Multimodal Feedback in First Encounter Interactions

Kristiina Jokinen

University of Helsinki, Finland
kristiina.jokinen@helsinki.fi

Abstract. Human interactions are predominantly conducted via verbal communication which allows presentation of sophisticated propositional content. However, much of the interpretation of the utterances and the speaker's attitudes are conveyed using multimodal cues such as facial expressions, hand gestures, head movements and body posture. This paper reports some observations on multimodal communication and feedback giving activity in first encounter interactions, and discusses how head, hand, and body movements are used in conversational interactions as means of *visual interaction management*, i.e. unobtrusive ways to control the interaction and construct shared understanding among the interlocutors. The observations and results contribute to the models for coordinating communication in human-human conversations as well as in interactions between humans and intelligent situated agents.

Keywords: multimodal interaction, feedback, nodding, head movements.

1 Introduction

Natural communication does not only include verbal utterances but a wide variety of non-verbal means, ranging from physiological displays to paralinguistic vocalisations and head movement, hand gestures, and body posture. They all have important functions in the communication as whole, as they provide tacit cues of the interlocutors' emotional state and also allow the speaker to control the conversation and manage feedback and turn-taking. Physiological signals can effectively indicate the interlocutor's emotional state, and usually they are unintentional (e.g. we blush for embarrassment, or our pupils dilate for surprise), while hand gesturing, body and head movements, gaze, and facial expressions appear as more controlled means of communication, although they also can indicate the speaker's emotions and focus of attention in a spontaneous and automatic manner. The speakers need not be fully aware about their behaviour, e.g. tilting one's head, beating one's hand, or swaying one's body can be typical behaviours for a speaker, but not necessarily something that the speaker intentionally aims to act like. [1] talks about the degrees of intentionality and awareness in bodily communication, and discusses the concepts of *indicate, display,* and *signal* to specify three different degrees of intentional behaviour. "Indicate" denotes the agent's actions lacking conscious intentionality (automatic reactions like blushing), while the two others are associated with greater degrees of awareness and intentionality: "display" refers to the agent showing something intentionally, and

M. Kurosu (Ed.): Human-Computer Interaction, Part IV, HCII 2013, LNCS 8007, pp. 262–271, 2013.

"signal" is a second-order display, implying that the communicator does not only display a meaning, but also their intention that the partner understands their intention to display the meaning. Because of its more spontaneous nature, non-verbal communication is often regarded as more truthful or authentic expression of the speakers' meaning than their verbally expressed utterances. Although it is difficult to identify behavioural cues that would reliably indicate the speaker's truthfulness (or deceit) in general, it is possible to detect non-verbal behaviour patterns that characterize individual speakers as a whole and then to look for changes in their behaviour that can be used as an indication of their emotional, intentional, and attentional state. In fact, such indirect lie detection methods have been successfully used in recent deception studies and they have produced more accurate results than the attempts to identify specific behaviours that are thought to be related with lying. For instance, [8] report that their subjects could distinguish liars from truth-tellers more accurately, if they were asked to identify changes in the people's behaviour rather than explicitly asked to look for liars.

In everyday interactions the processing of multi-modal information is necessary for smooth communication, and much of this socially conditioned. For instance, recognition of social signals affects the interpretation of the partner's message as a humorous or a sarcastic comment, and helps the construction of shared context and mutual understanding. The signals are also important to take into account when planning one's own contribution and intending to coordinate the flow of conversation. Relevant signals in this respect include gesturing to catch the partner's attention to a particular element of interest in the context, turning one's head to the speaker to show one's willingness to be engaged in the interaction, or looking away from the partner, to provide indirect cues of one's non-understanding or lack of interest in the presented message. Such social signalling models are useful for various human-computer applications where the goal is to build more natural interactive systems. We can say that multimodality increases the system's *affordance,* the concept brought to HCI by [21] and suggested by [9] to be used especially with respect to natural language interactive systems: the users need not spend extra time wondering how to operate the interface in order to get their task completed, as multimodal natural language techniques *afford* interaction and lend themselves to the intuitive use of the system.

