MULTIMODAL SIGNALLING: THE INTERPLAY OF ORAL AND VISUAL FEEDBACK IN CONVERSATION

Malin Spaniol[1], Alicia Janz[2], Simon Wehrle[2], Kai Vogeley[1], Martine Grice[2]

[1]Dept. of Psychiatry, University Hospital Cologne, Germany; [2]I*f*L Phonetik, University of Cologne, Germany
{malin.spaniol, kai.vogeley}@uk-koeln.de; {alicia.janz, simon.wehrle, martine.grice}@uni-koeln.de

ABSTRACT

Human communication comprises a complex and dynamic interplay of verbal and nonverbal communication channels. It is an intrinsically multimodal, interactive, time-sensitive and highly coordinated process. However, the multimodal nature of face-to-face interaction needs further study. Here, we present a novel approach to studying dyadic face-to-face conversation with dual mobile eye-tracking glasses. Our pilot data are from four conversations between German speaking interlocutors in different contexts, comprising introductory small talk, task-oriented dialogue (Tangram task), and free discussion (about the Tangram task). Our exploratory analysis on oral feedback and gaze behaviour reveals that tokens indicating Passive Recipiency (backchannels) tended to involve more gaze directed at the conversational partner than tokens indicating Incipient Speakership (signals with the intention to take the floor). It also offers interesting insights into behavioural differences as a function of communicative context.

Keywords: face-to-face interaction, feedback, social gaze, backchannels, turn-taking

1. INTRODUCTION

In any face-to-face interaction, interlocutors continuously exchange a variety of signals, both verbal and non-verbal, the majority of which are of social relevance [1]. This continuous two-way exchange is possible because interlocutors are able to see, hear and react to each other. The timing of visual and verbal signals is of crucial importance [2, 3].

Language coordination has been claimed to be an inherently redundant process [4]. It is known that e.g. turn-taking cues, relevant for a high degree of coordination, can be transmitted on multiple communication channels.

Verbally, feedback signals such as backchannels function as turn-taking signals and are used to demonstrate understanding and acknowledgement. More specifically, they can indicate either Passive Recipiency (PR), which supports the ongoing turn of the interlocutor, or Incipient Speakership (IS), which signals the intention of taking over the floor from the current speaker [5, 6]. Previous research has found complex but consistent evidence for the relationship between a feedback token's lexical form, its intonation contour and its function as either turn-taking (IS) or turn-yielding (PR) [7]. Although phonetic cues have been shown to generally play a role in signalling the function of feedback signals, there is no one-to-one mapping of phonetic form to turn-taking function [7, 8].

In addition to spoken cues, nonverbal signals such as gaze have been shown to substantially influence the coordination of turns [9, 10]. Gaze serves both a sensing and a signalling function [11]. Early research on gaze-direction in social interaction has shown that even small differences in study design, regarding e.g. task or setting, can have considerable effects on gaze behaviour [12, 13] and substantially influence the coordination of turns [10].

More recent studies on gaze-direction in social interaction further emphasize the importance of gaze in turn-taking [14] and have shown that speakers generally tend to direct their gaze at their interlocutor towards the end of a turn, while more often averting it at the beginning [13, 15] in order to plan the rest of the current turn [16]. Direct gaze by the speaker, which entails mutual gaze, can create a so-called "gaze window" [17], proposed to function as a backchannel-inviting cue [23]. Furthermore, speakers likely direct their gaze to the interlocutor before the end of a turn in order to signal turn-yielding and facilitate a possible turn-transition [10, 15, 16, 18].

Although there is substantial inter-individual variability, previous results also indicate a general tendency for participants to gaze less at their conversation partner when they are in the speaker role as compared to the listener role [13].

Despite the fact that both verbal feedback and gaze behaviour are important cues for the coordination of speaking turns, only few studies have investigated the relation between these two signals in the context of turn-taking to date. Skantze et al. [11] have shown that direct gaze by the speaker (entailing mutual gaze) serves as a backchannel-inviting cue. However, the function of gaze behaviour by the listener uttering feedback remains unclear. More specifically, the relationship between Passive Recipiency vs. Incipient Speakership on the one hand, and aversion vs. direction of gaze on the other, remains unclear.

