Mutual Visibility and Information Structure Enhance Synchrony between Speech and Co-Speech Movements

Petra Wagner[1,2] *and Nataliya Bryhadyr*[1]

[1]Faculty of Linguistics and Literary Studies, Bielefeld University
[2]Center of Excellence Cognitive Interaction Technology (CITEC), Bielefeld University
Bielefeld, Germany

petra.wagner@uni-bielefeld.de, n.bryhadyr@uni-bielefeld.de

Abstract

Our study aims at gaining a better understanding of how speech-gesture synchronization is affected by the factors (1) mutual visibility and (2) linguistic information structure. To this end, we analyzed spontaneous dyadic interactions where interlocutors are engaged in a verbalized version of the game TicTacToe, both with and without mutual visibility. The setting allows for a straightforward differentiation of contextually given and informative game moves, which are studied with respect to their manual and linguistic realization. Speech and corresponding manual game moves are synchronized more often when there is mutual visibility and when game moves are informative. Mutual visibility leads to a slight precedence of manual moves over corresponding verbalizations, and to a tighter temporal alignment of speech and co-speech movements. Informative moves counter the movement precedence effect, thus allowing co-speech movement targets to smoothly synchronize with prosodic boundaries.

1. Introduction

Previous research has shown that visibility has a strong effect on the frequency of gesture production, especially on *communicative gestures* that aid conversational interaction or highlight information, while *representational gestures*, co-expressing verbal message's content, are produced similarly frequent both with or without mutual visibility (Alibali et al., 2001; Bavelas et al. 2008). Many other results stress the communicative, listener-oriented function of co-speech gesturing, as speakers adapt their gesture rate based on the knowledge that they are seen, rather than seeing the interlocutor (Mol et al., 2011) and with gestures being larger and less reduced under visibility (Hoetjes et al., 2015). Similarly, de Ruiter et al. (2012) argue that speech and co-speech gestures tend to stand in a redundancy relationship that makes communication, especially conversational *grounding*, more robust.

Unlike frequency and gesture shape, we hitherto know very little about the effect of mutual visibility on speech-gesture synchrony. Speech-gesture synchrony has been pronounced a key feature of co-speech gesturing (McNeill, 1992), but is notoriously difficult to measure, given the variability of potential temporal anchors in the verbal and gestural stream, and the vagueness of lexical affiliates (cf. discussions in Esteve-Gibert and Prieto, 2013; Wagner et al, 2014). For beat gestures and deictic gestures, the gesture apex aligns with accented syllables in the speech stream (e.g. Loehr, 2012; Jannedy and Mendoza-Denton, 2005; Leonard and Cummins, 2010, Esteve-Gibert and Prieto, 2013). Additionally, there is evidence for alignment of gesture and prosodic boundaries (Loehr, 2012, Krivokapic et al., 2017, Jannedy & Mendoza-Denton, 2005). Many studies find that gestures or gesture apices precede speech, but tend to not lag behind (e.g. Esteve-Gibert and Prieto, 2013). Gesture lags are perceived as more asynchronous to speech and impede comprehension (Leonard & Cummins, 2010; Özyürek, 2008), but these effects may be a function

of gesture type (Kirchhof , 2017).

If speech and co-speech gesture are as tightly linked as suggested and mainly serve communicative robustness, they should be temporally synchronized to facilitate perceptual integration. We therefore expect that speech and gestures are aligned more precisely given mutual visibility. We furthermore expect that a message's information load enhances this temporal alignment. We test these assumptions in a study where interlocutors are engaged in a verbalized game of TicTacToe (cf. Watson et al., 2007), while simultaneously playing the game on a board with and without mutual visibility. When interlocutors cannot see each other, we expect a certain decoupling of speech and game-related manual movements, as the latter no longer fulfil any communicative function. Despite the redundancy of information conveyed visually and verbally under visibility, we expect an increase in speech-co-speech synchrony, further enhanced by communicative needs such as the highlighting important information. Such highlighting occurs either if a move is unexpected (as in early stages of the game) or game-relevant (as in later stages of the game). In our setting, moves can also be uninformative, e.g. when the game situation inevitably leads to a tie.

We are aware that the manual movements we examined do not qualify as traditional co-speech gestures spontaneously produced alongside with speech. Rather, they are co-speech movements elicited by the task. Still, they obviously fulfil a communicative function in constituting the game moves and are fully co-extensive in semantic content with their corresponding verbalizations.

2. Methods

2.1 Participants

We recorded 20 native speakers of German (10 dyads, friends in their 20s, unpaid volunteers, no control for gender, speakers stayed in one dyad) engaged in a verbalized version of TicTacToe.

