Head and Hand Movements in the Orchestration of Dialogue

Stuart A. Battersby (stuart@dcs.qmul.ac.uk)
Queen Mary, University Of London
Interaction, Media & Communication Group
School Of Electronic Engineering & Computer Science
London, E1 4NS

Patrick G. T. Healey (ph@dcs.qmul.ac.uk)
Queen Mary, University Of London
Interaction, Media & Communication Group
School Of Electronic Engineering & Computer Science
London, E1 4NS

Abstract

Gaze and head orientation are considered to be the most important non-verbal cues people use to help manage the flow of conversation. However, if there are more than two participants, gaze and head orientation become problematic. People can only look at a single participant at a time. When speakers concurrently engage with more than one participant, they often make use of both head and hand orientation. We show two contrasts with existing findings. Firstly, people do not automatically look where the speaker is looking. Secondly, we demonstrate that hand movements are more important for the interaction than head movements. Specifically, changes in speaker hand orientation prompt quicker and more frequent responses from recipients than changes in head orientation.

Keywords: Dialogue; Non-verbal interaction; Multi-party; Gesture; Gaze; Simultaneous engagement;

Introduction

Consider the following situation: Ann, Bob and Claire are discussing a film that Bob and Claire went to see the previous night. Ann asks "Was it good"? Claire responds by saying "I really enjoyed it" while Bob simultaneously pantomimes a yawn. More than one person's responses are potentially relevant to the interpretation of the answer. Moreover, the orientation of each participant to those responses is also relevant. For example, it matters whether Bob is looking at Ann or Claire as he pantomimes a yawn and it also matters whether Claire is aware that he is looking at her when he yawns.

Putting puzzles about mutual-knowledge to one side, this example highlights the intuition that in multi-party interactions participants often face the challenge of simultaneously monitoring the responses of several people to each contribution (see Goodwin (1979)). People can also design their contributions in ways that directly convey how different participants stand in different relationships to what is said. In a variation of the example above, Claire might look at Ann and say "I really enjoyed it but Bob was bored" while simultaneously pointing toward Bob as she says his name (see Healey and Battersby (2009), for documented examples of this kind).

In the literature on non-verbal communication, a significant body of evidence has accumulated that shows gestures have managerial functions within dialogue (see Bavelas, Coates, and Johnson (2002) and Jokinen and Vanhasalo (2009)). However, eye gaze and, by association, head-orientation are normally singled out as the most important cues to the current orientation of participants in interaction (see, for example, Argyle (1975)). Kendon (1990) uses the term 'Face Address System' to make the claim that speakers use their gaze to identify the intended recipient of their utterance and Streeck (1993) observed that it is the speaker's gaze that addressees follow, potentially to the speaker's gesture. Langton, Watt, and Bruce (2000) reflect upon the claims about gaze and although they agree on its importance, suggest that gaze cues should be considered along with cues from the head orientation and hands.

Gullberg (2003) provides a quantitative estimate of the relative importance of a speaker's face and hands by measuring the eye-gaze patterns of addressees. Her live condition consisted of two people one of which had watched a cartoon. This person then retold the cartoon in narrative form to an addressee who had been configured with eye tracking equipment. The gaze patterns of this addressee were recorded. Only 7% of the speaker's gestures were fixated by the addressee. 96% of the time the addressee looked at the speaker's face; only 0.5% of the time was spent on their gestures with the remaining time spent looking at other objects in the room. Whilst this data points to a marked difference in the relative importance of the head and the hands, the interactional situation is different to open multi-party conversation.

Coordinating Multi-Party Interactions

Although eye gaze is an effective cue to focus of attention in dyadic (two-person) interactions it has more limited value in multi-party interactions. We can only look at one person at a time and we can only monitor the gaze of one person at a time. As Loomis, Kelly, Pusch, Bailenson, and Beall (2008) have shown, direction of eye gaze is difficult to estimate in the physical arrangements typical of conversation. In small group conversations people are only able to judge another's eye gaze direction with a maximum 4° retinal eccentricity whereas other people's head orientation can be judged effectively up to a 90° retinal eccentricity. This leads to the pre-

diction that, in multi-party conversations, auxiliary cues such as head and hand orientation should therefore be much more important to the conduct of the interaction.

