Listener head gestures and verbal feedback expressions in a distraction task

Marcin Włodarczak[1], *Hendrik Buschmeier*[2], *Zofia Malisz*[1]
Stefan Kopp[2], *Petra Wagner*[1]

[1]Faculty of Linguistics and Literary Studies
[2]Sociable Agents Group, CITEC and Faculty of Technology
Bielefeld University, Bielefeld, Germany

{zofia.malisz,petra.wagner,mwlodarczak}@uni-bielefeld.de,
{hbuschme,skopp}@techfak.uni-bielefeld.de

Abstract

We report on the functional and timing relations between head movements and the overlapping verbal-vocal feedback expressions. We investigate the effect of a distraction task on head gesture behaviour and the co-occurring verbal feedback. The results show that head movements overlapping with verbal expressions in a distraction task differ in terms of several features from a default, non-perturbed conversational situations, e.g.: frequency and type of movement and verbal to nonverbal display ratios.

Index Terms: communicative feedback; head gestures; dialogue; attentiveness; distraction task

1. Introduction

Head gestures "are both an integral part of language expression and function to regulate interaction" [1]. Often the resulting structure involves interactional synchrony where head movements between speakers are aligned in a rhythmic or quasi-rhythmic way [2]. Such temporal coordination of communicative actions on many levels and in many modalities facilitates turn-taking [3] and enhances communicative attention [4, 5]. Also, as [6] notes, feedback is an essential part of the grounding process where common ground is shared and achieved as a result of joint conversational activity. Head gestures are involved in updating the information status (grounding) and in establishing rapport.

In order to describe the form of a head gesture, a couple of features need to be taken into account: head orientation, speed and amplitude of movement [7]. Several different inventories of gesture forms were devised in the past by *inter alia* [8, 9]. Research on general head gesture kinematics was pioneered by [10] and [11]. [11] distinguished between linear and cyclic kinematic forms, equivalent to e.g. single and multiple nodding bouts and associated them with turn taking signals and responses to questions respectively. Moreover, phrasing and prominence information can be carried by head nodding along with other visual modalities [12, 13]. [11] noted that floor grabbing cues are usually expressed by wide and linear head movements (e.g. high amplitude single nods) while synchronisation with pitch accented syllables in the interlocutors' speech occurred in case of narrow, linear head gestures, e.g. low amplitude single nods. More importantly, the tendency of "yes" and "no" movements to be cyclic (multiple nods) was uniform and robust across speakers. In [9] feedback categories defined as "recognition-success" and "contents-affirmation" corresponding to backchannels and other affirmative responses respectively, were found to occur with "vertical head movements" of both large and small amplitude.

Claims were made by some of the above authors as to how the physical properties of head gestures relate to their communicative use. The function of a head gesture can be independent within the nonverbal modality or co-expressive with the accompanying linguistic content. [14] enumerates the criteria that are necessary to disentangle the meaning of nods. Additionally, she makes a distinction between how a meaning of a nod can be modified by the co-occurring linguistic context (such as preceding or overlapping feedback expressions) and/or simultaneous multimodal context (co-occurring facial displays, gaze behaviour or hand gestures [15]). In [16] it was shown that head nods of a listening agent were interpreted as "agree" and "understand" by participants; however, when combined with a smile they were interpreted as "like" and "accept". This and similar examples show that the exact level of evaluation and grounding can be modified by several modalities at once and that head gestures need to be interpreted in their multimodal context.

In our study we concentrate on the functional and timing relation between head movements and the overlapping spoken feedback expressions leaving the remaining co-occurring multimodal context, certainly able to modify the resulting function, to later study. Additionally, we investigate the effect of a distraction task on head gesture behaviour and the co-occurring verbal feedback. We also briefly look at the timing relations within sequences of nods.

2. Study design

In order to analyse feedback behaviour, we carried out a face-to-face dialogue study in which one of the dialogue partners (the 'storyteller') told two holiday stories to the other participant (the 'listener'), who was instructed to listen actively, make remarks and ask questions. Furthermore, similar to [17], the listeners were distracted during one of the stories by an ancillary task. They were instructed to press a button on a hidden remote control every time their dialogue partner uttered a word starting with the letter 's' (the second most common German word-initial letter). Participants also had to count the total number of 's-words' they heard. Storytellers told two different holiday stories and listeners only engaged in the distraction task for either the first (in even-numbered sessions) or the second story (in odd-numbered sessions). Participants were seated approximately three metres apart to minimise crosstalk. Interactions were recorded from three camera perspectives: medium shots showing the storyteller and the listener and a long shot showing the whole scene.