This paper studies the interlocutors' multimodal activity, such as hand, head and body movement, from the point of view of feedback and construction of shared context. The paper is structured as follows. Section 2 discusses previous studies and sets the scene for multimodal feedback studies. Section 3 presents the corpus of First Encounters, including annotations and our methodology. Section 4 discusses our studies concerning conversational feedback, and reports the results. Section 5 concludes the paper and provides future prospects within intercultural dimensions of the work.

2 Multimodal Feedback

Explicit feedback is important to signal that the speaker's message got through to the partner, but simultaneously it also displays the partner's willingness to maintain good contact and rapport with the speaker. Earlier research has emphasised that multimodal

signals serve social functions by creating bonds and shared understanding but also convey information by reflecting the speakers' attitudes, mood, and emotions [7].

Other important functions deal with their use to regulate the flow of information. For instance, [14] talks about *meta-discursive* function of hand gestures, and shows how different hand forms represent semantic themes which are motivated by different communicative needs on the utterance level, or by communication management. For example, pointing gestures can direct the partner's attention to an important piece of information in the utterance, they can be used to halt the conversation, and they can also mark the next speaker. Although pointing is fairly a distinct gesture, it can also vary in its form: it can be made by an extended finger, an extended hand with open palm, or by a tool such as a pencil that the participant can manipulate in their hand.

Also eye-gaze is an effective means to give and elicit feedback. Gaze indicates where the speaker's focus of attention is directed, and so looking at the conversational partner or looking away from the partner can indicate the partner's understanding of the presented information or willingness to continue interaction. Gaze is also important in indicating if the speaker wishes to keep the turn although hesitant in their wording, or if the speaker wishes to offer the turn to the partner [10]. Mutual gaze is needed to agree on smooth turn-taking and grounding of information [e.g. 4, 16,19].

One of the most important feedback signals is head movement. Nodding is a common way to give acknowledgement and agree with what the speaker says, while side turns effectively signal the change in the participant's focus of attention. [20] compared head nods in three Nordic languages and noticed statistically significant differences in their frequency. [23] discuss nods in Finnish interactions and point out a difference in up-nods and down-nods, the former being mainly used if the speaker presents information that is somehow new to the listener while the latter is neutral acknowledgement of the presentation.

As for the body posture, leaning forward often means interest while leaning backward signals withdrawing from the conversational situation, and they can thus be used to control and coordinate interaction [13]. Some body movements are also used to fill pauses in conversation: the speaker does not want to take the turn or is unable to take the turn. In multiparty interactions, spatial configurations of the participants can show the participants' relation to each other and distinguish the primary and secondary recipients of the speaker [5].

2.1 Cooperation and Shared Context

The interlocutors cooperate with each other and construct a shared context by exchanging information [6, 9]. Communication can thus be seen as cooperative activity in which the interlocutors are engaged in. Cooperation can manifest itself on several levels, from tight task-based collaboration in order to achieve a particular goal, to behaviour patterns that occur simultaneously when the interlocutors interact. An important component of this process is to construct a shared context in which to achieve the shared goal. The construction of the shared context takes place via interactive evaluations of the partner's contributions, whereby the agents give feedback to each other on the current state of communication: if they are willing to continue in the

interaction and what is the level of understanding and agreeing with the partner on the information content exchanged. In the widest sense, feedback refers to the agent's response to the partner's utterance in general, i.e. it is a conscious or unconscious reaction to the changes that take place in the agent's communicative environment. Feedback in this sense is used to refer to the agent's evaluation of the basic enablements for interaction, so it is synonymous to the agent monitoring the interaction in general. Often feedback is understood in a narrower sense as a particular expression used to give feedback to the partner or to elicit feedback from the partner, on some communicatively relevant aspect of interaction. Multimodal feedback has been mainly regarded as displaying certain aspects of the speakers' cognitive and emotional state and thus they allow the interlocutors to monitor each other's emotions and understanding in a natural way.

As mentioned, the interlocutors provide feedback by head, hand, and body, besides explicit verbal feedback. Gestures, facial expressions, and eye-gazing are an unobtrusive means to construct shared context effectively, so that successful interaction can take place. In recent years, a number of studies concerning synchrony and cooperation has increased [12, 17, 18, 22]. Copying of each other's movements, gestures and body postures often occurs in conversations, and this kind of behaviour where the participants align their behaviour with each other can be understood as signalling cooperation between the participants building a common ground. In psycholinguistic and social interaction studies such synchronous behaviour is usually called alignment [22] or mimicry or copying [e.g. 18].

