Automatic detection of participant status and topic changes in natural spoken dialogues *

Nick Campbell, NiCT/ATR

1 Introduction

The goal of this work is to automatically tag an interactive spoken dialogue for participation states for each participant and to detect topic changes and areas of high engagement or empathy in the discourse ('moriagari' in Japanese 対話の盛り上がり). The detection of participant status in a spoken discourse is a relatively new area of interest for speech research and has particular application in the construction of interactive dialogue systems. Recent work in this area includes AMI [1] and CHIL [2] as well as SCOPE [3] (the Robot's Ears Project [4]) and this paper extends the above SCOPE findings through Kaken-funded [5] work with multicultural subjects in a round-table environment, interacting in English, for multimodal conversational data collection and analysis.

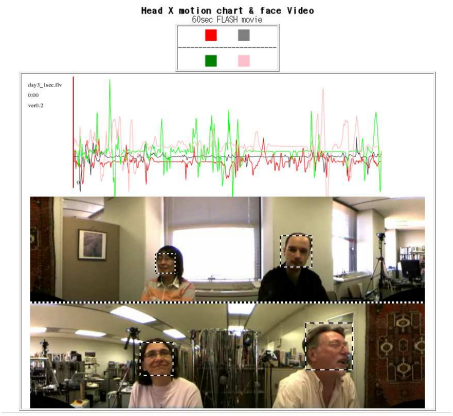

Fig. 1 The data capture environment and display (This figure can be viewed interactively, with sound, at http:feast/nonverbal/moriagari/)

The materials for this study consist of multimodal recordings of three 90-minute conversations in which the participants were from four countries, each speaking a different native language but all conversing in English, as guests in a research laboratory in Western Japan (see Figure 1). The conversations were free and unstructured, and the interactions were recorded in audio and multichannel video, primarily using a small unobtrusive centrally-placed 360-degree capture device which tracked the movements of their heads and upper-bodies throughout. The data and annotations are all available on the web, but access requires a password which can be obtained from the author on request, subject to conditions of confidentiality.

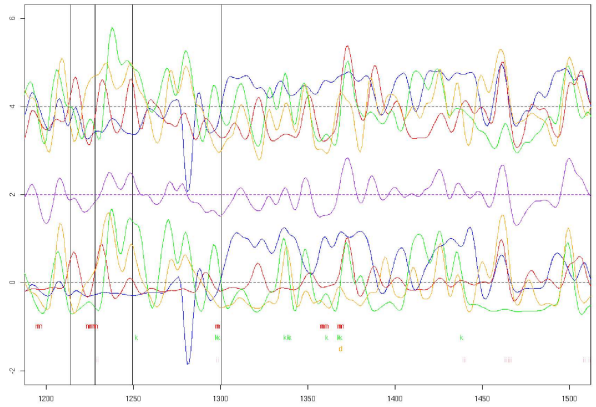

Fig. 2 Display of 5-minutes of captured data. See text for a detailed explanation of the plots.

This paper describes ongoing work into the detection of topic boundaries in the conversational speech using multimodal sensors to detect and evaluate synchrony in the physical movements of the participants in order to produce an estimate of their participation status and degree of engagement in the conversation.

The situation of the conversations was such that all participants were seated approximately equidistant from a small and unobtrusive centrally-placed 360-degree video capture device that employed face-spotting technology to first determine the number of active participants and then to record the movement or activity measured in the region of each detected face and in the equivalent regions 2.5 times the width of each head and immediately below it. The data was produced in real time and consisted of six streams of real x, y, and z values describing the head and body movement of each participant with a 10 fps frame rate, which is now considered optimal for fast processing of speech-based information.

Simultaneously with this video data capture, a centrally-placed large-diaphragm high-quality dynamic microphone collected sound levels to estimate speech activity from each participant. The purpose of the sound-stream information is not to feed a speech recogniser, since no 'intelligent' processing is done by the proposed apparatus, but just to note the on-off nature of speech activity for each participant, for use in conjunction with the video data to estimate synchrony in participation.

Participation status can be considered as an ordered series of discrete states, ranging from 'inactivity' through 'listening' to 'talking', and including such intermediate states as 'waiting to speak', 'thinking', 'agreeing', 'actively listening', etc.