Healey and Battersby (2009) describe how in three-way task-oriented dialogues speakers frequently use combinations of head and hand orientation to enable simultaneous engagement with two other participants. These moments of simultaneous engagement occurred on average once every 25 seconds. However, it is unclear what the consequences of the events are for the other participants in the interaction. Specifically, do these head and hand movements have any demonstrable impact on the responses of the other participants?

This paper addresses the question of whether changes in a speaker's orientation reliably prompt changes in the behaviour of the other participants. It also compares the relative impact of head and hand movements on other participants.

Method

Materials

All data was gathered in the Augmented Human Interaction (AHI) lab at Queen Mary. This lab houses a Vicon optical motion capture system consisting of an array of 12 infra-red cameras which track reflective markers attached to the clothing of participants. Each participant wears an upper body motion capture suit and a baseball cap with reflective markers attached. The motion capture system records the precise 3D coordinates of each marker at a rate of 60 frames per second (see Battersby, Lavelle, Healey, and McCabe (2008) for more details). Video cameras are placed above and to either side of the participants and are time synchronised with the motion capture system. Audio is recorded on the video cameras. Motion capture data from each interaction is time synchronised with the video data. A custom piece of software reads the motion capture data and integrates it with hand-annotation data from ELAN.

Participants

Participants were recruited from undergraduate and masters courses at Queen Mary and either received pay or modules credits and pay for their time. 33 participants (19 female and 14 male) aged between 18 and 30 took part. Each group consisted of three people meaning that the data presented are from 11 triads.

Task Description

Six tuition tasks were developed that consisted of a description of either a short Java program or a description of a system of Government. They were designed to involve an abstract hierarchy with no direct visual analogue. All material was text based with no graphical descriptions.

Procedure

Each group completed three rounds, based on either three Java or three Government tasks. On the first round one member of the triad is randomly assigned to a 'learner' role and the other two participants are assigned 'instructor' roles. These roles are then rotated on subsequent rounds so that each participant is as a learner once and an instructor twice. The instructors are asked to collaborate to teach the leaner the structure described in the task description. The learner is removed from the group to another room whilst the instructors are given the descriptions of the task for next round. Once the instructors signal that they understand the task, the descriptions are returned to the experimenter and the learner rejoins the group. All three participants are seated on pre-positioned stools in the AHI lab (see Fig. 1).

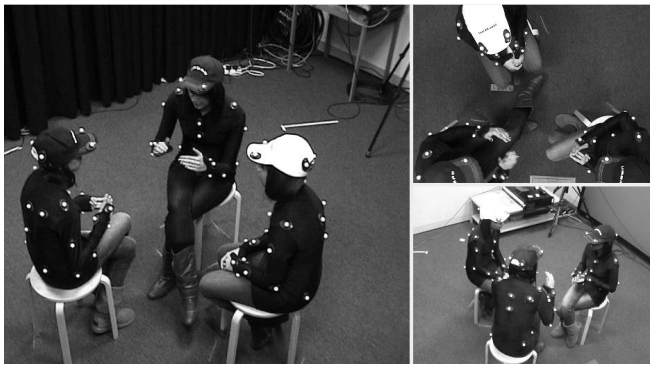

Figure 1: Three participants wearing upper body motion capture suits.

There was no time limit on the tuition stage and no restrictions on the interaction other than they were not allowed to use pens or paper. The participants notified the experimenter when they finished each round of tuition. To motivate the participants to adequately teach and learn the material a post completion test (comprising of a drag and drop arrangement of the classes for the computer program, or some multiple choice questions for the government material) was given after each round. Tasks were systematically rotated across groups and the order of the printed sheets of paper was randomised before each round.

Hand Coding of Target Events

All interactions were recorded on video, with cameras above and either side of the group, using synchronised video recording. ELAN was used to hand code these videos. The recordings were coded for all instances of simultaneous engagement in which a speaker who is making a gesture visibly changes the orientation of their hands or head with respect to the other participants. For example, by turning their hand from one participant to another or changing their head orientation. These changes were coded as:

- Head Moves: Here the head orientation changes but the gesture remains stationary
- Hand Moves: Here the gesture orientation changes, but the head orientation remains stationary
- Both Move: Here both the gesture and the head orientation shift

Motion Data Analysis

Taking the hand coded target events for the speaker as the starting point, the motion capture data was used to provide quantitative measures of recipients' responses to target events.

