How far can non-verbal information help us follow a conversation? Preliminary experiments with speech-style and gesture tracking.

Nick Campbell
ATR Network Informatics Laboratories,
Keihanna Science City, Kyoto, 619-0288, Japan.
nick@atr.jp

November 26, 2004

Abstract

This paper introduces a new project funded by the Japanese Ministry of Public Management, Home Affairs, Posts & Telecommunications, under the Strategic Information and Communications R&D Promotions Programme (SCOPE). With the working title "Listening Machines; following the flow of a human conversation", it aims to produce non-verbal information processing technology for use in robots, meetings, and as an intelligent human-machine interface component.
Keywords: non-verbal communication, speech & gesture, robots' ears & eyes, conversation-flow, gestural primitives, emergent understanding

1 Introduction

One of the interesting scientific findings arising from the recent JST/CREST Expressive Speech Processing Project [1, 2] is that people communicate as much by use of tone-of-voice and changes in their speaking style as they do by use of linguistic information when they speak. This may be considered common sense, and certainly not new in itself, but it is an area of oral communication that is not at all addressed by current speech technology.

The work of the ESP project has resulted in computer algorithms for the automatic quantification of differences in settings of the vocal tract that correlated well with changes of interlocutor and speaker-state. This work also resulted in a model of non-verbal communication [?] in which discourse events are realised subject to the constraints of a framework of speaker-state and speaker-listener relationships. It was further found that less than half the number of utterances in a very large conversational-speech corpus were used to convey linguistic content, and that many were primarily expressing discourse-flow, speaker-state and relationship information.

The present work, on "Listening Machines" extends the JST/ESP research, which was focussed on speech production, towards speech understanding. In particular, it aims to produce a technology that is capable of processing this non-verbal information in order to follow the flow of a conversation. From simple oral and gestural primitives, a representation will be formed of the roles of the participants and the effect of their utterances on the social dynamics of the interactive situation. The system makes no use of linguistic knowledge, but instead uses information about gestures related to speech events, making inferences from low-level primitives derived from the acoustic and video signals. From the representation, it should be possible to determine which utterances or parts of an utterance need a linguistic analysis and which can be processsed per-se.

2 Data Collection

The initial constraint on the conversation is that it be in round-table format, with participants (currently 2 - 6) seated within clear view of two centrally-mounted 360-degree cameras (see figure 1). We use one camera at table-level and one higher-up for a birds-eye view of speakers' movements. Audio is currently captured by use of studio-quality chest-worn radio microphones, but our intention is to use an array of centre-mounted microphones for the eventual application, in order not to encumber the speakers in any way.

3 Data Analysis

We are currently processing data from one such conversation, which lasted for an hour and took place in a quiet, white-walled room. The video and audio streams are collected, synchronised, and labelled manually using Transcriber for the initial segmentation of the speech, and Wavesurfer for subsequent labelling of the video-related information. The poster presentation will give details of the transcription conventions and show samples of the types of non-verbal indformation that can indicate conversational flow and participant involvement. It remains as future work to map between the perceptual information and the derived primitives.

