A Multi-media Database for Meetings Research

Nick Campbell

ATR Media Information Science Laboratories
Keihanna Science City, Kyoto 619-0288, Japan
nick@atr.jp

Abstract

At ATR, we are collecting and analysing 'meetings' data using a table-top sensor device consisting of a small 360-degree camera surrounded by an array of high-quality directional microphones. This equipment provides a stream of information about the audio and visual events of the meeting which is then processed to form a representation of the verbal and non-verbal interpersonal activity, or discourse flow, during the meeting. This paper describes the resulting corpus of speech and video data which is being collected for the above research. It currently includes data from 12 monthly sessions, comprising 71 video and 33 audio modules. Collection is continuing monthly and is scheduled to include another ten sessions.

1. Introduction

There has recently been considerable interest in the analysis and modelling of meetings from the point of view of multimodal information processing. Curently available corpora include those of ISL (audio-only) [1], ICSI (audio-only) [2], with a dialog act annotation extension [3], NIST (audio-visual) [4], M4 (audio-visual) [5], AMI (audio, video, slides, whiteboard and handwritten notes) [6], and VACE (audio, video, and motion) [7]. A good overview of this work can be found in the online proceedings of the second Joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms [8] which was held in Edinburgh this year. Similar work has recently started in Japan [9, 10].

The VACE team [7] summarise the goal as follows:

> "From a psycholinguistic and social psychology viewpoint, the management of turn taking, floor control, and speaker dominance (even if not speaking) are crucial variables, and the prospect of instrumentally recording clues to these kinds of things could be the basis for valuable interdisciplinary work. These descriptive features are the reality of the meeting to which instrumental recording methods need to make reference. The automatic or semi-automatic monitoring of meetings needs to be related to the actual events taking place in the meeting at the human, social level, and our coding is designed to provide an analytic description of these events. The coding emphasizes the multimodal character of the meeting, attending equally to speech, nonverbal behavior and the use of space, and the aim of the collaboration is to test which (if any) recoverable audio and video features provide clues to such events, thus warranting human inspection. If I may speak for our total VACE collaboration, we believe that some instrumentally detectable traces of turn taking, floor control, dominance and marginality will prove to be feasible."

McNeill describes [11] two key theretical concepts for grounding this work; 'Growth Points' and the 'hyperphrase', highlighting the need for a combination of both speech and video modalities:

> "The cognitive states participants align in multiparty conversations are, in theory, growth points. A growth point (GP) is a mental package that combines both linguistic categorial and imagistic components. Combining such semiotic opposites, the GP is inherently multimodal, and creates a condition of instability the resolution of which propels thought and speech forward. The concept, while theoretical, is empirically grounded. GPs are inferred from the totality of communication events with special focus on speech-gesture synchrony and co-expressivity.
>
> A second theoretical idea, the 'hyperphrase', is crucial for analyzing how these alignments, registrations and maintenances are attained in complex multi-party meetings. A hyperphrase is a nexus of converging, interweaving processes that cannot be totally

> untangled. We approach the hyperphrase through a multi-modal database structure comprising verbal and non-verbal (gaze, gesture) data".

It is our goal in the ATR SCOPE "Robot's Ears" project [10] to produce a similar model of multimodal interaction, and to develop technology that will allow us to identify such key points in a meeting so that we can produce a representation of both the 'flow' of the discourse and of the 'degree and type of participation' of each of the participants at any point in time throughout the meeting.

Our approach is also based on the findings of Kendon [13] and Condon [14, 15, 16] who studied interpersonal synchrony and the functions of interpersonal coordination to signal attention (the latter analysing in great detail the the micromovements of dialogue speakers from a painstaking frame-by-frame analysis of video sequences). Condon claims:

> 'Your body's locked precisely with your speech. You can't break out of this no matter what you do. Your eyes even blink in synchrony with your speech'. Movements appear to begin, change, or end on the same film frame that a new vowel or consonant begins - within about four-hundreths of a second in the new sound. 'The synchrony of the listener with the speaker is just as good as my own synchrony with myself'. An auditory-motor reflex in the central nervous system might allow, even force, a listener's movements to synchronize with a speaker's voice far faster than any conscious reaction time. 'We're almost in auditory touch'.

