An Audio-Visual Approach to Measuring Discourse Synchrony in Multimodal Conversation Data

Nick Campbell

Trinity College Dublin

nick@tcd.ie

Abstract

This paper describes recent work on the automatic extraction of visual and audio parameters relating to the detection of synchrony in discourse, and to the modelling of active listening for advanced speech technology. It reports findings based on image processing that reliably identify the strong entrainment between members of a group conversation, and describes techniques for the extraction and analysis of such information.

Index Terms: discourse interaction, active listening, synchrony, image and audio processing, conversation

1. Introduction

The technical analysis of synchrony in active listening might be said to have started with the work of Eliot Chapple 70 years ago [1]. He and his co-workers adapted a manual typewriter (a machine pre-dating the word-processor and consisting mainly of levers, gears, and cogs) by fitting a small electric motor to the rubber shaft so that an operator could accurately record changes in subject activity over time, using a roll of adding-machine paper, for subsequent analysis. By means of this device, Chapple was able to observe the discourse actions of two individuals and to obtain a series of durations of their actions. His concern was to understand the way in which this sequence of durations is arranged, but flaws in his technique made it difficult to quantify overlapping speech and silent periods within the speech of one uninterrupted partner. However, this was the first recorded sequential analysis of human discourse behaviour, though an earlier study had made similar observations with the goal of obtaining only percentage behavioural data [2].

Subsequent schools of Conversation Analysis and later Discourse Analysis have focussed on such sequence organisation as the underlying core of their work [3]. Goffman [4] was perhaps the first to observe the systematic, socially organised procedures underlying the ways in which social actors move into mutually ratified participation in an encounter, which more recently Kendon has referred to as 'frame attunement' [5].

Kendon places such research in the realms of human behaviour, rather than linguistics or conversational content analysis, and thus eliminates the necessity for an understanding or even a processing of the linguistic content of such interactions:

> The first task of a human ethologist, like that of an ethologist who sets out to study a bird or a fish or a monkey, must be systematic description. He must set out to see what behavioural structures the human being has ... In doing this with people it would seem best to begin with those aspects of behaviour which are most likely to be shared with other animals ... Thus while detailed analyses of language ... must eventually find a place in human ethology, these do not seem to be the best aspects of human behaviour with which to start (from [6] as cited by [7]).

In describing the sequence of moves in social conversational interaction, Kendon explains that "... they each contribute to the emergence of a jointly sustained system of coordinated action patterns and [that] the emergent 'Common Understanding' is or may be the cognitive consequence of this" [8]. The present work aims to produce technology for the modelling of such discourse moves in conversational speech, and focusses on the *behaviour of the participants* in developing a technology to make inferences about their discourse participation status, rather than focussing on the text or interpretation of their speech.

More recently, large amounts of technology have been developed with the goal of monoitoring and modelling human social interaction through talk. International research projects such as AMI [9] and CHIL [10] have institutionalised such research, and have developed sophisticated apparatus for observing, recording, and analysing human spoken interactions from many simultaneous viewpoints and modalities. They have produced a technology and methods for the automatic processing and archiving of such data. Their work is perhaps the latest technological offspring from a long series of more subjective attempts by Conversation Analysts to produce detailed descriptions of how people manage their social interactions through structured discourse participation. This might be considered as 'language' independent.

Figure 1: *The 360-degree lens assembly taken from a SONY RPU-C251device (left), and fitted to a PointGrey Flea2 industrial video camera (right) and a similar Palnon lens on a small firewire Firefly camera (bottom left).*

2. Technology for listening to a conversation

This paper extends previous work presented in [11] and [12] on patterns of speech and silence, and patterns of overlapping speech, observed in a corpus of 100 30-minute telephone conversations [13], and links it with subsequent analyses of a corpus of 3 90-minute round-table conversational interactions [14] that were recorded using a range of multimodal sensors including a small 360-degree table-top video camera (shown in Figure 1).

Figure 2 shows a typical screenshot from an interactive web page[1] showing discourse behaviour as observed in the recorded telephone conversations. The bars in the figure show the speech activity of two people during the first thirteen minutes of their sixth telephone conversation and it is clear that Pink (the upper bars) dominates the conversation after the first seven minutes, while Blue (the lower bars) takes the lead during the first six minutes of the 30-minute conversation.