Although it is apparent that the complex interplay of different conversation modalities and channels is still not fully understood, a recent renewed interest in the dyad as a fundamental unit of social behaviour, aided by recent technological and methodological advancements that enable fine-grained multimodal analysis, can help in furthering our understanding.

In an attempt to contribute to this progress, we have developed a novel method for analysing multimodal communication. In a highly naturalistic setting, we record gaze using mobile dual eye-tracking glasses and speech via lapel microphones while capturing the scene with an external camera. This setup allows for the simultaneous measurement of multiple verbal and non-verbal signals. The glasses are equipped with integrated cameras and are capable of automatic gaze detection. Additional face-detection on the videos recorded by the eye-tracking glasses enable us to automatically detect gaze into the facial region of the interlocutor.

The current study is based on pilot data recorded with this novel setup. The exploratory analysis presented here is focussed on verbal feedback signals, gaze behaviour, and the interplay between the two.

Because verbal feedback and directed gaze from the listener both function as feedback signals for the speaker, and as signals are expected to be redundant, we expect to see directed gaze towards the speaker during backchanneling. However, one hypothesis for averted gaze at the beginning of turns is that it helps to keep cognitive load low while planning the turn ahead. Verbal feedback with the function of Incipient Speakership always occurs at the beginning of a speech turn. Thus, it can be expected that direct gaze during Incipient Speakership tokens will be reduced compared to Passive Recipiency backchannels (where we expect to find gaze directed towards the speaker). The previous literature has shown that all of the relevant behaviours are task-dependent, and therefore, we also expect to find differences between conversational contexts.

We investigated three different communicative contexts, and the following questions guided our exploration: Do listeners show different gaze patterns when producing Incipient Speakership signals, as compared to Passive Recipiency backchannels? If so, do these patterns differ according to communicative context?

2. METHOD

2.1. Apparatus

Two participants were seated opposite each other wearing Pupil Invisible mobile eye-tracking glasses [19] and lapel microphones. The glasses were connected to remote devices which were wirelessly connected to the experimenter's computer. The eye-trackers were manually adjusted for each participant using gaze offset correction. The lapel microphones were attached to the eye-tracking glasses. Microphones and a scene camera were connected to the experimenter's laptop via cable. Figure 1 shows an overview of the experimental setup. The experimenter was in a separate room for the entire duration of the experiment.

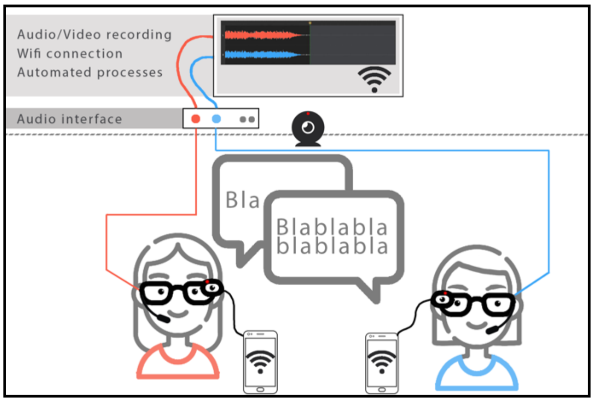

Figure 1: Multimodal dyadic conversation setup.

2.2. Participants

Four dyads matched for age and gender participated in this study (8 participants: 6 female, 2 male; age M = 29.63, SD = 5.21, range = 23-38). All were native German speakers. The study was approved by the ethics committee of the University Hospital Cologne.

2.3. Procedure

After giving their written informed consent, participants were introduced to each other before being seated and equipped with eye-tracking glasses and microphones. Conversations lasted between 26 and 48 minutes and encompassed three communicative contexts, in a fixed order: 1. Introduction (and getting to know each other); 2. Tangram task; 3. Discussion (of the Tangram task).

2.4. Task and recordings

Two assistants familiarised the participants with the recording set-up before giving them instructions for the first context and the opportunity to ask any questions. The assistants then left the room and the conversation started. After 10 minutes, an assistant came back into the room and provided instructions for the Tangram task. One participant saw four different figures, of which one was highlighted by an arrow. The other participant saw only one figure. The task for the dyad was to find out if the single figure was identical to the highlighted one; an example can be found on the *OSF* project page. The Tangram conversations comprised 10 iterations and had no time limit. After the task was completed, an assistant entered the room again to give instructions for the final, ten-minute long discussion.

