Working with Very Sparse Data to Detect Speaker and Listener Participation in a Meetings Corpus

Nick Campbell & Noriko Suzuki

Department of Cognitive Media Informatics,
ATR Media Information Science Labs, Keihanna Science City, Kyoto, 619-0288, Japan,
{nick,noriko}@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. In this paper we show that simple primitives can provide a rich source of information.

1. Introduction

Several laboratories around the world are now collecting and analysing "meetings data" in an effort to automate some of the transcription, search, and information-retrieval processes that are currently very time-consuming, and to produce a technology capable of tracking a meeting in real-time and recording and annotating its main events. One key area of this research is devoted to identifying and tracking the active participants in a meeting in order to maximise efficiency in data collection by processing inactive or non-participating members differently. [1, 2, 3, 4, 5, 6, 7, 8].

At ATR we are now completing the second year of a three-year SCOPE funded project to collect and analyse such data. This paper reports an analysis of material collected from one such meeting in terms of speaker overlaps and conflicting speech turns. Our goal is to determine whether it is necessary to track multiple participants, or whether processing can be constrained by identifying the dominant member(s) alone. The results show that in a clear majority of the cases, only one speaker is active at any time, and that the number of overlapping turns, when two or more psrticipants are actively engaged in speaking at the same time, amount to less than 15% of the meeting. This encourages us to pursue future research by focussing our resources on identifying the single main speaker at any given time, rather than attempting to monitor all of the speech activity throughout the meeting.

The second part of the paper shows that a change in speaker might be predicted from the amount and types of body movement. These movements are speaker-specific and not uniform, but systematically increase in the time immediately before onset of speech. By observing the bodily movements of the participants, we can form an estimate of who is going to speak next, and prepare to focus our attention (i.e., the recording devices) accordingly.

2. Categories of Speech Activity

We have regularly been recording our monthly project meetings, where research results and project planning are discussed, to provide a database of natural (non-acted/no role-playing) speech and interaction information.

The number of members attending each monthly project meeting can vary between four and twelve. Participation

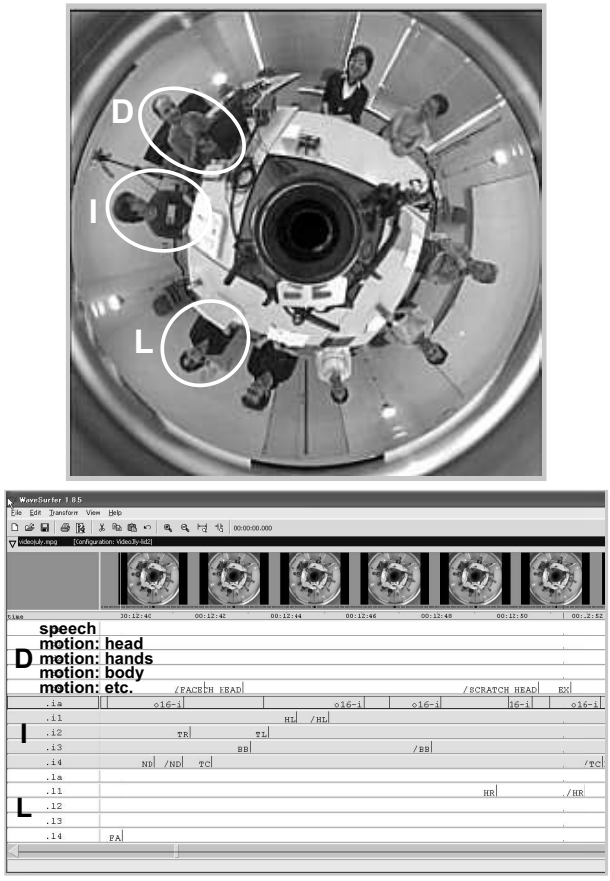

Figure 1: The camera's-eye view of a meeting (top), showing the annotated movement data for three participants (D,I,L) using the wavesurfer video plugin (bottom)

is voluntary, but since the research is being carried out by three teams at different locations (ATR, NAIST, and Kobe University) the meetings provide an essential focus-point for coordinating the research activities.

All meetings are recorded on both video and audio, using purpose-built equipment that has been described elsewhere [9, 10, 11]. All visible body-movements of the participants (head, hands, and torso) are annotated from observation of the video recordings, topic changes are noted, and the categories of speech activity are tagged by human labellers working interactively with the data.