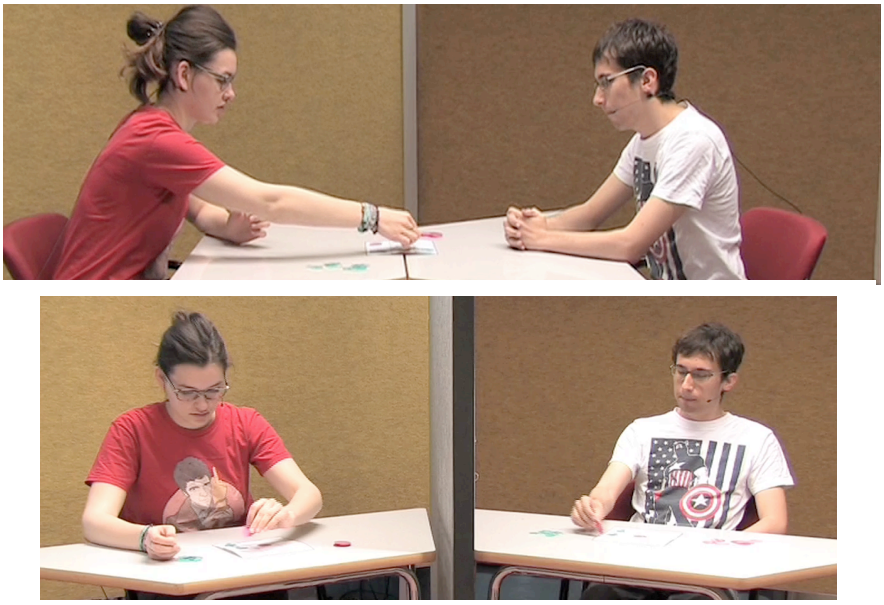

Figure 1: Recording setting with (top) and without (bottom) mutual visibility

2.2 Recording setup

Each dyad was recorded at our faculty's recording studio using Sennheiser neckband microphones (audio) and a studio camcorder (video) in two different recording conditions (cf. Figure 1):

- Visibility Condition: The players were seated facing each other, with a shared TicTacToe game board placed in the middle.
- Invisibility Condition: The players were seated on separate tables and were parted from each other by the movable wall, each of them having his or her own game board.

Each player received a set of cut outs in the form of a *tree* (ger. "Baum") and a *ball* (ger. "Ball") to make their moves. To control for order effects, we used an alternating initial recording condition with each newly recorded dyad, and per condition, 4 consecutive games were played. The game board looked like a normal TicTacToe grid, however with every cell being numbered. This enables the interlocutors to unambiguously refer to the different cells on the game board. A typical verbalized move is produced by placing a sentence accent on the target of the move, which corresponds to one of the numbers available on the game board – these accented verbalizations of numbers are later analyzed for their prosodic realization, e.g.

Ich lege einen Baum auf Feld **FÜNF.**
(engl.: I put a tree on field **FIVE.**)

Prior to each game, the players were informed about a preset first move. Also, the players alternated in setting the first move. On average, each recording session lasted 8.57 minutes per dyad, resulting in roughly 1.5 hours of recorded speech in total.

2.3 Annotations

The verbalized target moves, i.e. the sentence accented number realizations, were manually annotated using Praat (Boersma & Weenink, 2008).

The corresponding manual target moves were annotated with ELAN (Brugman & Russel, 2004) starting from a resting position or a position in which there is no target-oriented move, but in which the players hold the cut-out to be placed in their hand. The moves end with the full contact of the cut out on the game board. In cases where the players hold their target-oriented move before making full contact with the game board, the time of first contact is annotated as ending point of the gestural game move.

As the first move was preset by the experimenter and made known to both players before the game, it is annotated as *given*. In the case of a tie, the last move was annotated as *given*. The remaining moves were annotated as *informative*, as they led to a win, blocked a potential winning move, or were unpredictable (cf. Watson et al., 2008).

2.4 Measuring speech-movement synchrony

We analyzed the synchronization between the points at which the movement targets are reached and two prosodic anchors: pitch accent peaks and prosodic boundaries. As an estimate of gesture-boundary synchrony, we calculated the difference between the delay between the time the gestural moves reached their target on the game board and the corresponding end of a verbalized move, coinciding with a prosodic boundary (= gesture delay to prosodic boundary).

To estimate the synchrony between co-speech movements and pitch accents, we calculated the point of maximal pitch excursion for each verbalized target move using the pitch tracking functions of Praat. We then calculated the delay between pitch accent location and the time of corresponding co-speech movements. (= gesture delay to pitch accent)

Notice that a gesture preceding the prosodic anchor has a negative delay (lead), a gesture succeeding the verbalization a positive delay (lag).