The three conversational contexts where chosen for being complementary, but not unrelated. Getting to know each other was used as an example of a relatively frequent everyday situation. The Tangram task was chosen, in contrast, as a form of highly structured, task-oriented dialogue, and also because it has been used in other studies on speech and gaze in interaction [20, 21]. The final discussion about the Tangram task was designed as a conversational context that was as open and relaxed as possible, while also guaranteeing a conversation on the basis of shared knowledge and experience.

2.5. Data

The pilot dataset consists of 149 minutes of conversation in total. Eye-tracking data were logged at 200 Hz. Audio data were recorded with *Audacity* (version 2.4.2) at a sampling rate of 44100Hz (32-bit). The scene video was recorded with *Quicktime* player, and the videos recorded by the eye-tracking glasses were logged with 30 frames per second.

2.6. Processing and analysis

For the processing of the eye-tracking data, we used the dedicated *PupilCloud* software [19]. For annotation of audio files, Praat [22] was used. Further data processing and visualisation was conducted using *Python* and *R*. Eye-tracking and speech data were pre-processed separately. Eye-tracking data were stored as time series data in a .csv file, audio data independently as .wav files.

For each participant, eye-tracking behaviour was automatically merged with the video data of the eye-tracking glasses and dummy-coded as *gaze_in_face* using the fixation detection and face detection enrichment in *PupilCloud*, which detects the facial region of the interlocutor. Both *gaze_in_face* files were merged via a *Python* script by calculating a time offset relative to the start of recordings. For all consecutive timestamps, the mean of both files was used, leading to a negligible latency between the two *gaze_in_face* data points of up to half a frame.

Speech data were annotated in *Praat*, semi-automatically for interpausal units (with a minimum silence of 200 milliseconds) and manually for tasks, turns, backchannel types, backchannel tokens and backchannel functions. The resulting TextGrids were merged into the *gaze_in_face* time series format. All pre-processing and analysis scripts can be found on the *OSF* project page https://osf.io/su5nx.

3. RESULTS

Overall, we found that for the discussion context, PR backchannels were more often accompanied by gaze directed towards the interlocutor than was the case for IS tokens, but this did not hold true for the other two conversational contexts (overall—PR: 39%, IS: 33.7%; introduction—PR: 52.2%, IS: 56%; Tangram—PR: 12.9 %, IS: 15.1%; discussion—PR: 48.8%, IS: 35.7%). Figure 2 shows an example of the interaction between gaze and verbal feedback behaviour from one dyad (Dyad 2).

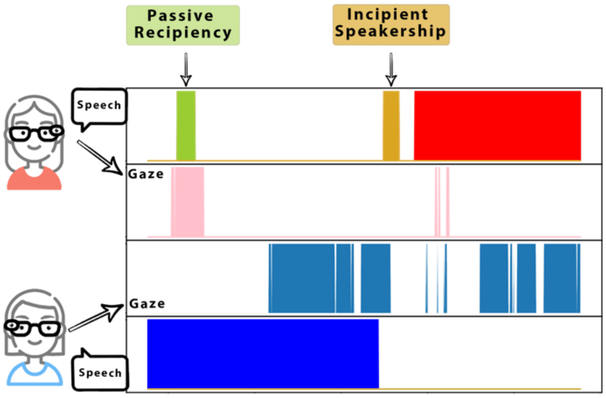

Figure 2: Example turn-plot showing 5 seconds of a conversation from Dyad 2. PR in green, IS in yellow.

In total, the data set contained 1208 verbal feedback tokens. The overall proportion of dialogue featuring verbal feedback was 4.95% (PR: 4.03 %, IS: 0.93%). This proportion was relatively stable for the different dyads (Dyad 1: 6.27%, Dyad 2: 4.67%, Dyad 3: 5.3%, Dyad 4: 4.31%). Overall, the proportion was highest in the introduction and similar (but somewhat lower) in the discussion, but far lower in the Tangram context (introduction: 6.23 %, Tangram: 3.69%, discussion: 5.65%). In all tasks, more PR than IS tokens were produced, but this difference was more pronounced for the introduction and discussion (~ 5 times more PR than IS) than for the Tangram task (~ 3 times more PR than IS).