On the other hand, the function of multimodal signals can vary depending on the context. For example, forward leaning can be related to adjusting one's position, but it can also be interpreted as the partner finding the situation uncomfortable and wanting to leave. Backward leaning can display a relaxed participant in a happy listener position, or withdrawal from the conversation. Knowledge of the context is thus a key factor in understanding the function of multimodal signals in interactive situations. [14] argues that the meaning and function of gestures depend on the different relations they have to the surrounding context, i.e. their meaning is created in interaction of the linguistic and dialogue contexts in which they occur, see also [10].

2.2 Referential Schemas

Speakers make frequent verbal and non-verbal references to situation and language context. Their interpretations of linguistic expressions, and of the whole interaction, are influenced by observations from the spatial and visual environment. We can assume that when processing linguistic expressions, human cognition operates on the basis of particular action representations which representations can be defined as general schemas. The schemas can then be employed in different contexts with different multimodal means, and be formulated as scripts or frames or patterns.

Much research is being conducted concerning how senso-motor activity constitutes linguistic representations. Discussion has concerned their origin being based on experience or an innate skill that directs the interpretation. [24]. We emphasise the interaction between linguistic expressions and experience: the meaningful units, either

verbal or non-verbal elements, are created in interaction of the agents with their environment and the other agents. Every action is characterized as goal-oriented, and the signals which have been successfully used in communication to achieve a particular goal are likely to be repeated in future situations too. Their repetitive success reinforces their usefulness Referential schemas are learned though acting, observation and imitation. They are useful in that they allow the agent to plan their actions and estimate the outcome of their actions: based on their previous experience, the agent can select appropriate actions with the desired outcome, and anticipate the results of their actions.

Concerning feedback, a question is related to the agents' understanding of the relation between language and the physical environment, the embodiment of linguistic knowledge via action and interaction. The relation between high-level communication and the processing of sensory information is under intensive research, but it is obvious that motoric activity accompanies speech, e.g. [14] discusses synchrony of gestures and speech, and how gesture peaks coincide with stress in spoken utterances.

3 Data

The video recordings were collected within the Finnish part of the Nordic project NOMCO [20]. The goal of the project is to study the relation and correlations between speech and multimodal signals in face-to-face communication situations in the Nordic countries (Danish, Finnish, and Swedish data; later on a similar project was also started on Estonian data), and to compare conversational strategies in culturally rather similar yet linguistically different (Finnish vs. Danish and Swedish) contexts. To provide a comparable basis for data analysis, special focus was put on the similar collection setup (non-verbal communication between interlocutors meeting for the first time), and also on the use of a uniform annotation scheme [2], which enables

Fig. 1. Two participants interacting with each other for the first time

Fig. 2. The center camera view of the interactants

similar annotation categories and features to be used across the corpora. The Finnish data consists of a total of 16 interactions between 8 female and 8 male participants, average age about 23 years, each taking part in two separate conversations with different partners. The participants did not know each other in advance, and their task was to talk and to get to know each other. Each encounter was about 6-10 minutes long, and it started when one of the participants entered the video recording room where the other was waiting. Each participant was recorded by a separate video camera, and a third camera was positioned so that it recorded both partners simultaneously. Figure 1 shows screenshots of two participants interacting with each other, and Figure 2 is the corresponding center camera view of them.

The data was then annotated using the Anvil software [15] and the NOMCO annotation scheme [2], with respect to head, hand, and body movements that were considered carrying communicative functions. Of all the 645 head movements, 576 (89 %) were related to feedback: 375 gave feedback and 201 elicited feedback. Of all the 402 hand movements, 295 (73%) occurred simultaneously with feedback: 90 to give feedback and 205 to elicit feedback. Finally, of the 96 body movements, 68 (71%) were feedback related: 38 gave feedback and 30 elicited feedback. The other movements were related to turn management or other unspecified functions. Annotation features and their statistics are given in Table 1.

