Using MacVissta, researchers at the University of Chicago are currently focusing on annotating gesture and gaze patterns in meetings. Gesture onset, offset, and stroke are coded, as well as the semiotic properties of the gesture as a whole, in relation to the accompanying speech. Because gesture is believed to be as relevant to a person's communicative behavior as speech, by coding gesture, we are attempting to capture this behavior in its totality. In addition, gaze is coded for each speaker in terms of the object of that gaze (who or what gaze is directed at) for each moment. Instances of shared gaze (or "F-formations") are then extractable from the transcript, which can inform a turn-taking analysis. More fine-grained analyses include the coding of mimicry (in both gesture and speech), and the tracking of lexical co-reference and discourse cohesion, which permits the analyst to capture moments where speakers are negotiating how to refer to an object or event. These moments appear to be correlated with shared gaze.

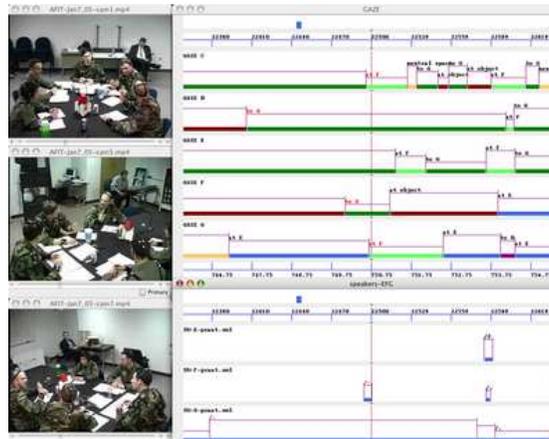


Fig. 4. A snapshot of the MacVissta multimodal analysis tool with multiple videos shown on the left and gaze annotations shown on the right.

4.2 Preliminary Investigation of Floor Control

Participants in a conversation have different roles in the conversation. The dominant participant, who is responsible for moving the conversation forward, is said to have control of the floor. In meetings, participants compete and cooperate for floor control distribution. This information is critical for fully understanding the structure of a meeting. Investigation of floor control has attracted researchers from linguistics, psychology, and artificial intelligence for decades [28, 29]. Non-verbal behaviors play an important role in coordinating turn-taking and the organization of discourse [30, 29]. Gaze is an important cue for floor control [31, 32]. Near the end of an utterance, the dominant speaker may gaze at an interlocutor to transfer the control [33]. The interlocutor who is gazed at has the advantage for taking the floor because he knows that the speaker is conceding the floor to him/her. Kendon [34] observed that utterances that terminate without gaze cues more frequently had delayed listener response. Gesture and body posture are also important cues for floor control decisions. Posture shifts tend to occur during an utterance’s beginning and ending, as well as at various discourse boundaries [35]. Pointing at a participant is a direct way to assign the control of floor to that person.

In the rest of this section, we examine an excerpt of a planning meeting on the testing of a foreign weapon from the VACE corpus to highlight the importance of investigating multimodal cues for meeting room floor control investigations. From the recording done on January 7th 2005, we have extracted a short segment for processing. Following [10], we used the forced alignment of transcribed words to obtain speaker activity profiles shown in Figure 5. Using the seating chart in Figure 2, the five participants are denoted C, D, E, F, and G. We segment the audio for each speaker into bins 100 msec in length, and if the participant speaks