These findings of a very close interaction between movement and speech are reinforced by Tannenhaus [17], who has been using a lightweight head-mounted eye-tracker to monitor subjects' eye-movements as they follow spoken instructions to manipulate real objects (e.g., "Put the candle that's on the towel into the box"). His work confirms that eye-movements to the objects are precisely time-locked to relevant information in the instruction as it unfolds, thereby making it possible to study the comprehension of spoken language in real-time with natural tasks in real-world contexts.

Accordingly, in the SCOPE project, we are concentrating less on the actual content of the dialogues; i.e., we currently employ no speech recognition or transcription of individual utterances but instead, and at the cost of some fine detail in the physical data, focus on processing combinations of low-level primitives of 'sound' and 'movement'. This approach will allow us to employ simple, non-invasive, monitoring devices to integrate the gestures and 'utterance noises' of the participants, while at the same time as ensuring greater naturalness of interaction between the participants involved.

2. Meetings

The participants at the meetings are all members of the SCOPE project, which incorporates three different research institutions, and the meetings comprise actual monthly progress & planning sessions, which also happen to be recorded. The members are usuallly balanced between the sexes but can be of varying degrees of seniority, experience, and committment to the project. Participation is voluntary and all members are aware that the sessions are being recorded. No scripting or any form of role playing is used.

Meetings may last longer than an hour, but we typically only keep no more than one hour's worth of recordings from each meeting. The core capture is on video, using a 360-degree camera (see figure 1), writing directly to a hard-disk, but the sound recording usually starts earlier and finishes later than the core video recording. Backup video is recorded onto 90-minute tapes using up to six cameras placed around the periphery of the meeting room. No recording devices are worn by the participants, but the device described in section 3 is placed centrally on the table and speakers sit around it in relatively fixed positions (i.e., on chairs without wheels, see figure 2). Numbers of participants vary, but meetings typically include between four and nine people (see figure 3). Table 1 shows participation details and milestones.

The first meeting in November was principally to setup, position, and test the various recording devices. We soon found that artificial lighting was necessary to maintain consistent video quality, as natural sunlight moves and changes in intensity frequently. The third meeting allowed us to test different microphone combinations (see figure 5) and confirmed skin-tone detection to be effective for tracking movements, revealing hands and faces clearly (see figure 4, though note that the effects of summer tank-tops were not then considered!). By the fourth meeting we had developed vectorisation algorithms to separate the individual speakers, and we settled on the 'windmill' assembly described below for a central table-top data collector. We then moved the location to various actual meeting rooms in order to test the effects of different environments on data quality. Most recently, we have been testing the use of individual notebook-based sensors for comparison. We are also testing a miniature assembly to replace the bulky initial equipment for a 'coffee-cup' sized wireless sensor (figure 6).

Table 1: Details of meeting participation and location. (There was no meeting in Dec04). The third column shows number of participants. See text for an explanation of the historical notes.

When	where	who	notes
Nov04	NAIST	6	first setup trial
Jan05	NAIST	8	room lights added
Feb05	NAIST	4	skin-tone-detection
Mar05	NAIST	7	vector segmentation
Apr05	NAIST	9	windmill (all present)
May05	ATR	6	main conference room
Jun05	ATR	8	long meeting table
Jly05	NAIST	9	integrated software
Aug05	ATR	4	tabletop notebooks
Sep05	ATR	8	lunchtime meeting outside

3. A Platform for Data Collection

We are using a Merlin digital video camera with a 360-degree mirror lens (figure 1) and a set of high-end domestic cam-corders (Sony DCR-HC1000) for the video capture. The latter are primarily for labelling purposes, to provide a fuller record of the overall meeting environment than that captured by the 360-degree camera alone.