Table 1. Frequency count of head, hand, and body annotation features

Head movements	**Count**	**Hand movements**	**Count**
Backward	38	**Handedness**	
Forward	77	Both hands	265
Nod (up and down)	345	One hand	137
Tilt	95	**Hand movement repetition**	
TurnSide	76	Repeated	174
Waggle	11	Single	220
Other	3	Other	8
Total (head)	**645**	**Hand movement interpretation**	
Body movement		Deixis	8
Backwards	15	Emphasis	19
Forward	37	Rhythm	185
LeaningAwayFromPartner	41	Standup	2
Other	3	Not classified	188
Total (body)	**96**	**Total (hand)**	**402**

Annotation was done by two annotators independently and checked by an expert annotator. Intercoder agreement between the two annotators on annotation categories is shown in Table 2. Kappa (κ) is calculated using Anvil's coding agreement facility which automatically compares two annotation files frame by frame (frame = 0.4s) and calculates intercoder agreement with respect to the joint segmentation and categories, and the overall agreement. The annotators showed very good agreement on body annotation, good or fair agreement on face features, and almost perfect agreement on hand category agreement. The surprisingly low agreement on hand segmentation is obviously due to the annotators' difference in deciding on the start and end points of complex hand gestures, and whether these are classified as one or many gestures.

Table 2. Kappa and %-agreement of some of the annotation categories

Track	Segmentation		Category		Overall	
	κ	%	κ	%	κ	%
Face	0,48	74	0,71	79	0,65	72
Hand	0,09	41	0,95	98	0,11	41
Body	0,86	93	0,79	84	0,91	93

4 Results

Simple correlations were calculated on the basis of the data, to find out how hand, head, and body movement correlate with the feedback function. Detailed analyses are discussed in [13] while here we focus on the combined effects of the different modalities on feedback. These are shown in Figure 3. The first four columns on the left concern body movement, the six last ones on the right concern head movement, and the 10 columns in the middle describe hand movements. The columns are normalised with respect to the frequency counts in Table 2.

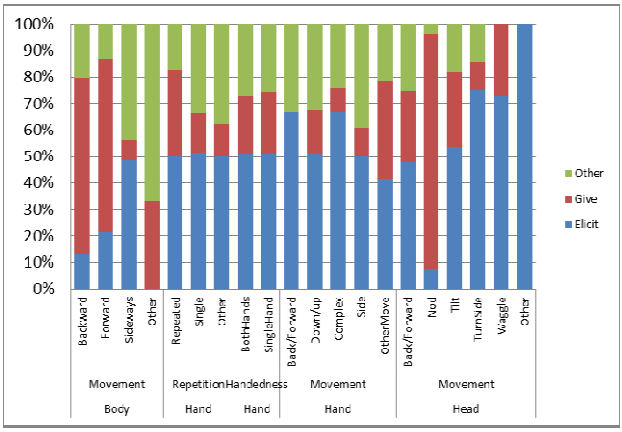

Fig. 3. Give and Elicit feedback in different modalities. Other refers to movements not related to feedback.

We can see that back/forward body movements are mainly used to give feedback and sideway movements to elicit feedback. Sidway movements rarely function to give feedback, but they can also occur in other functions not related to feedback at all. However, other, unspecified body movement types can also function to give feedback. Hand is mainly used to elicit feedback, i.e. it is used to engage the partner into the conversation. In particular, back/forward gestures are related to feedback elicitation. Concerning repetition of hand gestures, almost 40 % of the repeated gestures but only about 20% of single hand gestures give feedback, i.e. twice as many repeated gestures function as feedback giving signals as single gestures. As for handedness, independently from whether the gestures are produced by single hand or both hands, a similar pattern appears: feedback elicitation is twice as common as feedback

giving. Also, one third of all the gestures are used for other functions than providing feedback.

Head movements are also used to provide feedback, and in general they show a similar relative distribution between feedback giving and eliciting functions as hand movements: about third of the movements function as feedback giving. Back/forward movement, tilting, side turns, and waggle are more common in eliciting feedback, but nodding, however, is almost exclusively used to give feedback. Nodding can also be divided into up-nods and down-nods depending on which way the movement starts. Although this may not always be clear, in [23] it was noticed that the difference is related to the interpretation of nodding: up-nods are used when the speaker has presented information that is new to the listener in the given context, while down-nods signal neutral acknowledgement of the information that is expected in the context.