Overall, the proportion of gaze directed towards the interlocutor was again highest during the introduction, similar (but somewhat lower) during the discussion and far lower during the Tangram task (introduction: 78.7%, Tangram: 24.2%, discussion: 73.5%).

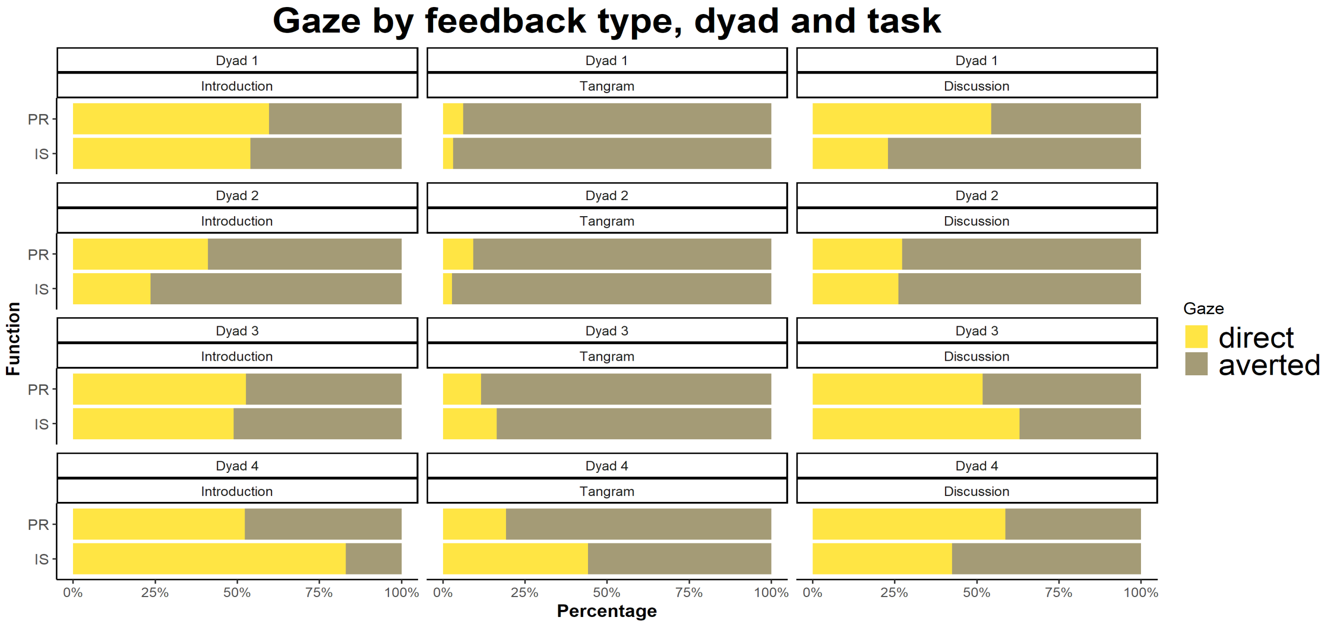

Figure 3: Proportion (on the x-axis) of gaze directed towards the partner (in yellow) vs. averted gaze (in grey) during the production of Passive Recipiency (PR) and Incipient Speakership (IS) types of verbal feedback (on the y-axis), as a function of dyad and communicative context.

Similarly, face-directed gaze *during the production of verbal feedback* was most frequent in the introduction, slightly less frequent in the discussion and very rare in the Tangram task.

On the dyad level, we observed differences in how much time was spent on backchanneling, and in how often gaze was directed towards the partner while providing verbal feedback; see Figure 3. All four dyads showed changes in gaze behaviour according to context, and most dyads used more direct gaze during the production of PR as compared to IS across the contexts, although there were exceptions, notably Dyad 4 in the introduction and Tangram task and Dyad 3 in the Tangram task and the discussion.

4. DISCUSSION AND CONCLUSION

The goal of the present study was to explore verbal feedback signals (e.g. backchannels), visual feedback signals (i.e. gaze), and their interplay in face-to-face interactions with different conversational contexts and using a novel experimental setup.