This type of plot reveals much information about the structure of the conversation without requiring any indication of what was being said. Long continuous stretches of speech activity most likely indicate parts of the conversation with high propositional content; while the short bursts of overlapping activity probably indicate backchannel utterances showing agreement or interest. Previous work has attempted to model this activity and the balance of dominance throughout the 100 conversations [15]. It is interesting to note that the concepts of 'turn' and of 'utterance' are particularly difficult to define when examining this type of natural and highly-overlapping conversational data.

[1]The page (http://feast.atr.jp/cgi-bin/mnatr/taba/esp_c/...) has disappeared temporarily due to the closure of the ATR SLC Labs in Japan, but this and similar data will soon be made publicly available again from the SSPNET Social Signal Processing web site [16].

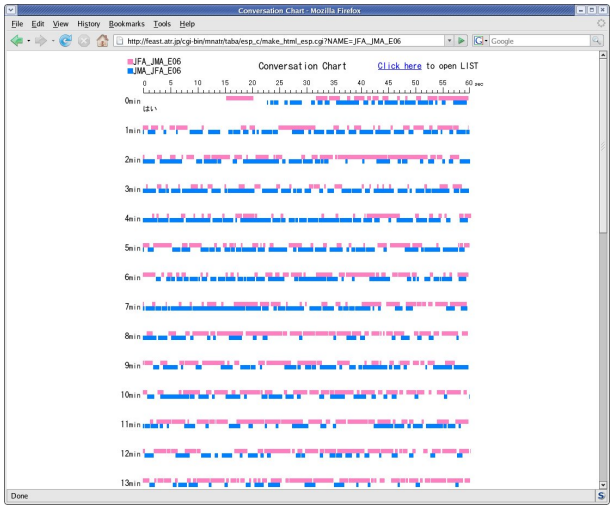

Figure 2: *Schematic speech on/off activity, one-minute per line, from a telephone conversation, showing the high degree of overlap and the shifting dominance patterns throughout the conversation.*

3. Active Listening

Traditional approaches to Multimodal Interface design have tended to assume a 'ping-pong' or 'push-to-talk' approach to speech interaction wherein either the system or the interlocuting human is active at any one time. This is contrary to many recent findings in conversation and discourse analysis, where the definition of a 'turn', or even an 'utterance' is found to be very complex; people don't *take turns* to talk in a typical conversational interaction, but they each contribute actively to the joint emergence of a 'common understanding' through a process of 'Active Listening' (AL) in which both participants *mutually* interact, frequently overlapping their speech.

As defined in the Wikipedia, AL is "a structured way of listening and responding to others [which] focuses attention on the speaker. Suspending ones own frame of reference and suspending judgment are important in order to fully attend to the speaker." specifically, "Having the ability to interpret a person's body language allows the listener to develop a more accurate understanding of the speaker's words" ([17] quoted in Wikipedia).

In contrast to the popular definition above, which concerns *attention paid by the listener* and presupposes special 'techniques of listening', the present author would instead claim that AL is more a process whereby *both* participants actively engage in the discourse in an overlapping and complementary manner, as illustrated in Figure 2 above, and where the focus is on contributory and participatory discourse actions, rather than on the cognitive attention states of the listener. These are physical observables that can readily be measured.

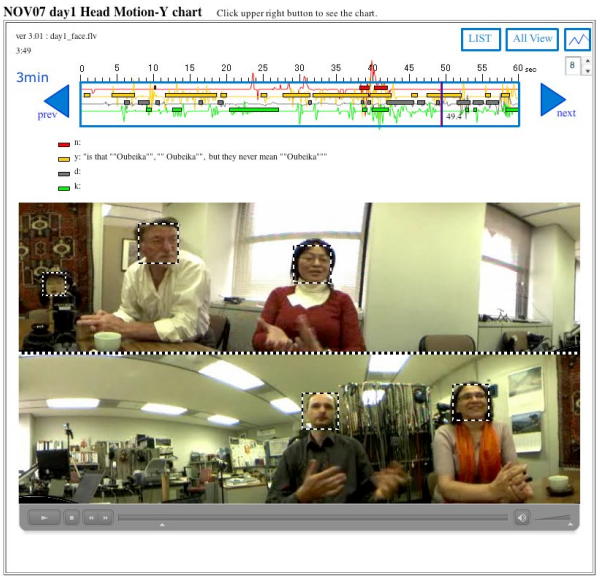

Figure 3: *Multimodal conversation data capture and analysis, showing colour-coded speech activity per person in an interactive flash movie framework. Subtitles and activity plots are the result of manual transcription.*

4. Synchrony

Figure 3 shows a scene from day 1 of the three multimodal conversation recordings. The cursor on the data-plot shows a moment, just after a burst of activity, when there is particular engagement in the discussion. The figure shows not only that all participants are focussed on a common space (represented by the hands of speaker top-right) but also that they are close to taking a common pose. Careful extended examination of this material for a Mumin analysis [18] has revealed a remarkably high number of times when all four participants are in close synchrony of movement, pose, and action.

