Eye tracking and the perception of gestures in face-to-face interaction *vs* on screen

Marianne Gullberg
Max Planck Institute for Psycholinguistics, Nijmegen
marianne.gullberg@mpi.nl
and
Kenneth Holmqvist
Lund University Cognitive Science, Lund
kenneth@lucs.lu.se

Introduction

There is growing evidence that recipients attend to and retain information expressed in gestures (e.g. Cassell et al. 1999). However, there is a conspicuous lack of corresponding perceptual data, and we know very little about the relationship between perceptual and cognitive attention to gestures (and to Sign). In a previous eye-tracking study (Gullberg & Holmqvist 1999) we showed that recipients fixate only a minority of gestures (9 %) in face-to-face interaction, instead maintaining eye contact and chiefly perceiving gestures through peripheral vision. Only gestures performed in peripheral gesture space or fixated by speakers themselves were fixated. Thus, gestural performance features seem to compete with social norms for maintained eye contact in determining gesture fixation. In the absence of any social pressure for eye contact, as in a video setting, fixation behaviour towards gestures might therefore change, as suggested by the results from three video-based studies (Nobe et al. 1998, 2000, Rimé et al. 1988). However, these studies also differed with respect to agent (human vs. non-human) and degrees of speech comprehensibility. The present study therefore aimed to isolate the effect of the medium of presentation on gesture fixations by comparing recipients' fixations of naturally occurring co-speech gestures (McNeill 1992) in story retellings under two conditions, a live face-to-face and a video condition.

The study

We specifically set out to test if recipients fixate a) the speaker's face less often on video than live; b) more gestures overall on video than live; c) different gestures on video than live. We considered three gestural performance features that may affect fixations.

Place of performance in gesture space. A standard assumption in the visual literature is that peripheral movement captures visual attention. In an interactional pers-

pective, it is presumed that recipients fixate gestures in the periphery to ensure good perception of the gestural information.

Gestural hold (a momentary cessation of movement). Nobe et al. (1998) showed that poststroke holds are likely to attract fixations on video. We therefore test the attraction force of this feature in a live setting.

Autofixation (speakers fixate their own gestures). It has been claimed that speakers can direct recipients' gaze towards gestures intentionally in this way to achieve joint attention (e.g. Streeck 1993).

In the live condition, eight speakers retold a printed cartoon in Swedish to eight recipients facing them and wearing a head-mounted SMI iView eye-tracker. In the video condition, 16 new Swedish recipients were shown video recordings of the first set of speakers on a video screen. An SMI iView remote eye-tracker was placed between the recipients and the screen. This overall design allowed us to collect fixation data for the same gestures presented live and on video. Figure 1 shows an example of the data.

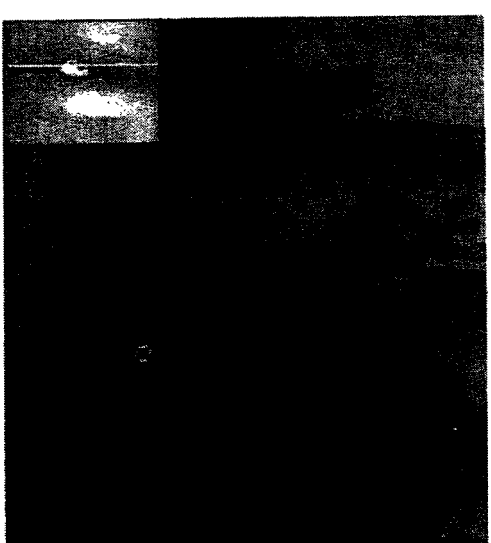

Figure 1.- Example of data showing the recipient's field of vision (the speaker en face), the recipient's fixation (the white circle), and an inlaid picture of the recipient's eye.

All gestures were coded for place of performance using McNeill's space schema (1992), for hold, and autofixation. All instances of autofixation were cases of enactment; i.e. the speakers looked at their own gestures acting as characters in the story. Speech was transcribed and checked for deictic expressions referring to gestures ('it was this long'). No such expressions were found. A spatial and a temporal

criterion determined fixations. The fixation marker had to remain in an area the size of the marker itself for at least 120 ms to count as a fixation (cf. Bruce & Green 1990). The fixation data were coded for object fixated (gesture, cup, etc.), and for duration.