The time spent producing verbal feedback was relatively similar across the three conversational settings, but the proportion was lower overall in task-oriented conversation (Tangram) as compared to more spontaneous speech (cf. [23–25]). Conversational context also had an influence on the type of verbal feedback produced: a higher proportion of IS tokens was used during the Tangram compared to the more spontaneous contexts. This increased use of Incipient Speakership tokens, signaling the intention to take over the speaker role, might reflect a functional motivation to complete the task efficiently.

The proportion of gaze directed at the interlocutor differed substantially between conversational contexts. The low proportion of face-directed gaze we found during the Tangram task was expected, as participants were holding a piece of paper and had to look at it regularly to complete the task. Regarding the interplay of gaze and verbal feedback, we found that only during the discussion, participants directed their gaze more towards the partner for PR as compared to IS. This effect was heavily influenced by the behaviour of Dyad 4, which showed opposite patterns for the two least natural contexts (introduction and Tangram).

This exploration of our pilot data set suggests that gaze and verbal feedback behaviour differs depending on the communicative context, and that gaze behaviour also changes depending on the type of verbal feedback used. In future work, we will use this highly promising, novel methodology of exploring multi-modal behaviour via dual mobile eye-tracking glasses to explore further conversational behaviours and different groups of speakers. We will focus in particular on the intonational realisation of verbal feedback, on analysing head nods and overall kinematic energy, and on the behaviour of autistic adults, a population that is known to have difficulties in social interaction and to use social cues, such as eye gaze and verbal feedback, in a different way.

5. ACKNOWLEDGEMENTS

The study was funded by the Deutsche Forschungsgemeinschaft (German Research Foundation) – Project-ID : 281511265 – SFB 1252.

6. REFERENCES

[1] L. V. Hadley, G. Naylor, A. F. De, and C. Hamilton, ‘A review of theories and methods in the science of face-to-face social interaction’, *Nature Reviews Psychology 2022 1:1*, vol. 1, no. 1, Art. no. 1, Jan. 2022, doi: 10.1038/s44159-021-00008-w.

[2] J. R. Kelly and J. E. McGrath, *On Time and Method*. Newbury Park: SAGE Publications, Inc, 1988.

[3] J. McGrath and F. Tschan, *Temporal Matters in Social Psychology: Examining the Role of Time in the Lives of Groups and Individuals*. Washington: American Psychological Association, 2004.

[4] B. Winter, ‘Spoken language achieves robustness and evolvability by exploiting degeneracy and neutrality’, *BioEssays*, vol. 36, no. 10, pp. 960–967, 2014.

[5] K. Drummond and R. Hopper, ‘Back Channels Revisited: Acknowledgment Tokens and Speakership Incipiency’, *Research on Language and Social Interaction*, vol. 26, no. 2, Art. no. 2, Apr. 1993, doi: 10.1207/s15327973rlsi2602_3.

[6] M. Savino, ‘Degrees of (un) certainty in Bari Italian yes-no question intonation: perceptual evidence’, in *The communication of certainty and uncertainty: Linguistic, psychological, philosophical aspects*, in 1, vol. 12. In Zuczkowski, A., Bongelli, R., Riccioni, I., & Canestrari, C., pp. 52–67.

[7] S. Sbranna, E. Möking, S. Wehrle, and M. Grice, ‘Backchannelling across Languages: Rate, Lexical Choice and Intonation in L1 Italian, L1 German and L2 German’, *Proc. Speech Prosody 2022*, pp. 734–738, 2022.

[8] S. Wehrle, ‘A Multi-Dimensional Analysis of Conversation and Intonation in Autism Spectrum Disorder’, PhD Thesis, University of Cologne, Cologne, Germany, 2021.

[9] M. Argyle and M. Cook, *Gaze and Mutual Gaze*. Oxford: Cambridge University Press, 1976.

[10] Z. Degutyte and A. Astell, ‘The Role of Eye Gaze in Regulating Turn Taking in Conversations: A Systematized Review of Methods and Findings’, *Front. Psychol.*, vol. 12, p. 616471, Apr. 2021, doi: 10.3389/fpsyg.2021.616471.

[11] G. Skantze, A. Hjalmarsson, and C. Oertel, ‘Turn-taking, feedback and joint attention in situated human-robot interaction’, *Speech Communication*, vol. 65, p. 66, 2014, doi: 10.1016/j.specom.2014.05.005.

