A multimedia database of meetings and informal interactions for tracking participant involvement and discourse flow

Nick Campbell, Toshiyuki Sadanobu, Damien Douxchamps, Masataka Imura, Suzuki Noriko and Mayumi Bono

ATR Media Information Science Laboratories, Nara Institute for Science and Technology, Kobe University, 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. Currently 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].

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.

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.

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 [11, 12], 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 [13, 14]. 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 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.
2. 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 different 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 microphones) 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.

3. 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. Dhillion, 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 pre-announcement”, 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] The JST/CREST Expressive Speech Processing project, introductory web pages at: http://feast.his.atr.jp
[12] 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.
[13] 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.
[14] 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.