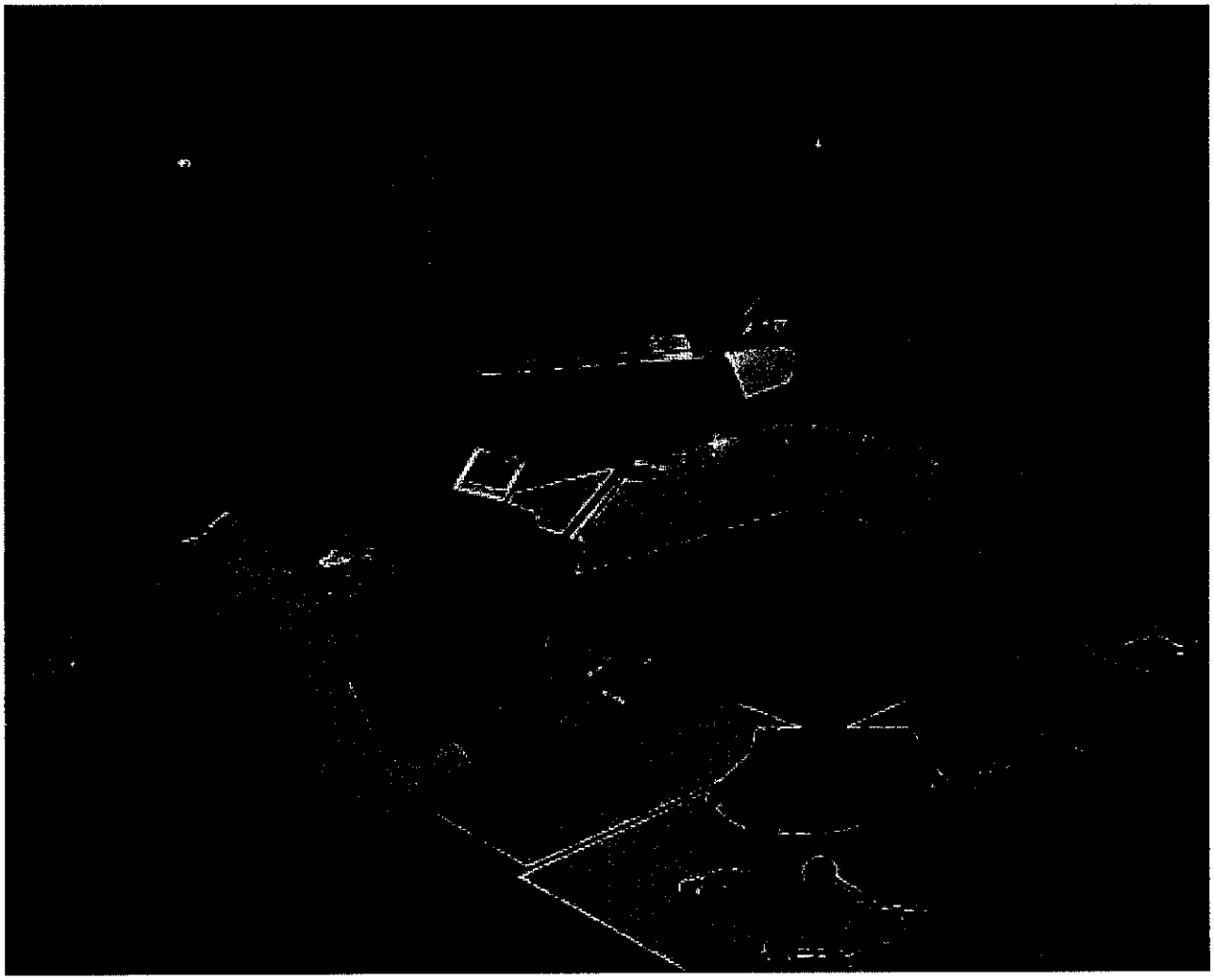

Figure 1: The "round-table" setup, showing 2 vertically-mounted omni-directional cameras and audio equipment

Acknowledgements

This work, which is supported by a grant from the Japanese Ministry of Public Management, Home Affairs, Posts & Telecommunications (Soumu-sho). extends previous research funded by both the National Institute of Information & Communications Technology, and the Japan Science & Technology Agency under CREST Project #131.

References

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

[2] Campbell, W. N., "Recording Techniques for capturing natural everyday speech", in Proc Language Resources and Evaluation Conference (LREC-2002), Las Palmas, Spain, 2002.

[3] Campbell, W.N., "Databases of Emotional Speech", in Proc ISCA (International Speech Communication and Association) ITRW on Speech and Emotion, pp. 34-38, 2000.

[4] Campbell, W. N., "Voice Quality; the 4th prosodic parameter", in Proc 15th ICPhS, Barcelona, Spain, 2003.

[5] Mokhtari, P, & Campbell, W. N., "Automatic detection of acoustic centres of reliability for tagging paralinguistic information in expressive speech.", in Proc LREC 2002.

[6] ESP web-pages: http://feast.his.atr.jp/esp