[12] B. Butterworth and G. Beattie, ‘Gesture and Silence as Indicators of Planning in Speech’, in *Recent Advances in the Psychology of Language: Formal and Experimental Approaches*, R. N. Campbell and P. T. Smith, Eds., in NATO Conference Series. Boston, MA: Springer US, 1978, pp. 347–360. doi: 10.1007/978-1-4684-2532-1_19.

[13] A. Kendon, ‘Some functions of gaze-direction in social interaction’, *Acta Psychologica*, vol. 26, pp. 22–63, 1967, doi: 10.1016/0001-6918(67)90005-4.

[14] P. Auer, ‘Turn-allocation and gaze: A multimodal revision of the “current-speaker-selects-next” rule of the turn-taking system of conversation analysis’, *Discourse Studies*, vol. 23, no. 2, pp. 117–140, Apr. 2021, doi: 10.1177/1461445620966922.

[15] Ho, T. Foulsham, and A. Kingstone, ‘Speaking and Listening with the Eyes: Gaze Signaling during Dyadic Interactions’, *PLOS ONE*, vol. 10, no. 8, Art. no. 8, Aug. 2015, doi: 10.1371/journal.pone.0136905.

[16] D. R. Rutter, G. M. Stephenson, K. Ayling, and P. A. White, ‘The timing of Looks in dyadic conversation’, *British Journal of Social and Clinical Psychology*, vol. 17, no. 1, Art. no. 1, Feb. 1978, doi: 10.1111/J.2044-8260.1978.TB00890.X.

[17] J. B. Bavelas, L. Coates, and T. Johnson, ‘Listener Responses as a Collaborative Process: The Role of Gaze’, *Journal of Communication*, vol. 52, no. 3, pp. 566–580, Sep. 2002, doi: 10.1111/j.1460-2466.2002.tb02562.x.

[18] J. Holler, K. H. Kendrick, M. Casillas, and S. C. Levinson, *Editorial: Turn-Taking in Human Communicative Interaction*, vol. 6, no. DEC. 2015. doi: 10.3389/fpsyg.2015.01919.

[19] M. Tonsen, C. K. Baumann, and K. Dierkes, ‘A High-Level Description and Performance Evaluation of Pupil Invisible’. arXiv, Sep. 01, 2020. Accessed: Dec. 05, 2022. [Online]. Available: http://arxiv.org/abs/2009.00508

[20] J. Holler, J. Bavelas, J. Woods, M. Geiger, and L. Simons, ‘Given-New Effects on the Duration of Gestures and of Words in Face-to-Face Dialogue’, *Discourse Processes*, vol. 59, no. 8, pp. 619–645, Sep. 2022, doi: 10.1080/0163853X.2022.2107859.

[21] M. Savino, L. Lapertosa, and M. Refice, ‘Seeing or Not Seeing Your Conversational Partner: The Influence of Interaction Modality on Prosodic Entrainment’, in *Speech and Computer*, A. Karpov, O. Jokisch, and R. Potapova, Eds., in Lecture Notes in Computer Science. Cham: Springer International Publishing, 2018, pp. 574–584. doi: 10.1007/978-3-319-99579-3_59.

[22] Boersma and Weenink, ‘PRAAT: Doing phonetics by computer (Version 6.1.40)’, 2021. http://www.praat.org/ (accessed Jul. 30, 2021).

[23] A. Janz, ‘Navigating Common Ground Using Feedback in Conversation- A Phonetic Analysis’, MA thesis, University of Cologne, Cologne, Germany, 2022.

[24] C. Dideriksen, R. Fusaroli, K. Tylén, M. Dingemanse, and M. H. Christiansen, ‘Contextualizing conversational strategies: backchannel, repair and linguistic alignment in spontaneous and task-oriented conversations’, in *CogSci'19*, Cognitive Science Society, 2019, pp. 261–267.

[25] R. Fusaroli, K. Tylén, K. Garly, J. Steensig, M. H. Christiansen, and M. Dingemanse, ‘Measures and mechanisms of common ground: Backchannels, conversational repair, and interactive alignment in free and task-oriented social interactions’, 2017.