The figure shows a complex data plot, with colour-coded talk & silence bars indicating speech activity for each participant overlaid on a movement display (in this case horizontal head movement) similarly colour-coded. This latter data stream is produced fully automatically in real time by image-processing based on head-detection using a modified Viola-Jones algorithm [19]. Head detection is very reliable, and bodies are assumed in the space 2.5 times the head-width immediately below. All movement in these areas is tracked in two dimensions, with an estimate also made for forward-backward movement from changes in detected head size [20].

Figure 4 illustrates these traces (low-pass filtered in this case) showing how precisely the movements align. Whereas we were expecting a cascade of movements as one participant reacts to an event from another, it appears that a window-width of 1 frame is sufficient to capture many of these synchronous movements.

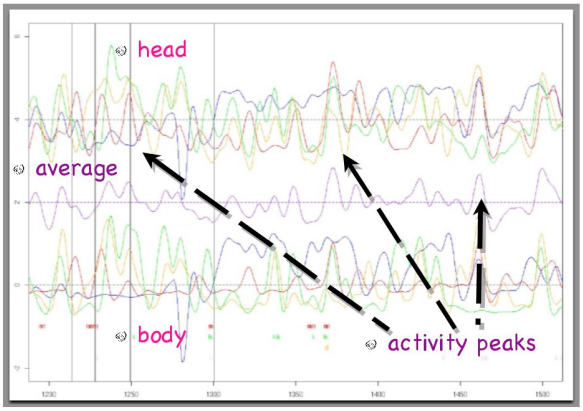

Figure 4: *Schematic diagram of joint action, showing colour-coded movement traces, derived automatically from the image processing, for head and body of each of the four participants, with an average 'overall movement' trace in the middle. Arrows show peaks of joint simultaneous movements. The plot covers about 5 minutes of interaction*

5. Discussion

Activity peaks in the movement indicate bursts of high interaction in the conversation, as illustrated from the manual transcriptions in Figure 5. There are clear sequences of propositional content, and bursts of high activity at the transitions. These are points of high engagement, indicating key points of the interaction. Figure 6, showing automatic measurements, is almost a perfect reflection of Figure 5 which required human judgement. These typically correlate at rates higher than 0.8, rendering manual transcription redundant.

6. Conclusions

This paper has presented material derived from multimodal recordings of multi-party conversational interactions, showing that participants engage positively in a discourse, synchronising their speech and movements to a very high degree, and frequently speaking and moving simultaneously at points of high engagement. Previous findings based on telephone conversations (in Japanese) were confirmed in the multi-party round-table conversation data (in English). Particular note should be taken of the high amount of overlapping speech at regular periods throughout the discourse, and of the use that can be made of these for detecting topic changes and participation status.

A key point made in the paper is that whereas previous work required lengthy and expensive manual transcription of the data, the proposed automatic procedures derived from simple image processing show a very high correlation with the transcribed speech activity.

Figure 5: *Manually obtained measures of discourse interaction. The areas where all participants are active are of particular interest as indicators of high discourse engagement, marking key events in the conversation*

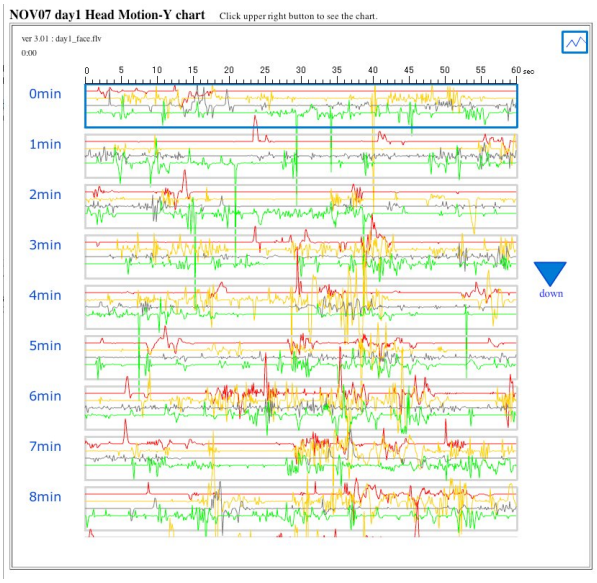

Figure 6: *Automatically derived measures of discourse engagement, based on image processing from the 360-degree capture device. Note the high correlation with data from the manual transcription.*