Most commonly feedback is related to spoken feedback particles and backchannelling vocalisations, and it is interesting to see correlations between verbal and non-verbal communication. The four most common verbal feedback signals in the Finnish corpus are *joo (yeah), nii(n) (yes but), okei (okay), aivan (exactly)*. The difference between *joo* and *nii* is related to the novelty value of the presented information: *joo* indicates that the speaker acknowledges the presented information, while *nii(n)* indicates that the speaker has some reservations about it, or can add more to it. *Okei* and *aivan* indicate the speaker's stronger agreement or commitment to the presented information. Figure 3 depicts co-occurrence frequency counts of the common verbal feedback and head movements, normalised with respect to time (in our case: one minute). We notice that nodding is common with both the neutral acknowledgement (*joo*) and the acknowledgement with reservations (*nii*), and in fact, the difference appears to be related to the direction of nodding as discussed in [23]: the neutral *joo* most often co-occurs with down-nods, while *nii* usually co-occurs with up-nods.

Back- and forward head movement is about twice as common with *nii* than with *joo* or *okei*, and also tilting of the head co-occurs more than twice as often with *nii* as with *joo* or *aivan*. We can thus assume that tilting is related to a surprise or novelty value of the presented information, and that the listener has some reservations about it that prevent the uptake of the information as such. Side turns, however, generally signal change of attention away from the partner, and this may explain why they are not common with any of the four feedback particles. They are extremely rare with *aivan* (exactly), which is understandable as the speaker would verbally express strong agreement with the presented information but non-verbally display contradiction or non-interest with it. Verbal and non-verbal feedback activity tend to be produced, and also get interpreted, as semantically aligned in smooth communication

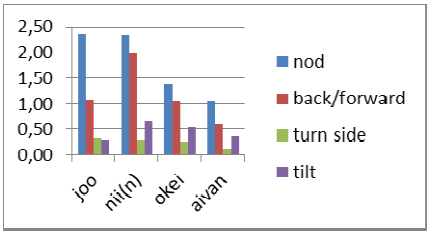

Fig. 4. Frequent co-occurrences of verbal feedback and head movement signals

5 Conclusion and Future Prospects

This paper has focussed on the head, hand, and body movements and their use in conversational interactions as means of visual interaction management. It is clear that hand gestures, head movement and body posture are important multimodal means for interaction management and for providing feedback of the current state of the interaction. We presented some observations of their co-occurrence and correlations in feedback giving and eliciting situations, and also provided interesting co-occurrence data of verbal and non-verbal feedback expressions.

These observations will be used for further multimodal and multicultural studies. We will investigate visual interaction management as part of human-computer interaction, and focus on the role of multimodal signals in the control and coordination of interaction. Experiments will concern the participant's engagement in conversational interactions, and we will use various features, especially multimodal, besides verbal features, to measure their conversational activity. We will also build models of the multimodal strategies that the interlocutors have at their disposal to construct shared understanding and to advance their goals, to be applied to interactive applications.

Since the NOMCO first encounter corpus is analogous to the other similar corpora collected at the other Nordic countries, it is possible to compare interaction strategies in different (cultural) contexts, especially in cultures and among languages which are closely related. Such comparison will allow us to deepen our understanding of the various multimodal signals and their impact and function in interactions as well as in intercultural communication.

References

1. Allwood, J.: An Activity Based Approach to Pragmatics. In: Gothenburg Papers In Theoretical Linguistics 76. Department of Linguistics. Göteborg University (2000)
2. Allwood, J., Cerrato, L., Jokinen, K., Navarretta, C., Paggio, P.: The MUMIN Coding Scheme for the Annotation of Feedback, Turn Management and Sequencing Phenomena. Multimodal Corpora for Modelling Human Multimodal Behaviour. Special Issue of the International Journal of Language Resources and Evaluation 41(3-4), 273–287 (2007)
3. André, E., Pelachaud, C.: Interacting with Embodied Conversational Agents. In: Jokinen, K., Cheng, F. (eds.) Speech-based Interactive Systems: Theory and Applications. Springer (2009)
4. Argyle, M., Cook, M.: Gaze and Mutual Gaze. Cambridge University Press (1976)
5. Battersby, S.: Moving Together: the organization of Non-verbal cues during multiparty conversation. PhD Thesis, Queen Mary, University of London (2011)
6. Clark, H.H., Schaefer, E.F.: Contributing to Discourse. Cognitive Science 13, 259–294 (1989)
7. Feldman, R.S., Rim, B.: Fundamentals of Nonverbal Behavior. Cambridge University Press (1991)
8. Hart, C.L., Fillmore, D.G., Griffith, J.D.: Indirect Detection of Deception: Looking for Change. Current Research in Social Psychology 1(9), 134–142 (2009)