The small 360-degree video camera is placed in the centre of the meeting table above a ring of directional microphones, in windmill configuration (figures 1, 2, and 5), to collect a stream of audio-visual information from which to characterise the discourse events of the meeting. The video signal is of relatively low resolution (see figure 3), so fine details such as eye-gaze and direction are not available to the system in its present design (and will perhaps not be necessary). Instead, gross movements are detected from the skin tone areas and, from these, a set of primitive features describing the body, hand and head movements is produced automatically.

The output from the co-axial ring array of microphones is synchronised by use of an Edirol FA-101 multi-input analogue-to-digital firewire device to produce a multichannel signal. Audio data is sampled at 48kHz, 24-bit precision. The amplitude (rms power) of each waveform is calculated using a sliding 1024-point Hamming window with a 50-millisecond step-size and the relative amplitude of each line is used to provide an indication of the local variations in overall sound quality around the table. From this information, in conjunction with the video primitives, we are attempting to produce an estimate of each speaker's activity throughout the meeting.

Currently several hardware configurations are being tested, with a close-to-perfect audio signal being captured by a set of four Sennheiser MKH-60 P48 shotgun microphones arranged in a windmill configuration (figure 5) to provide the reference signal for calibration purposes. Four small Audio-Technica directional radio-microphones, placed in parallel to these, are used for system-level 'intermediate sound quality' capture. In addition to these centrally-placed devices, we are also testing small notebook-mounted Sony ECM-Z590 stereo microphones which are closer to the consumer-level recording quality that we envisage using in the final device.

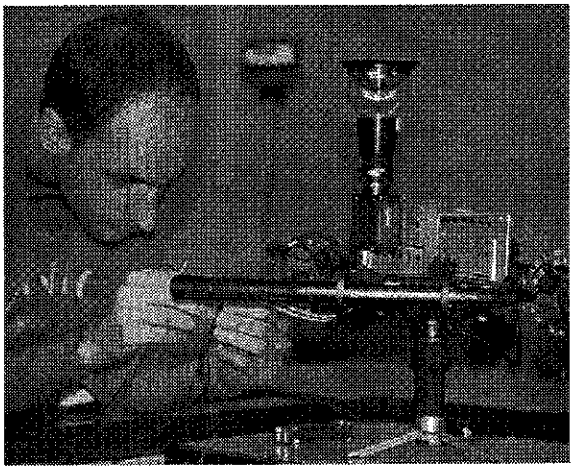

Figure 1: A 360-degree camera surrounded by a ring of microphones provides a view of all participants at the meeting

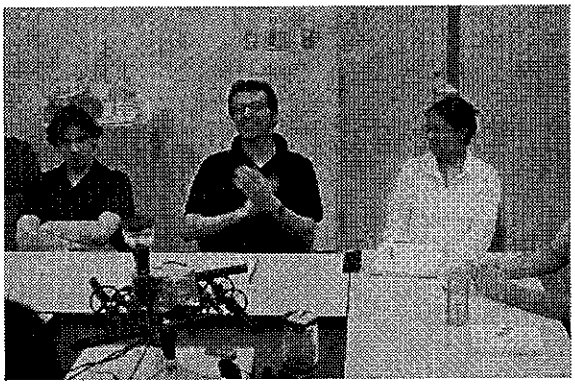

Figure 2: The device is relatively unobtrusive in actual use and does not hamper the interactions

3.1. Processing the Signals

Output from the video capture device is currently at 15 frames per second, for tracking face and hand movements only, and the audio is summarised at 100 frames per sec-