7. Acknowledgements

Much of this work was carried out in Japan while the author was employed by the National Institute of Information and Communications Technology (NiCT), and by the Advanced Telecommunications Research Institute (ATR) in Kyoto. It is being continued at Dublin University, Trinity College, with grateful thanks to the Science Foundation Ireland (SFI) and with partial support from the NiCT through the Japanese Government 'kaken' funding. With particular thanks to Tabatan for her skillful programming.

8. References

[1] Chapple, Eliot, D., (1939) "Quantitative analysis of the interaction of individuals", pp.58-67 in Proceedings of the National Academy of Science U S A. February; 25(2).

[2] Thomas, Dorothy A., Loomis, Alice M., and Arrington, Ruth, E., (1934) "Observaional studies of social behaviour", pp.194-195 in Annals of the American Academy of Political and Social Science, Vol. 172, No. 1.

[3] Schegloff, Emanuel A. (2007) *Sequence Organization in Interaction: A Primer in Conversation Analysis*, Volume 1, Cambridge: Cambridge University Press.

[4] Goffman, Erving (1961) *Encounters*. Indianapolis: Bobbs-Merrill.

[5] Kendon, Adam, (1990) *Conducting Interaction: Patterns of Behaviour in Focused Encounters*. Cambridge: Cambridge University Press.

[6] Kendon, Adam and Andrew Ferber (1973) "A description of some human greetings". In R. P. Michael and J. H. Crook (eds.) *Comparative Ecology and the Behaviour of Primates*. London: Academic Press. 591668.

[7] Hutchby, Ian, (1999) "Frame attunement and footing in the organisation of talk radio openings", *Journal of Sociolinguistics* 3/1, pp.41-63.

[8] Kendom, Adam (2008) personal communication to the author.

[9] AMI: Augmented Multi-party Interaction (http://www.amiproject.org)

[10] CHIL: Computers in the Human Interaction Loop (http://chil.server.de/)

[11] Campbell, Nick (2007) "On the Use of Nonverbal Speech Sounds in Human Communication", pp.117-128 in *Verbal and Nonverbal Communication Behaviors*, LNAI Vol.4775, .

[12] Campbell, Nick (2008) "Individual traits of speaking style and speech rhythm in a spoken discourse", pp.107-120 in Esposito, A., et al (Eds) *Verbal and Nonverbal Features of HH and HM Interaction*, Springer-Verlag, Berlin.

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

[14] Campbell, Nick (2006) "A multimedia database of meetings and informal interactions for tracking participant involvement and discourse ow", in Proc LREC 2006, Lisbon.

[15] Campbell, Nick (2008) "Multimodal Processing of Discourse Information; The Effect of Synchrony," ISUC, pp.12-15, 2008 Second International Symposium on Universal Communication, Osaka, Japan.

[16] SSPNET: Social Signal Processing (http://www.sspnet.eu/)

[17] Atwater, Eastwood (1981). *I Hear You*. Prentice-Hall. p. 83. ISBN 0-13-450684-7.

[18] The MUMIN coding scheme for the annotation of feedback, turn management and sequencing phenomena by: Jens Allwood, Loredana Cerrato, Kristiina Jokinen, Costanza Navarretta, Patrizia Paggio, in *Language Resources and Evaluation*, Vol. 41, No. 3. (16 December 2007), pp. 273-287.

[19] Viola, P., and Jones, M., "Robust Real-Time Face Detection", International Journal of Computer Vision, vol. 57, no. 2, pp.137-154, May 2004.

[20] Nick Campbell, Damien Douxchamps (2007) "Robust real time face tracking for the analysis of human behavior", pp.1-15, in *Machine Learning & Multimodal Interaction*, Springers LNCS series, 4892.