[*] The first three authors contributed to the paper equally.

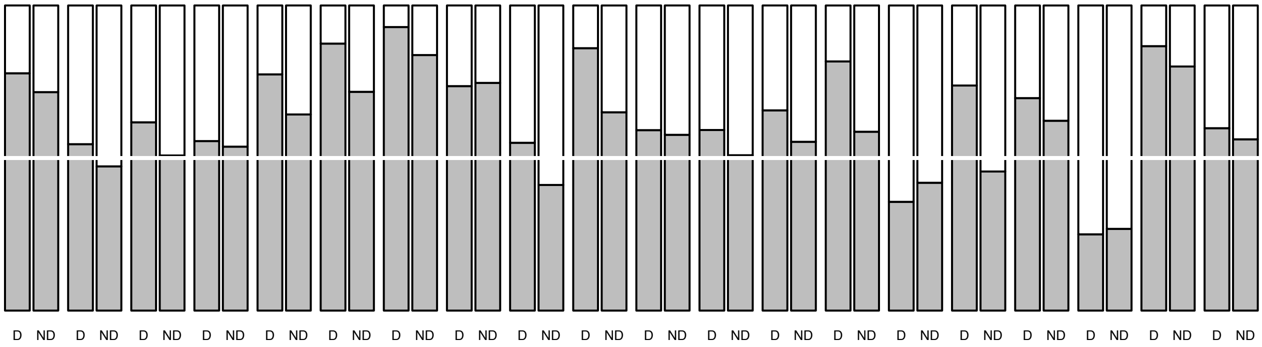

Figure 1: The ratio between head gesture units (grey bars) and verbal feedback expressions (white bars) across 20 dialogue sessions and two experimental conditions (D: distracted; ND: non-distracted)

3. Multimodal annotation

3.1. Verbal feedback

Feedback utterances and head gesture units were segmented and transcribed for 20 sessions in the corpus. A feedback function annotation scheme was devised (see [18] for full description) in which feedback levels largely correspond to definitions by [19]. Our category P1 corresponds to backchannels understood as 'continuers', category P2 signals successful interpretation (understanding) of the message, and category P3 indicates acceptance, belief and agreement. These levels can be treated as a hierarchy with increasing value of judgement, "cognitive involvement" or "depth" of grounding. Feedback expressions were labeled according to German orthographic conventions. Feedback functions were annotated independently by three annotators taking communicative context into account. Majority labels between annotators were then calculated automatically and problematic cases (185; ca. 9%) were discussed and resolved.

3.2. Head gestures

Head gesture annotation was based on *head gesture units* (HGUs). We defined an HGU as a perceptually coherent and continuous movement sequence. Any perceived pauses either before a rest (no movement) or between units were marked as unit boundaries. The exact onset and offset of an HGU was determined by close inspection of the video in ELAN. Each HGU was annotated for movement types (*nod*, *jerk*, *tilt*, *turn*, *protrusion* and *retraction*) and the number of movement cycles. The movement type inventory was arrived at incrementally while inspecting the dataset. In case of nods, one "down-up" movement was counted as one cycle. In comparison, for jerks, one "up-down" movement was counted as one cycle.

The following features were extracted for each gestural phrase: *duration*, *complexity* (the number of subsequent gesture types in the phrase), *cycles* (the total number of cycles of all gestures in the phrase) and *frequency* (the number of cycles divided by the duration of the unit). For example, the label "Nod-2+Tilt-1-Right+Pro-1" has the complexity degree of 3 (nod, tilt-right, protrusion) and its total number of cycles equals 4 (2 nod cycles + 1 tilt-right cycle + 1 protrusion cycle).

Additionally, for phrases overlapping with short verbal feedback expressions, the exact function (P1, P2, P3) of the expression, the *overlap onset* (the time between the beginning of the gestural phrase and the feedback expression), and movement types of the head gesture were recorded.

4. Results and discussion

4.1. Verbal and nonverbal feedback

The proportion of all HGUs compared to verbal feedback will be examined first. To perform the analysis we excluded head movements coinciding with longer utterances not marked as feedback. It is not possible to determine how long a gap between multimodal expressions can be in order to be perceived as a functional unit without conducting a separate study on timing relations. Therefore, *barely* non-overlapping verbal and HGU relations were included in the 'non-overlapping' category. Figure 1 presents the total number of HGUs (both overlapping and non-overlapping) related to the total number of verbal feedback expressions. The results are presented as a ratio between the two variables and are split into the two experimental conditions (distracted vs. non-distracted) within single dyad sessions.