Table 1: Topics that arose during the July meeting, with durations, showing the division between researcher-centred and technology-centred discussions

id	topic	seconds
t-o2	progress-update(s1)	45
t-o9	progress-update(s2)	205
t-o15	progress-update(s3)	64
t-o23	progress-update(s8)	76
t-o12	self-introduction(s5)	191
sub-total		(738)
t-o6	data-tagging results	15
t-o8	data-preparation	157
t-o10	tanktops-and-skin-tones	142
t-o14	equipment-settings	82
t-o16	NAIST responsibilities	119
t-o18	reporting procedures	160
t-o20	Kobe Uni. responsibilities	58
t-o24	kinematics	148
t-o29	chameleon-eye-lens	564
t-o22	translation	11
t-o27	choice-of-camera	7
sub-total		(1306)
total		2044

The speech is not yet being transcribed verbatim, but tags are assigned per topic and per activity type. We consider it necessary to distinguish (i) "on-topic" speech from (ii) "personal" speech, and also (iii) "backchannel utterances" and (iv) "laughter". We had also proposed (v) "yes" and (vi) "no" as relevant categories, but our experience with annotating these further two types of speech event suggests that they will not be easily recognisable using automatic processing, and we currently limit our tagging of speech activity to types i-iv above.

3. Overlapping Speech

This paper reports the results of an analysis of one such meeting. Eight members were present at the meeting, which was held at NAIST in July 2005. They included the research director (s1), two team leaders (s3,s8), two researchers (s2,s4) two administrative assistants (s6,s7) and a guest researcher visiting from Ireland (s5). An observer was also present to monitor the recordings. The statistics of speech activity reported below clearly reflect the different roles of the participants, and the importance (in terms of time devoted to each) of the various topics.

Topics of discussion (see Table 1) included (a) progress-updates (approx. 36%) where one speaker tended to dominate, with the others listening and asking occasional questions, and (b) technical topics (approx. 64%), where more members became involved in the discussions.

There were 2513 different "speech events" in the meeting, which lasted approximately 45 minutes altogether. Here, a speech-event is defined as a block of continuous speech, bounded by a cessation of speech activity, from one speaker, as indicated by '+' = start and '-' = end markers in the columns of figure 2. A brief silence after a burst of speech is marked by the '-' label.

```
0.72    .|.|.|.|.|.|.|-p      938.49
0.78    .|+o.|.|.|.|.|.|      939.27
0.23    .|=o.|.|.|.|.|.|+p    939.51
0.47    .|-o.|.|.|.|.|.|=p    939.99
1.4     .|.|.|+p.|.|.|.|-p    941.39
0.75    .|.|.|-p.|.|.|.|.|    942.15
0.04    .|.|.|+p.|.|.|.|.|    942.2
0.82    .|.|.|-p.|.|.|.|.|    943.02
0.55    .|.|.|+p.|.|.|.|.|    943.58
0.58    .|.|.|-p.|.|.|.|.|    944.16
1.04    .|.|.|.|.|.|.|.|+b    945.21
0.15    .|.|.|.|.|.|.|.|-b    945.37
0.18    +o.|.|.|.|.|.|.|.|    945.55
0.92    =o.|.|.|.|.|.|.|+p    946.48
1.32    -o.|.|.|.|.|.|.|=p    947.8
0.01    +o.|.|.|.|.|.|.|-p    947.82
0.39    -o.|.|.|.|.|.|.|+p    948.21
0.21    .|.|.|.|.|.|+b.|=p    948.42
1.25    .|.|.|.|.|.|-b.|-p    949.68
0.03    +o.|.|.|.|.|.|.|+p    949.72
2.3     -o.|.|.|.|.|.|.|=p    952.03
0.95    .|.|.|.|.|.|+o.|-p    952.98
0.29    .|.|.|+p.|.|-o.|.|    953.27
0.89    .|.|.|-p.|.|.|.|.|    954.17
0.65    +o.|.|.|.|.|.|.|.|    954.82
0.41    =o.|.|+p.|.|.|.|.|    955.24
1       -o.|.|=p.|.|.|.|+p    956.24
1.2     +o.|.|-p.|.|.|.|=p    957.45
0.03    =o.|.|.|.|.|.|.|-p    957.49
1.75    -o.|.|.|.|.|.|.|.|    959.24
0.07    +o.|.|.|.|.|.|.|.|    959.32
0.08    -o.|.|.|.|.|.|.|.|    959.4
0.41    +o.|.|.|.|.|.|.|.|    959.82
```

Figure 2: A sample of the audio labelling, showing three categories of speech sctivity: o=opinion, or public speech, p=private or personal speech, and b=backchannel utterances. A '+' indicates onset of speaking, '=' continuation, and '-' cessation of speech. The time in seconds of each event is shown on the left, and absolute time on the right

Table 2: Counts of speech events per participant

s1	s2	s3	s4	s5	s6	s7	s8
759	587	106	127	522	64	75	138

The distribution of events per speaker is shown in Table 2. Tables 3 and 4 detail the types of speech activity and times spent on each per speaker. Mean event duration is 0.7 seconds (sd=0.78), with the longest recorded event being 17 seconds. The 25th quantile of event durations is at a quarter of a second, and the 75th at 1 second. There were in addition 1730 points throughout the meeting during which no-one spoke.