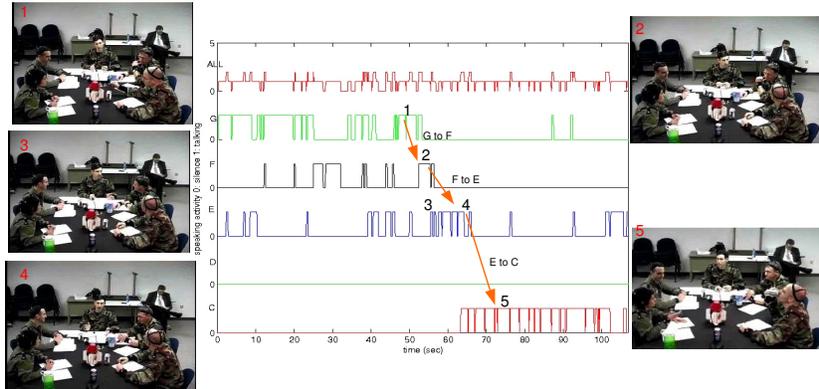


Fig. 5. Multimodal activities of participants in a meeting

at any time during this interval, the associated activity value is 1, otherwise 0. In Figure 5, from top to bottom, the speech activities are displayed as follows: all participants, G, F, E, D, and C. From the speech activity plot, it is easy to see smooth floor control transitions between 55 to 70 seconds, going from G to F to E and finally to C. Important additional details of this transition can be observed from video cues as can be seen in the key frames in Figure 5. In each video frame, the five participants C, D, E, F and G are arrayed from the left bottom corner to the right bottom corner. E, the coordinator of the meeting, sits in the middle of the other participants. C and D were to his right and F and G are to his left. The key frames shown highlight the importance of gesture and gaze: (1) G initially holds the floor; (2) F provides G with some additional information (notice that G turns his head to F); (3) E grabs the floor from G and F. He first uses his left hand to point to F and G, and then (4) turns his head to the right and used right hand to point to C and D to give them the floor; (5) C takes control of the floor and continues speaking. We are currently defining an annotation specification for speaker floor change.

5 Conclusions

In this paper, we have reported on the infrastructure we have developed to support our research on multimodal cues for understanding meetings. With our focus on multimodality, we investigate the interaction among speech, gesture, posture, and gaze in meetings. For this purpose, a high quality multimodal corpus is being produced. Each participant is recorded with a pair of stereo calibrated camera pairs so that 3D body tracking can be done. Also an advanced motion tracking system is utilized to provide ground truth. From recorded audio and video, research on audio processing and video tracking focus on improving quality of low features that support higher level annotation and modeling efforts.

6 Acknowledgments

We thank all team members for efforts to produce the VACE corpus: Dr. Yingen Xiong, Ying Qiao, Bing Fang, and Dulan Wathugala from Virginia Tech, Dr. Sue Duncan, Irene Kimbara, Matt Heinrich, Haleema Welji, Whitney Goodrich, and Alexia Galati from University of Chicago, Dr. David Bunker, Jim Walker, Kevin Pope, Jeff Sitler from AFIT. This research has been supported by the Advanced Research and Development Activity ARDA VACEII grant 665661: *From Video to Information: Cross-Model Analysis of Planning Meetings*. Part of this work was carried out while the seventh author was on leave at NSF. Any opinions, findings, and conclusions expressed in this paper are those of the authors and do not necessarily reflect the views of NSF or ARDA.

References

- [1] Burger, S., MacLaren, V., Yu, H.: The ISL meeting corpus: The impact of meeting type on speech type. In: Proc. of Int. Conf. on Spoken Language Processing (ICSLP). (2002)
- [2] Morgan, N., et al.: Meetings about meetings: Research at ICSI on speech in multiparty conversations. In: Proc. of ICASSP. Volume 4., Hong Kong, Hong Kong (2003) 740–743
- [3] Garofolo, J., Laprum, C., Michel, M., Stanford, V., Tabassi, E.: The NIST Meeting Room Pilot Corpus. In: Proc. of Language Resource and Evaluation Conference. (2004)
- [4] McCowan, I., Gatica-Perez, D., Bengio, S., Lathoud, G., Barnard, M., Zhang, D.: Automatic analysis of multimodal group actions in meetings. IEEE Trans. on Pattern Analysis and Machine Intelligence **27** (2005) 305–317
- [5] Schultz, T., Waibel, A., et al.: The ISL meeting room system. In: Proceedings of the Workshop on Hands-Free Speech Communication, Kyoto Japan (2001)
- [6] Polzin, T.S., Waibel, A.: Detecting emotions in speech. In: Proceedings of the CMC. (1998)
- [7] Stiefelhagen, R.: Tracking focus of attention in meetings. In: Proc. of Int. Conf. on Multimodal Interface (ICMI), Pittsburg, PA (2002)
- [8] Alfred, D., Renals, S.: Dynamic bayesian networks for meeting structuring. In: Proc. of ICASSP. Volume 5., Montreal, Que, Canada (2004) 629–632
- [9] Gatica-Perez, D., Lathoud, G., McCowan, I., Odobez, J., Moore, D.: Audio-visual speaker tracking with importance particle filters. In: Proc. of Int. Conf. on Image Processing (ICIP). Volume 3., Barcelona, Spain (2003) 25–28
- [10] Renals, S., Ellis, D.: Audio information access from meeting rooms. In: Proc. of ICASSP. Volume 4., Hong Kong, Hong Kong (2003) 744–747
- [11] Ajmera, J., Lathoud, G., McCowan, I.: Clustering and segmenting speakers and their locations in meetings. In: Proc. of ICASSP. Volume 1., Montreal, Que, Canada (2004) 605–608
- [12] Moore, D., McCowan, I.: Microphone array speech recognition: Experiments on overlapping speech in meetings. In: Proc. of ICASSP. Volume 5., Hong Kong, Hong Kong (2003) 497–500
- [13] Tu, J., Huang, T.S.: Online updating appearance generative mixture model for mean-shift tracking. In: Proc. of Int. Conf. on Computer Vision (ICCV). (2005)