4.2. Verbal and nonverbal feedback per condition

Overall, there is more nonverbal than verbal feedback in both conditions. Consequently, listeners use the nonverbal channel to signal feedback more often. It has been noted that head movement is present almost incessantly in human interactive communication. Moderately involved, polite listener behaviour, however, can be hypothesised to feature less speech and manual gesture but lots of eye contact and head movement. In our setting, a comparison between the "default" non-distracted condition and the distracted condition provides a platform for studying levels of involvement and attention in the feedback giving context in listeners. Indeed, Figure 1 suggests a tendency for 17 out of 20 listeners to produce more more HGUs when distracted experimentally. Overall, 65% nonverbal to 35% verbal signals in the distracted and 57% nonverbal to 43% verbal signals in the non-distracted condition was observed ($\chi^2 = 22.3$, $p < 0.001$). As shown in [18] less verbal feedback was displayed by distracted than non-distracted listeners. This was corroborated in the present study, where six more sessions from the same corpus were added to the analysed dataset.

No significant differences between conditions in the proportion of time spent gesturing with the head and in the number of HGUs were found. Overall, subjects spent 17.5% of time gesturing (19% in the distracted condition, 16% in the non-distracted). Similarly, no evidence of a distraction effect on absolute HGU counts (also normalised by the dialogue duration) was found. The effect seems to be only evident in the interaction between verbal and nonverbal channels.

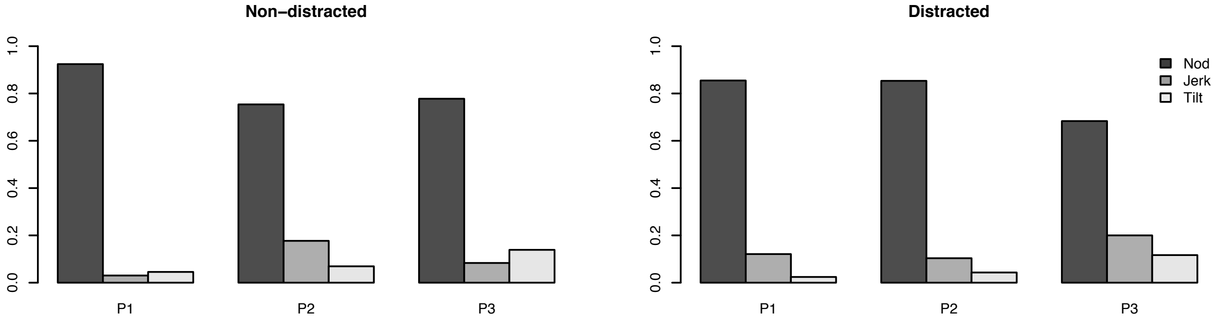

Figure 2: Conditional probabilities of three head gesture types given the function of the overlapping verbal expression (P1: backchannel, P2: understanding, P3: agreement/acceptance) in the two experimental conditions.

4.3. Gesture types across dialogue act categories

We assumed that HGUs overlapping with verbal feedback expressions share the same feedback function. Consequently, gesture units overlapping with more than one feedback expression ("multiple overlaps") were excluded because the functional relation between head movements and verbal feedback could not be determined for these cases. Figure 2 presents the conditional probability of the most frequent gesture types given the function of the overlapping verbal expression (P1, P2, P3). The left panel corresponds to the non-distracted condition.

Nods predominate both when overlapping with feedback expressions and on their own (81.5% of non-overlapping cases). The probability of nods decreases when one moves up the feedback function hierarchy, while other head movement types are more likely to occur. Specifically, the probability of the tilt correlates positively with the feedback function. For example, tilts are twice as probable in the "acceptance/agreement" function (P3) than in the "understanding" function (P2) and three times as probable as in backchanneling (P1). We also observe the probability of the jerk occurring in P2 is four times as high than in P1 and more than two times higher than in P3. Jerks are characteristic as displays of understanding and surprise, especially with the meaning of "I have finally understood", e.g.: after check questions [20].

German speakers tend to produce more repeated nods across feedback categories (P1 = 74.5%, P2 = 81%, P3 = 86%). Results in [6] (for Swedish speakers) are therefore corroborated on a larger dataset. Insofar as our feedback function inventory corresponds to the category of "yes and no responses" in [11], our result is also in agreement with their conclusions that those involve cyclic (multiple cycle) movements. However, in P1, when compared to higher feedback functions (P2 and P3), the backchannel function (comparable to "ContinuationYouGoOn" used by [6]) exhibits a lower percentage of multiple nods.