Both total utterance counts and overall speaking times indicate that s1 (the project leader), and s2 (a guest researcher expert in graphics processing) dominated the meeting. It is also evident from tables 2 & 3 that s5, the observer, also took an active part in the discussion. The administrative assistants spoke least at this research-based meeting.

Table 3: Utterance timings for each participant for three categories of activity: O; on-topic talk, P: private talk, B: backchannel utterances. All timings are rounded to whole seconds.

	s1	s2	s3	s4	s5	s6	s7	s8
o	344	32	49	47	212	27	22	44
p	5	7	-	4	14	5	30	2
b	64	11	2	10	17	3	2	1

Table 4: Number of events for each speaking type

on-topic	backchannel	private	laugh
2110	207	196	406

The count of participants actively speaking during each turn is given in Table 5. It shows that by far the majority of turns are single-speaker events. It is 6.5 times more likely that any given utterance will be single-speaker, and only 15% likely that more than one speaker will be active. There is only a 7% chance of more than 2 people speaking at any time in this meeting of 8 researchers. These figures may of course be culture-specific, and even meeting-specific.
It might be supposed that backchannels contribute to the majority of overlapping utterances, but a count of single-speaker backchannel utterances (n=134) versus a count of multi-speaker, overlapping backchannel utterances (n=74) shows this not to be the case. If we exclude from this s1's overlapping backchannels to s2 (n=19) then the ratio becomes 55:134, and it is 2.5 times more likely that a bakchannel utterance will be spoken without overlap.

Table 5: Number of participants active at each turn

silent	solo	two	three	four
1730	2000	291	15	1

4. Speech & Movement

It has often been observed (e.g., [12, 13, 14]) that people move more when they speak. To determine whether these two types of activity had any useful correlation, we also examined the physical activity of all participants that was visible to the camera. We looked both at activity prior to speaking, and at activity while speaking. Since all were seated around a table, this study is limited to upper-body movement.
Figure 1 shows the multi-tiered annotation that we use for labelling body movements which are apparent to a human observer when viewing the 360-degree camera output. In addition to a speech-related tier, separate tiers are available for "head", "hands", "body", and "other", where the last can be used for complex gestures such as "play with pencil", "scratch head", "fix glasses", "stroke beard", etc.
For this paper, we simply counted the number of active labels at each moment of time and categorised them as follows: "Motion 1": only one body part is moving (e.g., the head or a hand), "Motion 2": two body parts moving