---

*対話の盛り上がりとトピック境界の自動抽出について

ニック キャンベル NiCT/ATR-SLC, Keihanna Science City, Kyoto, 619-0288, Japan,

Correlation	**P1**	**P2**	**P3**	**P4**
head/body	0.797	0.809	0.808	0.722

Table 1 Head - body correlations within speaker

Body	b1	b2	b3	b4
P1	-	0.289	0.082	0.436
P2	-	-	-0.308	-0.036
P3	-	-	-	0.408
Head	h1	h2	h3	h4
P1	-	0.53	0.233	0.239
P2	-	-	0.081	-0.204
P3	-	-	-	0.221

Table 2 Head and Body correlations between 4 partners. Note especially P1-b4 and P1-h2.

In order to produce estimates of such participation features, the audio and video data is processed to determine synchrony features that indicate different types of involvement of the participants. This paper will focus primarily on processing of the video component.

Although the apparatus produces estimates of head and body movement separately, one can expect a high correlation between the head and body movements of each individual member. Table 1 shows these correlations between head and body movements for the four speakers in conversation 3 of the series to be in the order of $r = 0.8$.

Table 2 shows correlations of head and body movements between participants. Here, much lower correlations are found, as the participants take turns leading the dialogue and showing different degrees of interest and agreement in each topic and section.

It was originally assumed that the high correlation between head and body movement for each participant reduced the relevance of this information difference, but as Table 2 shows, different pairs of participants show higher head or body coordination, so both measures can be considered useful.

2 Experiments and results

A measure of normalised activity was calculated for each participant, by taking the log of each movement activity, scaling it by z-score to normalise the ranges, and smoothing the result with a Butterworth filter. Plots of the resulting data are shown in Figure 2, with head data at the top, body at bottom, and a global measure in the centre showing summed activity of the group. Vertical lines in the plot show topic boundaries as assigned by manual annotation.

It could be expected that each topic starts with low participant activity and ends with higher agreement (and correspondingly more synchrony in the movements) between the participants. However, Table 4 shows this not to be the case. By calculating the slope of a linear regression through the activity plots across each topic section, we find that as many are negative-valued as are positive, with the median being flat.

0%	25%	50%	75%	100%
4.49	22.04	46.01	78.40	276.45

Table 3 Quantiles durations (measured in seconds) of the 79 topic segments in a ninety-minute conversation.

	min	25%	50%	75%	max
bm1	-2.74	-0.10	-0.00	0.15	6.65
bm2	-1.10	-0.09	-0.01	0.09	4.15
bm3	-5.62	-0.24	-0.00	0.20	2.52
bm4	-6.95	-0.33	0.00	0.32	5.93
hm1	-6.63	-0.36	0.02	0.24	10.24
hm2	-3.71	-0.26	0.00	0.17	9.61
hm3	-7.04	-0.45	-0.03	0.26	2.49
hm4	-5.82	-0.29	-0.02	0.30	3.49

Table 4 Quantiles of the slope of a linear regression fit measured across all topic segments in a ninety-minute conversation.

Whereas slope across the topic segment appears to be flat on average, Figure 2 clearly shows that there are points in the dialogue where all participants come together in peaks of group activity. These peaks of empathy or 'moriagari' frequently occur immediately prior to a topic boundary, or lull in the conversation.

Present work involves detecting these peaks by using strong local correlations such as can be found e.g., at 1240, 1375, and 1460 seconds in Figure 2, and relating these local group maxima to topic boundaries. After smoothing, no time-lag calculation appears necessary. A count of these coincidences shows that more than half of the topic changes can be detected by means of these local maxima of joint activity.

3 Conclusion & Future Work

This paper has reported a method for using movement data to estimate discourse participation. This is still work in progress. A full quantitative analysis of these latest results will be presented in the oral version of this paper and published on the project web site at http://feast.atr.jp/nonverbal/moriagari/.

References

[1] AMI: Augmented Multi-party Interaction http://www.amiproject.org/

[2] CHIL - Computers In the Human Interaction Loop. http://chil.server.de/servlet/is/101/

[3] The SCOPE-funded ATR "Robots' Ears Project": http://feast.atr.jp/non-verbal/scope/lrec-campbell-et-al.pdf

[4] "Robust real time face tracking for the analysis of human behavior", Nick Campbell, Damien Douxchamps, in Machine Learning & Multimodal Interaction, Springer's LNCS series, 4892, pp.1-15, Dec. 2007.

[5] "On the use of Nonverbal Sounds in Human Communication" Grant in Aid #19300073, Ministry of Education, Culture, Sports, Science and Technology (MEXT), Japan.