For HGUs composed of head gestures other than head nods only, complexity and frequency tendentially falls with feedback function level, a significant difference was found between P1 and P3 (Mann-Whitney test, $p < 0.05$ and $p < 0.01$ respectively) for this data subset.

4.4. Gesture types per dialogue act category and condition

For HGUs overlapping with verbal feedback in the distracted condition, the probability of a nod co-occurring with an expression bearing the P2 function is closer to the probability for overlaps with the backchannel function (see Figure 2, right panel). In the default, non-distracted conversational situation on the other hand, the probability of nods comes closer to the probability for overlaps with the "acceptance/agreement" function (see Figure 2, left panel). Also, while bearing in mind the low number of annotated jerks overlapping with the "understanding" function (35 instances), we observe a tendency to decrease the use of jerks in the distracted condition when expressing understanding verbally. It is possible that the two phenomena are related: the characteristic "I understand" nonverbal expression, the jerk, is replaced with the nod, a more minimal, "default" response.

81.5% non-overlapping head gestures in the whole dataset ($N = 1328$) were nods ($N = 1083$). In the non-overlapping category we find no significant differences between distracted and non-distracted listeners in the number of cycles, the proportion of multiple vs. single nods and the duration of the nods.

Significant difference between the number of movement cycles and frequency in the distracted and non-distracted condition was observed for P2 ($p < 0.05$ and $p < 0.01$ respectively). The result indicates that more intense movement in the time domain is characteristic of distracted listeners expressing understanding.

4.5. Movement timing

We analysed overlaps between HGUs and single feedback expressions. The overlap onset is negative i.e. the HGU onset precedes the feedback expression onset. [21] found that nods in listeners preceded the corresponding speech by 175 ms. Most HGU onsets in our data were close in time to the overlapping feedback expression onset (median, non-distracted = 202 ms, SD = 380 ms) with a clear tendency for the HGU onset to precede the verbal expression.

Additionally, a regression analysis was performed on head nod durations in order to determine whether HGUs with multiple nod cycles show a linear trend with cycle increase. It turned out that from more than one nod cycle the duration of the HGU increases by 320 ms with each consecutive nod (adjusted $R^2 = 0.6$). A non-zero intercept indicates that as new nod cycles are added, the duration of the HGU increases non-cumulatively.

The trend can be explained by the dynamics of head motion that is continuously oscillating within a multiple nod phrase: adding nod cycles within a uninterrupted phrase takes less time than separate single nods. Definitely, the nature of the oscillatory process facilitates integration of kinetic energy so that speakers can use the momentum produced by a previous nod to produce the next one within (at least this might hold with the contrast between single nods and multiple ones).

Consequently, a significant nonlinear trend was evidenced when single nods were added to the regression analysis. We know from the nature of this biological system that there might be some visible damping towards the end of a nodding bout with multiple cycles. [10, 11] showed that movement amplitude decreases as its frequency increases as well as that the variability in amplitude within HGUs is high so the damping is not monotonous. Single head nods were described as linear and multiple head nods as cyclical in their study, which corresponds to the difference in the regression trend.

5. Conclusions and future work

Our results showed a significant difference between conditions, where the ratio of nonverbal to verbal feedback is higher in the distracted condition. In HGUs overlapping with verbal feedback expression, nods, especially multiple ones, predominated. Additionally, our results suggest that the tilt is more characteristic of higher feedback categories and that the jerk expresses understanding. The variation found here in the use of the jerk between experimental conditions is in accordance with our earlier result [18] that communicating ‘understanding’ (as in P2) is a marker of attentiveness.

The visual modality, as mentioned earlier, can influence and modify the interpretation of the feedback function. Perceptual evaluation of feedback functions including additional visual modalities needs to be conducted in the future in order to shed more light on the complex interaction of the verbal and nonverbal cues to feedback functions. Movement timing information will be used for the study of interactional synchrony that embodies attention processes and grounding.

Acknowledgements – This research is supported by the Deutsche Forschungsgemeinschaft (DFG) at the Center of Excellence in ‘Cognitive Interaction Technology’ (CITEC) as well as at the Collaborative Research Center 673 ‘Alignment in Communication’. We would also like to thank Joanna Skubisz for help with annotation.

6. References

[1] E. McClave, “Linguistic functions of head movements in the context of speech,” *Journal of Pragmatics*, vol. 32, pp. 855–878, 2000.