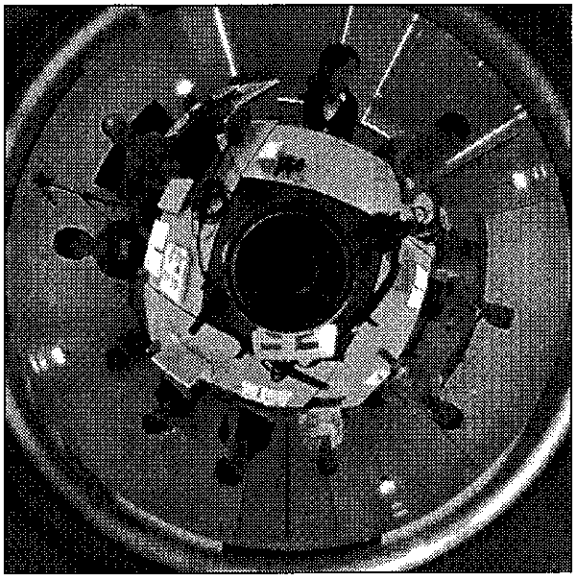

Figure 3: The resulting 360-degree view of the meeting

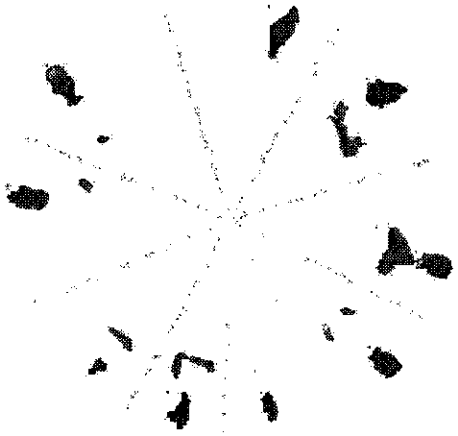

Figure 4: The 360-degree view after skin-tone detection. These are the moving parts that we track

ond. A smoothing algorithm is used to keep track of the video object IDs. The centre of gravity of each object is provided in a low bit-rate output data stream, along with information describing the object motion in rectified coordinates. Positive X motion indicates that the object moves to the left (e.g. the person looks or moves a hand to their left), positive Y motion indicates that the object moves up (e.g. the person moves his or her head up), and vice versa. This combined information is then vector-quantised and sent to a discrete HMM which has been trained using the manually-produced movement and discourse-event labels described in [10].

3.2. Alternative Collection Devices

While the core recording equipment remains unchanged, and the windmill microphones provide a consistent, high quality representation of the spoken interactions, we are also testing different combinations of both video and audio recording devices, and varying them from month to month.

For example, since it is now normal practice for participants to bring notebook computers to a meeting, we are also equipping each with a web-cam and a stereo microphone so that data can be collected per participant on a local rather than a centric-global basis. This results in consequent problems of data synchronisation but offers a non-invasive method that could be more closely linked to each participant. However, currently we find that people move around too much and that the closer camera suffers from the excessive ranges.

We are also testing different microphones, microphone combinations, and placement strategies. These are collected simultaneously with the core devices, and can be compared against their critical standard. Since our goal is to develop a device that is small, cheap, and unobtrusive, centrally-placed devices are perhaps to be preferred. We will not have the extremely high quality input that the core devices produce in a practical real-time system, but in this early development stage it is useful to be able to use the clean data as a baseline for comprison of the performance of the less-critical systems.

4. Discussion

We have limited our research context to that of a small business meeting, in which we track the flow of discourse and participant relationships in order to (a) produce a listing of those parts of the meeting for which a more detailed transcription might be necessary, and (b) produce a flow analysis independent of any linguistic information.

From our previous analyses of natural daily-conversational interactions [18, 19], we determined that a considerable portion of the spoken interaction that takes place between humans is primarily non-verbal, serving to express affect and to show current states of interpersonal relations [20, 21]. It is likely that even in a more formal meetings environment, such interactions will be found to be common.