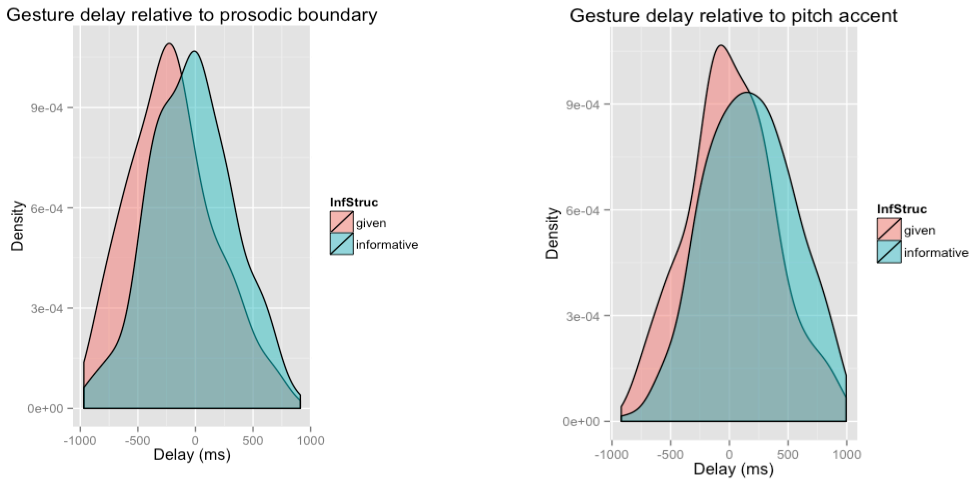

Figure 2: Delay between movement end points, prosodic boundaries (left) and pitch accents (right) in given and informative contexts.

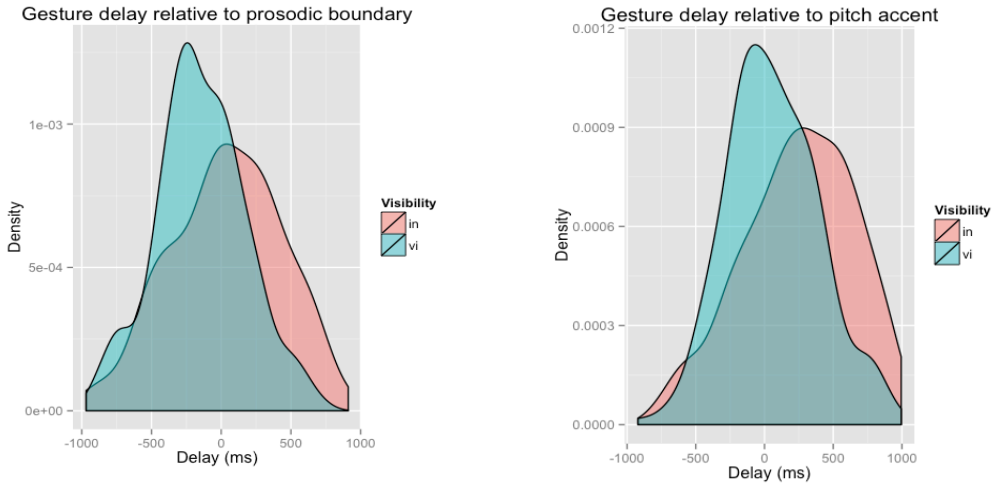

Figure 3: Delay between movement end points, prosodic boundaries (left) and pitch accents (right) with (vi) and without (in) mutual visibility.

3. Results

QQ-plots of the two delay measures (R package "car", Version 2.0–25) showed a violation of normal distribution due to outliers, e.g. instances of clear desynchronization between speech and co-speech movements. Based on the QQ-Plots, we define outliers as delays larger than ±1000ms. A closer analysis revealed that there are significantly more outliers in the absence of mutual visibility (18.6%) as compared to mutual visibility (4.0%, $\chi^2(1)=21.4, p<0.001$). We also found that there are significantly more outliers when the moves are not informative (24.7%) as compared to informative moves (8.6%, $\chi^2(1)=9.5, p<0.01$). After removing outliers, both types of delays show near-normal distributions. We subsequently constrained our analyses to delays between *-1000ms* (gesture end precedes prosodic boundary/pitch accent by *1000ms*) and *+1000ms* (gesture end follows prosodic boundary/pitch accent by *1000ms*).