Assigning Recipient Role The motion data was used to provide an operational definition of recipient role. Recipients are either primary or secondary recipients. This role is judged by the head orientation of the speaker. We project a vector from the middle of the forehead for each speaker. The orientation of this vector is compared to a centre line between the two recipients. The primary recipient is defined as the recipient who is on the same side of the centre line as the speaker's current head orientation (see Figure 2).

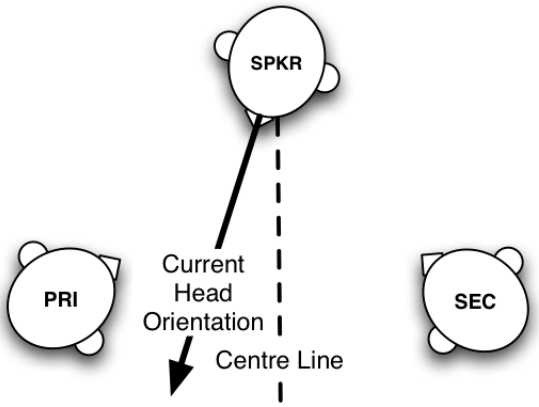

Figure 2: Defining primary and secondary recipients

Indexing Head Orientation Responses It is impossible to be sure exactly which head movements correspond to changes of orientation by the recipients. Instead, we set a criterion for counting movements as changes of orientation based on a vector projected from each recipient's head as it was for the speaker. A change in orientation is thus defined as a shift of head orientation that crosses the centre line between the speaker and the other recipient (see Figure 3).

Indexing Nod Responses A second index of responses, 'nods', was also generated from the motion capture data. As for changes in head orientation it is impossible to be sure when a head movement really constitutes a nod or is simply a shift in position or unintentional motion of some kind. As for

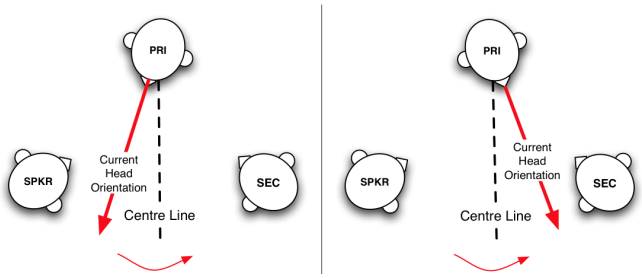

Figure 3: Indexing head orientation responses

changes in head orientation we set a criterion level of movement of a single frontal head marker in the vertical axis (see Figure 4 for some sample movement). Only movement with a frequency of between 2Hz and 8Hz is used. This removes some of the effects of gross body sways (below 2Hz) and very minor body shakes or fluctuations in data from the cameras (above 8Hz). Movements with an amplitude greater than 5cm are removed as these could likely be a result of shifts in position. The resulting signal, which is smoothed using a window size equivalent to 0.5 seconds, is used to represent periods of head movement that approximate nodding.

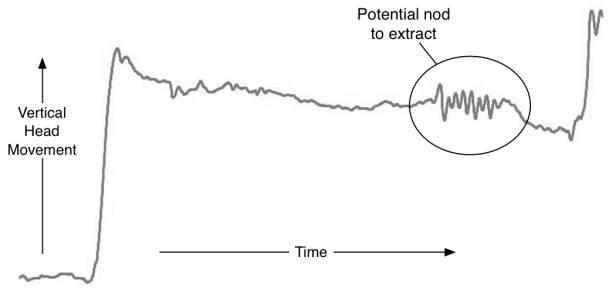

Figure 4: Raw head movement motion capture data. An area of potential nodding is circled.

In order to analyse frequency of responses to each simultaneous engagement event by the speaker we create a 5 second window after the event and score, for each recipient, whether a head re-orientation and whether a head nod occurs in that window. In order to provide a measure of response latency we record the first change of head orientation or nod that occurs after the target event and before another target event occurs.

Baseline Response Rates In order to interpret the measures of responses to the target events, it is important to know what the baseline likelihood of a recipient nodding or changing orientation is. To provide this a control comparison sample was created by randomly selecting points where someone was speaking but not producing a target event. Recipient

responses after these control points were then analysed in exactly the same way as for the target events.

Results

The total time for all the dialogues was 2 hours and 54 minutes, each task took on average 5 minutes and 16 seconds. A total of 287 target, simultaneous engagement events involving a change in orientation of the speaker were identified.