In application of this knowledge to current technology, both robots and embodied conversational agents can

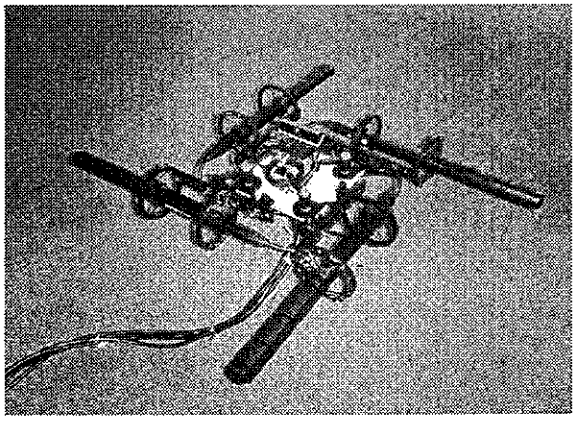

Figure 5: The microphone windmill— by placing direction-sensitive microphones at the cardinal positions, we can also collect high-definition audio streams to assist in tracking the location of each sound source for the training phase

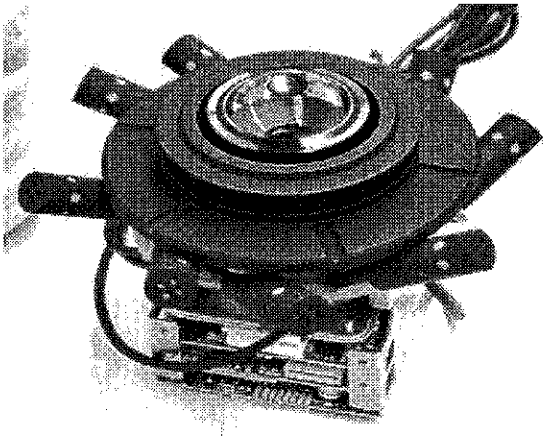

Figure 6: The new 'coffee-cup-sized' sensor - a SONY RPU-C251 Chamelion Eye with a ring of Audio Technica radio microphones

make immediate use of the resulting models and algorithms. To be believable, a life-like agent must appear to understand what is said to it, or what is being said around it, even if this is not actually the case. It must be able to follow a conversation and to understand what is happening in a discourse, even though the verbal content of the dialogue may be too complex (or too noisy) to be recognised. In this work, we are testing the platform design and details of the description language for such non-verbal speech-processing.

5. Conclusion

This paper has described a multi-media speech-and-video data collection that is being carried out for research into the dynamics of discourse processes and interpersonal interactions in a meetings context. The research is sponsored primarily by the Japanese Ministry of Internal Affairs and Communications, and is part of a 3-year project to develop technology for the processing of non-verbal information in human interactions. It is being carried out at ATR in Japan, in collaboration with graduate schools of Kobe University and the Nara Institute of Science and Technology.

The principal physical apparatus consists of a central 360-degree video camera and an array of up to eight coaxially-mounted directional microphones. This is backed-up by room video cameras and far-field microphones to aid in subsequent human analysis of the main data streams. Computer processing of the data streams is carried out by statistically-based methods (currently HMMs) that are trained on the manually-labelled audio and video data.

The content of the multi-media data is a series of monthly project meetings where members from different institutions, and with diffferent roles in the project, meet to appraise progress and to plan technical steps as the project progresses. Thus, participants are not 'playing a role' but are actually personally involved, to different degrees, in the progress of the meeting. The number of participants, who are seated in relatively fixed positions around the table, varies between four and twelve. No invasive collection techniques (such as lapel-mounted microphpones) are used, but nonetheless a high quality set of video and audio signals is being collected.

The immediate goal of the project is to produce a model of (and a technology for the processing of) the discourse flow within a meeting, such that the main speaker can be identified, and listener attention, agreement and dissent, etc., can be detected from the audio-visual information stream by the processing of non-verbal primitives. It remains as future work to evaluate the performance of this model and to determine exactly how far up the discourse hierarchy such low-level interaction information can effectively be made use of.