The ends of co-speech movements tend to precede corresponding prosodic boundaries (*n=86,*

M=-204ms, SD=371ms) when the conveyed information is given as compared to when the message is informative, in which case the gestures reach their target almost perfectly aligned with the point of time where the verbalization is finished, showing only a minimal negative delay (*n=421, M=-30ms, SD=363ms*; cf. Figure 2). There is a similar difference for delays between co-speech movements and pitch accents (cf. Figure 2), with movements being earlier when the expressed information is given and closely aligned with pitch accents (*n=86, M=16ms, SD=371ms*) as compared to when the message is informative (*n=421, M=176, SD=376ms*). Across conditions, manual movements lag behind corresponding pitch accents, but precede prosodic boundaries.

The ends of co-speech movements precede corresponding prosodic boundaries (*n=271, M=-144ms, SD=320ms*) when there is mutual visibility, but slightly lag behind elsewhere (*n=236, M=37ms, SD=399ms*; cf. Figure 3) There is a similar effect for delays between co-speech movements and pitch accents (cf. Figure 3) with movements being earlier under mutual visibility (n=271, M=63ms, SD=336ms) as compared to lacking visibility (*n=236, M=247, SD=403ms*). However, the co-speech movements generally lag behind their corresponding pitch accents. F-tests to compare variances showed that speech-gesture synchrony is less variable under mutual visibility both relative to prosodic boundaries *(F(270,235)=0.65, p<0.001)* and pitch accents *(F(270,235)=0.70, p<0.01)*. However, gesture-speech synchrony was not affected by information structure.

The data collected in the recordings and subsequent annotations and acoustic measurements were further analyzed with the help of Linear Mixed Models using R (Version 3.1.2) (R Team, 2015) together with the R-packages lme4 (Version 1.1-7) and lmerTest (2.0-25). The resulting models contained the factors visibility (*visible—invisible*) and information structure (*informative—given*) as fixed, and word (1-9) and participant as random factors with random intercepts. Our two measures of delay (prosodic boundary delay, pitch accent delay) served as dependent variables. Both dependent variables were analyzed by reducing a maximal model including all fixed and random effects in a stepwise fashion, i.e. by removing all non-significant main fixed effects and interactions through log-likelihood ratio comparisons. This analysis showed no significant interactions. For the synchronization of gesture and prosodic boundary, a model comparison shows that the full model differs significantly from one not including visibility ($\chi^2(1)=46.5, p<0.001$) or information structure ($\chi^2(1)=21.8, p<0.001$). The model confirms the prior descriptive analyses that informative moves make co-speech movements occur significantly (*t(484)=4.7, p<0.001*) later *(+176ms, SE=37ms*) relative to prosodic boundaries, while visibility leads to a significantly (*t(484.3)=-7.0, p<0.001*) earlier production of co-speech movement *(-196ms, SE=28ms*). The model intercept also shows that on average, gestures precede prosodic boundaries *(-87ms, SE=58ms)* For the synchronization of gesture and pitch accent location, a model comparison shows that the full model including differs significantly from one not including visibility ($\chi^2(1)=46.2, p<0.001$) or information structure ($\chi^2(1)=14.9, p<0.001$). The model confirms the prior descriptive analyses that informative moves make co-speech movements occur significantly (*t(496.3)=3.9, p<0.01*) later *(+149ms, SE=38ms*) relative to corresponding pitch accents, while visibility leads to a significantly *(t(502.3)=-7.0, p<0.001)* earlier production of co-speech movements *(-203ms, SE=29ms)*. On average, gestures reach their targets after pitch accents *(149ms, SE=54ms)*

4. Discussion

Our outlier analysis revealed that clear violations of gesture-speech synchrony (delays larger ±1000ms) occur much more frequently if interlocutors cannot see each other and if the message conveyed is uninformative. This supports our assumption that both information structure and visibility increase speech-gesture synchrony. It also corroborates models claiming that speech-gesture alignment is to some extent caused by communicative needs. The fact that visibility decreases the variability in speech-gesture alignment further strengthens this finding. A possible interpretation is that the stronger "gesture lead" occurring under visibility aids the cross-modal integration on the side of the listener (Leonard & Cummins, 2010). Interestingly, informativity has the contrary effect and makes gestures appear a bit later as compared to when the message is not

informative. At first glance, this may endanger the communicative robustness probably achieved by the gesture-lead, at least under visibility conditions. However, a closer look at the distributions of speech-gesture alignment reveals that even though gestures may be later when the conveyed message is informative, they align almost perfectly with prosodic boundaries, but do not make the gesture strongly lag behind speech. Thus, it is unlikely that audiovisual integration on the side of the listener is endangered by this effect of information-structure on speech-gesture alignment. In line with our assumptions, it rather seems that informativity further strengthens speech-gesture synchronization. If the proximity of speech and co-speech movement is taken as an indicator of an anchor for speech-gesture alignment (though see the critical discussion in Leonard & Cummins, 2010), the pitch accent qualifies as the anchor for gesture apices/movement boundaries under visibility, and the prosodic boundary in case the message is informative. Still, the exact kind of this relationship needs to be investigated further, taking into account global aspects of prosody.