Inter-rater Agreement

In order to check the reliability of the hand coding of event types by the 1st author, a random sample of 25 events taken from experimental and control data was independently coded for event type by a second coder. The inter-rater reliability was good with Kappa = 0.78,($p < 0.001$). The number of each type of target event is show in Table 1.

Table 1: Number of Changes in Speaker Orientation

Event Class	Count
Head Moves	170
Both Move	86
Hand Moves	31

Recipient Responses

In analysing responses to changes of speaker orientation we distinguish the task role of the recipients (learner or instructor) and their recipient status at the time of the event (primary or secondary). In addition we code whether each recipient is oriented toward the speaker or the other recipient at the time the simultaneous engagement event, i.e. the change in speaker orientation, begins. These judgements are made using the motion capture data.

Recipient Orientation

As Table 2 shows, at the point when the speaker initiates a change of orientation, the primary recipient is more likely to be looking at the speaker than the secondary recipient. The secondary recipient, by contrast, is equally likely to be looking at the other participants ($\chi^2 = 16.9$, $p < 0.001$)).

Table 2: Initial Head Orientations

Recipient Role	Oriented To Speaker	Oriented To Other
Primary Recipient	68.0%	32.0%
Secondary Recipient	50.2%	49.8%

Response Frequency

In contrast to recipient orientation (and our preliminary findings (Healey & Battersby, 2009)), there was no difference between the response rates of the primary and secondary recipients. Both were equally likely to respond.

Combining the responses of the two recipients together we can compare the overall frequency of response to a target coordination event with the baseline response rate. For changes in head orientation the recipients' baseline response rate is 41.3% and their response rate to target events is 48.6%; a small but reliable difference ($\chi^2 = 5.75$, $p < 0.05$).

Table 3 illustrates the differences in response rate for each type of event.

Table 3: Response rates by type of event, measured by recipient reorientations

Event Class	Response Rate	Baseline Response Rate	Sig
Head Moves	43.1%	41.3%	Not Significant
Both Move	56.2%	41.3%	$\chi^2 = 10.26$, $p < 0.01$
Hand Moves	63.0%	41.3%	$\chi^2 = 8.14$, $p < 0.01$

For target events in which only the head changes orientation there is no significant increase in response rate (measured by a shift in recipient head orientation) relative to the baseline rate. However, for targets events that involve changes to both gesture and head orientation we see a significant difference of 14.9% between the baseline and the target event. Where only the gesture changes orientation there is a 21.7% increase in response rate.

A slightly different pattern is evident in the head nodding response measure. Combining target events, recipients respond 72.4% of the time compared to a background response rate of 66.0% ($\chi^2 = 5.08$, $p < 0.05$). The breakdown by type is shown in Table 4.

Table 4: Response rates by type of event, measured by recipient nodding

Event Class	Response Rate	Baseline Response Rate	Sig
Head Moves	70.0%	66.0%	Not Significant
Both Move	73.6%	66.0%	Not Significant
Hand Moves	87.0%	66.0%	$\chi^2 = 8.51$, $p < 0.01$

In order to provide a direct comparison of the recipient's relative sensitivity to changes in the speaker's head and hand orientation responses to 'Head Moves' events and 'Hand Moves' events can be compared. This shows a significant difference between the groups using the values for both head

re-orientations and head nods as a measure of response shown above ($\chi^2 = 6.43, p < 0.02$ and $\chi^2 = 5.75, p < 0.02$ respectively).

Response Latency

The time elapsed between a target event until the first response (nod or change of head orientation) for each recipient was analysed in a Mixed Model linear analysis with Recipients and Task as random factors and 'Condition' (Target Event vs. Baseline) and Task Role (Learner vs Instructor) as within subjects factors. This showed a reliable main effect of Condition ($F_{(1,1089)}$ = 14.88, p = 0.00) but no main effect of Task Role ($F_{(1,1088)}$ = 1.29, p = 0.25) and no Task Role $\times$ Condition interaction ($F_{(1,1078)}$ = 0.39, p = 0.53).

As Table 5 shows, recipients' responses to target events were approximately 1 second faster than the baseline responses.