[2] F. J. Bernieri and R. Rosenthal, *Interpersonal Coordination: Behavior Matching and Interactional Synchrony*. Cambridge, UK: Cambridge University Press, 1991.

[3] M. Wilson and T. P. Wilson, “An oscillator model of the timing of turn taking,” *Psychonomic Bulletin and Review*, vol. 12, pp. 957–968, 2005.

[4] W. S. Condon and W. D. Ogston, “Speech and body motion synchrony of the speaker-hearer,” in *Perception of language*, D. L. Horton and J. J. Jenkins, Eds. Columbus, Ohio: Merrill, 1971.

[5] A. Kendon, “Movement coordination in social interaction: Some examples described,” *Acta Psychologica*, vol. 32, pp. 100–125, 1970.

[6] L. Cerrato, “Investigating communicative feedback phenomena across languages and modalities,” Ph.D. dissertation, KTH Computer Science and Communication, Department of Speech, Music and Hearing, Stockholm, Sweden, 2007.

[7] D. Heylen, “Challenges ahead: Head movements and other social acts in conversations,” in *Proceedings of AISB 2005*, 2005, pp. 45–52.

[8] R. L. Birdwhistell, *Kinesics and Context. Essays on Body Motion Communication*. Philadelphia, PA: University of Pennsylvania Press, 1970.

[9] Y. Iwano, S. Kageyama, E. Morikawa, S. Nakazato, and K. Shirai, “Analysis of head movements and its role in spoken dialogue,” in *Proceedings of ICSLP’96*, 1996, pp. 2167–2170.

[10] U. Hadar, T. J. Steiner, E. C. Grant, and F. C. Rose, “Kinematics of head movements accompanying speech during conversation,” *Human Movement Science*, vol. 2, pp. 35 – 46, 1983.

[11] U. Hadar, T. Steiner, and C. F. Rose, “Head movement during listening turns in conversation,” *Journal of Nonverbal Behavior*, vol. 9, pp. 214–228, 1985.

[12] D. House, J. Beskow, and B. Granström, “Interaction of visual cues for prominence,” *Lund Working Papers in Linguistics*, vol. 49, pp. 62–65, 2001.

[13] M. Sargin, O. Aran, A. Karpov, F. Ofli, Y. Yasinnik, S. Wilson, E. Erzin, Y. Yemez, and M. A. Tekalp, “Combined gesture-speech analysis and speech driven gesture synthesis,” in *Proceeings of the IEEE International Conference on Multimedia and Expo*, Toronto, Canada, 2006, pp. 893–896.

[14] I. Poggi, F. D’Errico, and L. Vincze, “Types of nods. the polysemy of a social signal,” in *Proceedings of the 7th International Conference on Language Resources and Evaluation*, 2010, pp. 17–23.

[15] H. M. Rosenfeld and M. Hancks, “The nonverbal context of verbal listener responses,” in *The Relationship of Verbal and Nonverbal Communication*, M. R. Key, Ed. The Hague, The Netherlands: Mouton Publishers, 1980, pp. 193–206.

[16] E. Bevacqua, “Computational model of listener behavior for embodied conversational agents,” Ph.D. dissertation, Université Paris 8, Paris, France, 2009.

[17] J. B. Bavelas, L. Coates, and T. Johnson, “Listeners as conarrators,” *Journal of Personality and Social Psychology*, vol. 79, pp. 941–952, 2000.

[18] H. Buschmeier, Z. Malisz, M. Włodarczak, S. Kopp, and P. Wagner, “‘Are you sure you’re paying attention?’ – ‘Uh-huh’. Communicating understanding as a marker of attentiveness,” in *Proceedings of INTERSPEECH 2011*, Florence, Italy, 2011, pp. 2057–2060.

[19] S. Kopp, J. Allwood, K. Grammar, E. Ahlsén, and T. Stocksmeier, “Modeling embodied feedback with virtual humans,” in *Modeling Communication with Robots and Virtual Humans*, I. Wachsmuth and G. Knoblich, Eds. Berlin: Springer-Verlag, 2008, pp. 18–37.

[20] J. Allwood and L. Cerrato, “A study of gestural feedback expressions,” in *First Nordic Symposium on Multimodal Communication*, Copenhagen, Denmark, 2003, pp. 7–22.

[21] A. Dittmann and L. Llewellyn, “Relationship between vocalizations and head nods as listener responses.” *Journal of personality and social psychology*, vol. 9, p. 79, 1968.