- [14] Tu, J., Tao, H., Forsyth, D., Huang, T.S.: Accurate head pose tracking in low resolution video. In: Proc. of Int. Conf. on Computer Vision (ICCV). (2005)
- [15] Quek, F., Bryll, R., Ma, X.F.: A parallel algorithm for dynamic gesture tracking. In: ICCV Workshop on RATFG-RTS, Gorfu, Greece (1999)
- [16] Bryll, R.: A Robust Agent-Based Gesture Tracking System. PhD thesis, Wright State University (2004)
- [17] Quek, F., Bryll, R., Qiao, Y., Rose, T.: Vector coherence mapping: Motion field extraction by exploiting multiple coherences. CVIU special issue on Spatial Coherence in Visual Motion Analysis (submitted) (2005)
- [18] Strassel, S., Glenn, M.: Shared linguistic resources for human language technology in the meeting domain. In: Proceedings of ICASSP 2004 Meeting Workshop. (2004)
- [19] Huang, Z., Harper, M.: Speech and non-speech detection in meeting audio for transcription. In: EuroSpeech 2005 (submitted). (2005)
- [20] LDC: (Linguistic annotation tools)
- [21] Barras, C., Geoffrois, D., Wu, Z., Liberman, W.: Transcriber : Development and use of a tool for assisting speech corpora production. Speech Communication (2001)
- [22] Boersma, P., Weeninck, D.: Praat, a system for doing phonetics by computer. Technical Report 132, University of Amsterdam, Inst. of Phonetic Sc. (1996)
- [23] Chen, L., Liu, Y., Harper, M., Maia, E., McRoy, S.: Evaluating factors impacting the accuracy of forced alignments in a multimodal corpus. In: Proc. of Language Resource and Evaluation Conference, Lisbon, Portugal (2004)
- [24] Sundaram, R., Ganapathiraju, A., Hamaker, J., Picone, J.: ISIP 2000 conversational speech evaluation system. In: Speech Transcription Workshop 2001, College Park, Maryland (2000)
- [25] Quek, F., McNeill, D., Rose, T., Shi, Y.: A coding tool for multimodal analysis of meeting video. In: NIST Meeting Room Workshop. (2003)
- [26] Chen, L., Liu, Y., Harper, M., Shriberg, E.: Multimodal model integration for sentence unit detection. In: Proc. of Int. Conf. on Multimodal Interface (ICMI), College Park PA (2004)
- [27] Rose, T., Quek, F., Shi, Y.: Macvissta: A system for multimodal analysis. In: Proc. of Int. Conf. on Multimodal Interface (ICMI). (2004)
- [28] Sacks, H., Schegloff, E., Jefferson, G.: A simplest systematics for the organisation of turn taking for conversation. *Language* **50** (1974) 696–735
- [29] Duncan, S.: Some signals and rules for taking speaking turns in conversations. *Journal of Personality and Social Psychology* **23** (1972) 283–292
- [30] Padilha, E., Carletta, J.: Nonverbal behaviours improving a simulation of small group discussion. In: Proceedings of the First International Nordic Symposium of Multi-modal Communication. (2003)
- [31] Beattie, G.: The regulation of speaker turns in face-to-face conversation: Some implications for conversation in sound-only communication channels. *Semiotica* **34** (1981) 55–70
- [32] Kalma, A.: Gazing in trials - a powerful signal in floor appointment. *British Journal of Social Psychology* **31** (1992) 21–39
- [33] Argyle, M., Cook, M.: *Gaze and Mutual Gaze*. Cambridge Univ. Press (1976)
- [34] Kendon, A.: Some functions of gaze-direction in social interaction. *Acta Psychologica* **26** (1967) 22–63
- [35] Cassell, A., Nakano, T., Bickmore, T.W., Sidner, C., Rich, C.: Non-verbal cues for discourse structure. In: Proc. of annual meeting of Association of Computational Linguistics (ACL), Toulouse, France (2001) 106–115