6. Acknowledgements

This work is partly supported by the Japan Science & Technology Corporation (JST), partly by the National Institute of Information and Communications Technology (NiCT), and partly by the Ministry of Public Management, Home Affairs, Posts and Telecommunications, Japan. The author is grateful to the management of ATR for their continuing encouragement and support.

7. References

[1] S. Burger, V. MacLaren, and H. Yu, "The ISL meeting corpus: The impact of meeting type on speech style", in Proc. International Conference on Spoken Language Processing (ICSLP), Denver, Sept. 2002.

[2] 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.

[3] E. Shriberg, R. Dhillon, S. Bhagat, J. Ang, and H. Carvey, "The ICSI meeting recorder dialog act (MRDA) corpus", in Proc. HLT-NAACL SIGDIAL Workshop, Boston, Apr. 2004.

[4] V. Stanford, J. Garofolo, , and M. Michel, "The NIST smart space and meeting room projects: Signals, acquisition, annotation, and metrics", in Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), Hong Kong, 2003.

[5] 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. 305317, Mar. 2005.

[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 preannouncement", in Proc. Workshop on Machine Learning for Multimodal Interaction (MLMI), Edinburgh, Jul. 2005.

[7] L. Chen, R. Travis 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.

[8] 2nd Joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms, online proceedings: http://groups.inf.ed.ac.uk/mlmi05/techprog.html.

[9] M. Katoh, K. Yamamoto, J. Ogata, T. Yoshimura, F. Asano, H. Asoh, N. Kitawaki, "State Estimation of Meetings by Information Fusion using Bayesian Network", Proc Eurospeech, pp. 113–116, Lisbon, 2005.

[10] W. N. Campbell, "Non-Verbal Speech Processing for a Communicative Agent", Proc Eurospeech, pp. 769–772, Lisbon, 2005.

[11] D. McNeill, "Gesture, Gaze, and Ground", in Proc 2nd Joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms, Royal College of Physicians, Edinburgh, UK. July 2005.

[12] SCOPE homepage: http://feast.his.atr.jp/non-verbal

[13] A. Kendon, "Movement coordination in social interaction: Some examples described". Acta Psychologica, Amsterdam, 32(2): 101 125. 1970.

[14] W. S. Condon, "Synchrony Demonstrated between Movements of the Neonate and Adult Speech", Child Development 45: 456-462. 1974.

[15] W. S. Condon, "Multiple response to sound in dysfunctional children", Journal of Autism and Childhood Schizophrenia 5: 37-56. 1975.

[16] W. S. Condon, "Communication: Rhythm and Structure. Rhythm in Psychological, Linguistic and Musical Processes", J. R. Evans and M. Clynes. Springfield, Illinois, Charles C Thomas Publisher: 55-78. 1986.

[17] M. K. Tanenhaus, M. J. Spivey-Knowlton, K. M. Eberhard & J. E. Sedivy, "Integration of visual and linguistic information in spoken language comprehension". Science, 268, 1632-1634. 1995.

[18] The JST/CREST Expressive Speech Processing project, introductory web pages at: http://feast.his.atr.jp

[19] W. N. Campbell, "Listening between the lines; a study of paralinguistic information carried by tone-of-voice", pp 13-16 in Proc International Symposium on Tonal Aspects of Languages, TAL2004, Beijing, China, 2004.

[20] W. N. Campbell, "Expressive Speech - Simultaneous indication of information and affect", pp.49-58 in *From Traditional Phonology to Modern Speech Processing"(Festschrift for Professor Wu Zongji's 95th birthday*), eds G.Fant, H.Fujisaki, J.Cao & Y.Xu, 2004.

[21] W. N. Campbell, "Perception of Affect in Speech — towards an Automatic Processing of Paralinguistic information in Spoken Conversation", 8th International Conference on Spoken Language Processing, pp881-884, Jeju, Korea, 2004.