Table 5: Marginal Means for Recipient Response Times

Condition	Marginal Mean	Standard Error
Target Event	2.4 seconds	0.37
Baseline Event	3.4 seconds	0.35

Discussion

The results show two important contrasts with existing findings on non-verbal cues and the co-ordination of interaction. First, in the dialogues reported here people do not automatically look where the speaker is looking. In fact, in the cases where the speaker only changes their head orientation there is no reliable shift in recipient's head-orientation. The second key finding is that changes of hand orientation are significantly more likely to invoke a response from the recipients than changes in head orientation; the opposite of what would be predicted on the basis of previous work.

The results also show that recipients are demonstrably responsive to the target events, but with a pattern of responses that is different to that typically described in the literature. This provides support for the claim that they are distinctive and significant interactional events. Although it is difficult to generalise beyond the particular task we have used, it seems likely that these moments of simultaneous engagement are a response to the demands of co-ordinating a conversation with multiple participants. As Healey and Battersby (2009) note, they are also distinguished by relying on physical co-presence in mutually shared space as a specific resource for interaction. For example, they cannot be deployed in point-to-point video communication.

Our analysis suggests that recipient role (primary or secondary) can manifest itself non-verbally. Whilst hand movements are more marked than head movements in initiating recipient responses, we see differing patterns of recipient head orientation through the dialogue. The primary recipient is more likely to be looking at the speaker than they are to be looking at the secondary recipient before a simultaneous engagement event occurs. Secondary recipients do not share this pattern though, and are equally likely to be looking at either party. It is interesting that this distinction between roles is not found when measuring responses, perhaps suggesting that the target events unify the recipients' behaviour.

The clear difference between our data and that of previous findings is the introduction of the third person. It would be intuitive, and logical, to understand the conflicting results with the statement that multi-party dialogue is simply different to dyadic dialogue. Whilst this is true, there is also the possibility that multi-party dialogue only allows us to see fully the underlying process that is present in *all* dialogue; dyadic interaction simply masks them.

Conclusion

We examined a corpus of multi-party dialogues comprising of video and motion capture data. Moments where the speaker simultaneously engaged both recipients were coded for. These events were broken down by changes in the speaker's orientation of their head, their gesture or both and the significance of these changes for the recipients was examined. These changes in speaker orientation were shown to hold interactional significance. In contrast to existing findings in the literature, movements of the hands elicited a higher and faster response rate than movements of the head.

References

Argyle, M. (1975). *Bodily Communication*. Bristol: Methuen & Co. Ltd.

Battersby, S. A., Lavelle, M., Healey, P. G. T., & McCabe, R. (2008, May). Analysing Interaction: A comparison of 2D and 3D techniques. In *Conference on multimodal corpora.* Marrakech.

Bavelas, J. B., Coates, L., & Johnson, T. (2002). Listener Responses as a Collaborative Process: The Role of Gaze. *Journal of Communication*, *52*, 566–580.

Goodwin, C. (1979). The interactive construction of a sentence in natural conversation. In G. Psathas (Ed.), *Everyday language: Studies in ethnomethodology* (pp. 97–121). Irvington Publishers.

Gullberg, M. (2003). Eye movements and gesture in human face-to-face interaction. In J. Hyönä, R. Radach, & H. Deubel (Eds.), *The mind's eye: Cognitive and applied aspects of eye movements* (pp. 685–703). Oxford: Elsevier.

Healey, P. G. T., & Battersby, S. A. (2009). The Interactional Geometry of a Three-way Conversation. In *Proceedings of the 31st annual conference of the cognitive science society* (pp. 785–790). Amsterdam.

Jokinen, K., & Vanhasalo, M. (2009). Stand-up Gestures Annotation for Communication Management. In *Nodalida*

2009 workshop multimodal communication: from human behaviour to computational models (pp. 15–20).

Kendon, A. (1990). *Conducting Interaction: patterns of behavior in focused encounters*. University of Cambridge.

Langton, S. R. H., Watt, R. J., & Bruce, V. (2000). Do the eyes have it? Cues to the direction of social attention. *Trends in Cognitive Sciences*, *4*(2), 50–59.

Loomis, J. M., Kelly, J. W., Pusch, M., Bailenson, J. N., & Beall, A. C. (2008). Psychophysics of perceiving eye and head direction with peripheral vision: Implications for the dynamics of eye gaze behaviour. *Perception*, *37*, 1443–1457.

Streeck, J. (1993). Gesture as Communication I: Its Coordination with Gaze and Speech. *Communication Monographs*, *60*(275-299).