9. Jokinen, K.: Rational communication and affordable natural language interaction for ambient environments. In: Lee, G.G., Mariani, J., Minker, W., Nakamura, S. (eds.) IWSDS 2010. LNCS, vol. 6392, pp. 163–168. Springer, Heidelberg (2010)
10. Jokinen, K.: Pointing gestures and synchronous communication management. In: Esposito, A., Campbell, N., Vogel, C., Hussain, A., Nijholt, A. (eds.) COST 2102 Int. Training School 2009. LNCS, vol. 5967, pp. 33–49. Springer, Heidelberg (2010)
11. Jokinen, K., Furukawa, H., Nishida, M., Yamamoto, S.: Gaze and Turn-taking behaviour in Casual Conversational Interactions. ACM Trans. Interactive Intelligent Systems (2013)
12. Jokinen, K., Pärkson, S.: Synchrony and copying in conversational interactions. In: Proceedings of the 3rd Nordic Symposium on Multimodal Communication, vol. 15, pp. 18–24 (2011)
13. Jokinen, K., Wilcock, G.: Multimodal Signals and Holistic Interaction Structuring. In: Proceedings of the 24th International Conference on Computational Linguistics (COLING 2012), Mumbai, India (2012)
14. Kendon, A.: Gesture. Visible Action as Utterance. Cambridge University Press (2005)
15. Kipp, M.: Anvil - A Generic Annotation Tool for Multimodal Dialogue. In: Proceedings of the 7th European Conference on Speech Communication and Technology (Eurospeech), pp. 1367–1370 (2001)
16. Lee, J., Marsella, S.C., Traum, D.R., Gratch, J., Lance, B.: The Rickel Gaze Model: A Window on the Mind of a Virtual Human. In: Pelachaud, C., Martin, J.-C., André, E., Chollet, G., Karpouzis, K., Pelé, D. (eds.) IVA 2007. LNCS (LNAI), vol. 4722, pp. 296–303. Springer, Heidelberg (2007)
17. Levitan, R., Gravano, A., Hirschberg, J.: Entrainment in Speech Preceding Back-channels. In: Proceedings of ACL 2011, pp. 113–117 (2011)
18. Mancini, M., Castellano, G., Bevacqua, E., Peters, C.: Copying Behaviour of Expressive Motion. In: Gagalowicz, A., Philips, W. (eds.) MIRAGE 2007. LNCS, vol. 4418, pp. 180–191. Springer, Heidelberg (2007)
19. Nakano, Y., Nishida, T.: Attentional Behaviours as Nonverbal Communicative Signals in Situated Interactions with Conversational Agents. In: Nishida, T. (ed.) Engineering Approaches to Conversational Informatics. John Wiley (2007)
20. Navarretta, C., Ahlsén, E., Allwood, J., Jokinen, K., Paggio, P.: Feedback in Nordic First-Encounters: a Comparative Study. In: Proceedings of the Language Resources and Evaluation Conference (LREC-2012), Istanbul, Turkey (2012)
21. Norman, D.A.: The Psychology of Everyday Things. Basic Books, New York (1988)
22. Pickering, M., Garrod, S.: Towards a mechanistic psychology of dialogue. Behavioral and Brain Sciences 27, 169–226 (2004)
23. Toivio, E., Jokinen, K.: Multimodal Feedback Signaling in Finnish. In: Proceedings of the Human Language Technologies – The Baltic Perspective (2012); Published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License
24. Tomasello, M.: First verbs: A case study of early grammatical development. Cambridge University Press, Cambridge (1992)