Results

The results show that recipients predominantly fixate the speaker's face (on average 95 % vs. 92 % of the time) in both conditions. Second, in both conditions only a minority of gestures is fixated (on average 8 % vs. 3 %). Contrary to expectations, however, recipients fixate fewer gestures in the video condition, although not significantly so. Video recipients instead spend more time fixating objects in the room and immobile body parts. Third, there are some differences regarding which gestures are fixated across conditions. In both conditions, autofixated gestures are fixated significantly more often than other gestures (19 vs. 5 % and 6 vs. 2 %). Holds are fixated significantly more in the live condition (20 vs. 7 %), whilst non-holds are fixated more in the video condition (0 vs. 6 %). Gestures in peripheral space do not attract more fixations than central gestures in either condition (10 vs. 6 % and 3 vs. 3 %).

Discussion and conclusion

By and large, watching a human speaker on screen is surprisingly similar to watching one live. In fact, the main differences between the conditions do not seem to stem from the presence/absence of a live interlocutor, i.e. the purely social factor, but rather from the 'mechanical' effect of presentation size. The social factor clearly does not affect the dominance of the face, and its impact on overall amount of gesture fixations is weak. Furthermore, although the medium of presentation affects which gestures are fixated, this effect is unlikely to reflect the social factor per se. Remember that the interactional feature autofixation, by which recipients follow the speaker's gaze, operates in both conditions, whilst the mixed results apply only to the articulatory feature hold. The capacities of peripheral vision better explain these findings. Peripheral vision is good at motion detection, but bad at fine-grained texture. Given this design, peripheral vision would be insufficient for holds, as it can pick up neither motion nor configurational detail from them. Recipients have to fixate holds in order to retrieve any gestural information at all. This result is important, as it challenges received wisdom in the visuo-cognitive field: fixations are not attracted by movement, but rather by the lack of it – at least in the live condition. In the video condition, in contrast, the distance between the gesture and the fixation on the face is presumably short enough for peripheral vision to operate efficiently despite the lack of movement. The difference between our results for holds and those obtained by Nobe et al. (1998) is probably due to the difference in agent rather than to the medium of presentation itself. The lack of effect of place of performance was unexpected, given our previous findings. It suggests, however, that peripheral vision

was sufficient in both conditions, presumably providing motion information. It is not clear, however, why this feature should differ across our studies. Other qualitative, dynamic gesture features may play a role, but have not been considered here. Finally then, the only behaviour clearly affected by the presence/absence of a live speaker, and governed by social norms, is fixations of other things than the face or gestures, viz. the relative absence of body fixations in the live condition.

In sum, the effect of the medium of presentation on fixation behaviour towards gestures can be separated into a social and a more mechanical effect. The absence of a live speaker appears not to affect gesture fixations, since interactional features like the face and speakers' autofixations exercise the same force across conditions. Instead, they are affected by features related to the capacity of peripheral vision. A video-based paradigm for the study of gesture perception need thus not compromise ecological validity, provided that projection is life-sized, such that similar constraints are placed on peripheral vision across conditions.

References

Bruce, V. & Green, P. 1990, *Visual perception*, Hove, Erlbaum.

Cassell, J. et al. 1999, Speech-gesture mismatches: Evidence for one underlying representation of linguistic and nonlinguistic information, *Pragmatics & Cognition*, 7, 1-33.

Gullberg, M., & Holmqvist, K. 1999, Keeping an eye on gestures: Visual perception of gestures in face-to-face communication, *Pragmatics & Cognition*, 7, 35-63.

McNeill, D. 1992, *Hand and mind*, Chicago, Chicago University Press.

Nobe, S., et al. 1998, Are listeners paying attention to the hand gestures of an anthropomorphic agent? An evaluation using a gaze tracking method, in Wachsmuth, I. & Fröhlich, M. (eds), *Gesture and Sign Language in human-computer interaction*, Berlin, Springer, 49-59.

Nobe, S. et al. 2000, Hand gestures of an anthropomorphic agent: Listeners' eye fixation and comprehension, *Cognitive Studies. Bulletin of the Japanese Cognitive Science Society*, 7, 86-92.

Rimé, B. et al. 1988, Visual attention to the communicator's nonverbal behavior as a function of the intelligibility of the message, Paper presented at the 24th International Congress of Psychology, Sydney.

Streeck, J. 1993, Gesture as communication I: Its coordination with gaze and speech, *Communication Monographs*, 60, 275-299.