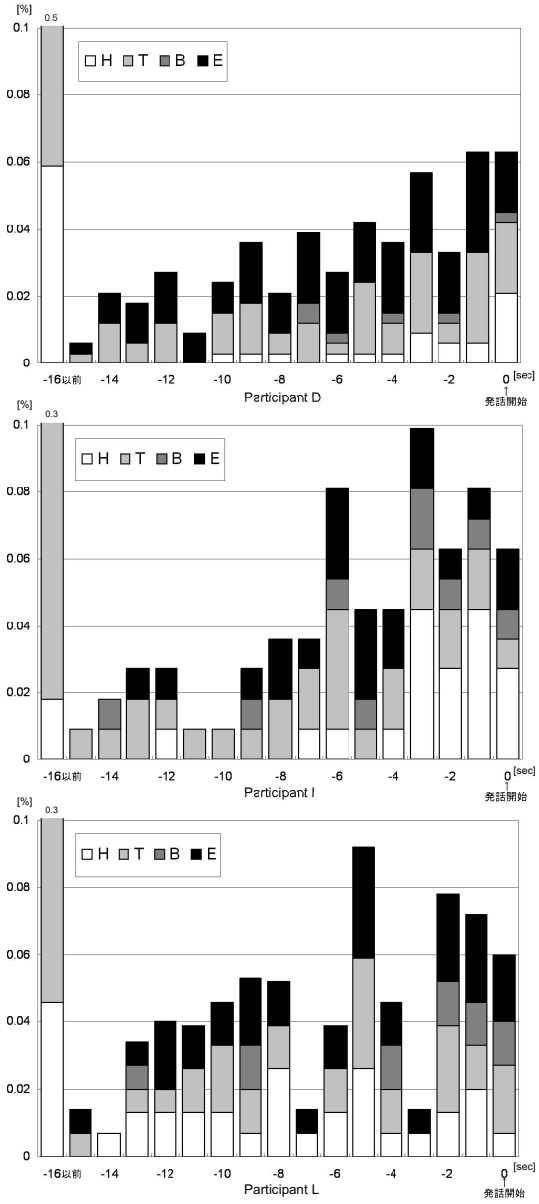

Figure 3: Rising amounts of bodily movement for 3 participants across a period of 16 seconds prior to onset of speech. Two speakers show a peak of activity a few seconds before speaking. Here "H" represents head movement, "T" represents hand movement, "B" represents body movement, and "E" represents particular gestures (see text for details). The rightmost column shows onset of speech, and the leftmost sums all movements since last speech event.

(e.g., head and hand, or two hands), "Motion 3": three body parts moving (e.g., head and hand and body), and "Motion 4": four or more body parts moving. The data from three speakers (those circled in the figure) were then compared for the periods immediately prior to onset of speech. Figure 3 clearly shows a rise in the amount of activity as the person prepares to speak. However, we can see individual differences, and it appears that two speakers reach a peak of activity shortly before speaking, while the third continues to increase up to the onset of speech.
We can also note differences in parts of the body moved: Participant I, for example (the centre portion of figure 3),

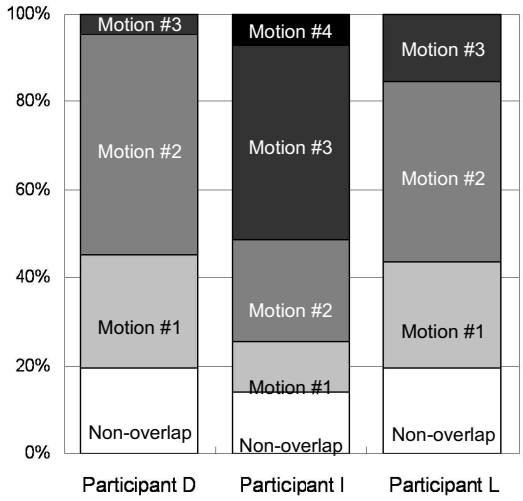

Figure 4: Number of major body parts that move while speaking. Non-overlap indicates that the speaker spoke while remaining relatively still. Participant I (centre) differs in moving more than the other two.

appears to move his head much more than the others (as indicated by the white portion of the bars). Figure 4 provides a breakdown of the types of activity per participant. It shows that for all speakers, the occurrance of speech having no overlap with body movement accounts for less than 20% of the total speaking time. It also shows that speakers behave differently; with all speakers moving 2 or more body parts at least 50% of the time, but one speaker (the centre column) moving 3 or more body parts more than 50% of the time while speaking.

5. Discussion

The above analysis of the audio data shows that in a clear majority of the cases, only one speaker is active in any given turn. This implies that we will only lose a small amount of relevant information if we limit our processing to the single most dominant member at any one time. This will considerably reduce the work-load of the processing.
Furthermore, from an examination of the video data, we confirmed that people do tend to move more when they speak, and found that there is a steady rise in the amount of movement of all participants particularly in the 10 to 15 seconds preceding the onset of speech.
From the two above findings, we conclude that it is feasible to design technology, based on the very simple presence or absence of speech noise and movement in the video signal, that will be able to detect and track the speakers in such a meeting situation. However, it will require development of separate technology to be able to determine the reactions of the other participants to any particular utterance or topic. This remains as future work.

6. Acknowledgements

This work is supported as part of the Strategic Information and Communications R&D Promotion Programme (SCOPE) by the Ministry of Internal Affairs and Communications, Japan and is being carried out as collaborative research between members of ATR, Kobe University, and the Nara Institute of Science & Technology.

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] 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.

[3] W. N. Campbell, "A multimedia database of meetings and informal interactions for tracking participant involvement and discourse flow", in Proc LREC 2006, Lisbon.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

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

[10] W. N. Campbell, "A Multi-media Database for Meetings Research", pp 77-82 in Proc Oriental COCOSDA, 2006, Jakarta, Indonesia.

[11] Project Homepage: http://feast.atr/non-verbal

[12] Zhang, D.,et al., "Multimodal group action clustering in meetings"., VSSN'04, 54-62, 2004.

[13] Katoh, M., et al., "State estimation of meetings by information fusion using bayesian network", INTERSPEECH2005, 113-116, 2005.

[14] 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.