Speech Overlaps in the ATR Meetings Data; an Initial Analysis of Multi-speaker Turns *

◎ Nick Campbell, ATR

1 Introduction

Several laboratories in Japan and 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 [1, 2, 4, 5, 6, 7].

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.

At ATR we are now in the second year of a Soumusho 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 speakers are actively engaged in discussion, amounts to less than 15% of the meeting.

This encourages us to pursue future research by focussing of our resources on identifying the single main speaker at any given time, rather than attempting to monitor all of the speech activity in the meeting.

2 Categories of Speech Activity

We have been regularly recording our monthly research project meetings, where research results and project planning details 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 is voluntary, but since the research is being carried out by three teams at different locations (ATR, NAIST, Kobe University) the meetings provide an essential focus-point for coordinating research activities.

All meetings are recorded on both video and audio, using purpose-built equipment that has been described elsewhere [8, 9, 10]. All observable 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 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 find 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 Analysis & Results

This paper reports the results from 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 from Ireland (s5) who is unrelated to the project but attended as an observer. 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

* ミーティング対話音声データの同時発話について
ニック キャンベル Department of Cognitive Media Informatics,
ATR Media Information Science Labs, Keihanna Science City, Kyoto, 619-0288, Japan,

Table 2 Counts of turns per meeting participant

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

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

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 active speaking turns in the meeting, which lasted approximately 45 minutes altogether. Here, a turn is defined as a continuous speech event, separated by a cessation of speech activity, from one speaker. The distribution of utterance turns 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 turn duration is 0.7 seconds (sd=0.78), with the longest recorded turn being 17 seconds. The 25th quantile of turn 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, informally questioning individual members about their goals and techniques.

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.

4 Discussion

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 solo.

The above analysis 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 considerably reduces the work-load of the processing. However, it will require separate technology to determine the reaction of the other participants to any particiular utterance or topic.

Table 4 Number of turns for each speaking type

on-topic	backchannel	private	laugh
2110	207	196	406

Table 5 Number of participants active at each turn

silent	solo	two	three	four
1730	2000	291	15	1

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. 305 ▪317, 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] 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.

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

[9] Project Homepage: http://feast.his.atr.jp/nonverbal

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