Overall, we have convincing evidence that both visibility and information structure enhance the temporal synchrony of speech and co-speech movements, thus supporting theories claiming that speech-gesture synchrony mainly fulfills a communicative function. However, due to the different distance of players to the game board across the two visibility conditions, our claims need further confirmation in follow-up studies. Also, the influence of the individual interlocutor dynamics (dyads) should be examined further. The co-speech movements we examined are no prototypical, spontaneously produced co-speech gestures, but their form and function closely resemble deictic gestures. At this point, we cannot say whether our results generalize to deictic gestures or to iconics or emblematics, for which speech-gesture synchrony may be less relevant (Kirchhof, 2017).

References

Alibali, M.W., Heath, D.C., & Myers, H.J. (2001). Effects of visibility between speaker and listener on gesture production: some gestures are meant to be seen. *Journal of Memory and Language 44*, 169–188.

Bavelas, J., Gerwing, J., Sutton, C., & Prevost, D. (2008). Gesturing on the telephone: independent effects of dialogue and visibility. *Journal of Memory and Language 58*, 495–520.

Boersma, P. & Weenink, D. (2008). Praat: Doing phonetics by computer (version 6.0.21). Retrieved from http://www.praat.org. [accessed Apr 7, 2017].

Brugman, H. & Russel, A. (2004). Annotating Multimedia/Multi-modal resources with ELAN. In: Proceedings of LREC 2004, Fourth International Conference on Language Resources and Evaluation, Lisbon, Portugal. Retrieved from: http://tla.mpi.nl/tools/tla-tools/elan/

Jannedy, S. & Mendoza-Denton, N. (2005). Structuring information through gesture and intonation. *Interdisciplinary Studies on Information Structure 3*, 199–244.

Esteve-Gibert, N. & Prieto, P. (2013). Prosodic Structure Shapes the Temporal Realization of Intonation and Manual Gesture Movements. *Journal of Speech, Language, and Hearing Research 56*, 850–864.

Hoetjes, M., Koolen, R., Goudbeck, M, Krahmer, E., & Swerts, M. (2015). Reduction in gesture during the production of repeated references. *Journal of Memory and Language 79–80,* 1–17.

Kirchhof, C. (2017). *The shrink point: audiovisual integration of speech-gesture synchrony*. Bielefeld: Universität Bielefeld. Retrieved from: https://pub.uni-bielefeld.de/publication/2908762.

Krivokapić, J., Tiede, M., & Tyrone, M. (2017). A Kinematic Study of Prosodic Gesture in Articulatory and Manual Gestures: Results from a Novel Method of Data Collection. Laboratory Phonology. *Journal of the Association for Laboratory Phonology 8*(1), 3. DOI: http://doi.org/10.5334/labphon.75 [accessed Apr 7, 2017].

Leonard, T. & Cummins, F. (2010). The temporal relation between beat gestures and speech. *Language and Cognitive Processes 26 (10),* 1457–1471.

Loehr, D. (2012). Temporal, structural, and pragmatic synchrony between intonation and gesture. *Laboratory Phonology*. *Journal of the Association for Laboratory Phonology 3*, 71–89.

McNeill, D. (1992). Hand and Mind: What Gestures Reveal about Thought. University of Chicago Press, Chicago.

Mol, L., Krahmer, E., Maes, A., & Swerts, M. (2011). Seeing and Being Seen: The Effects on Gesture Production. *Journal of Computer-Mediated Communiation 17(1),* 77–100.

Özyürek, A., Willems, R.M., Kita, S. & Hagoort, P. (2007). On-line integration of semantic information from speech and gesture: insights from event-related brain potentials. *Journal of Cognitive Neuroscience 19 (4),* 605–616.

de Ruiter, J.P., Bangerter, A., & Dings, P. (2012). Interplay between gesture and speech in the production of referring expressions: investigating the trade-off hypothesis. *Topics in Cognitive Science 4 (2),* 232–248.

R Team (2015). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org.

Wagner, P., Malisz, Z., & Kopp, S. (2014). Gesture and speech in interaction. *Speech Communication 57*, 209–232.

Watson, D.G., Arnold, J.E., & Tanenhaus, M.K. (2008). TicTacTOE: Effects of Predictability and Importance on Acoustic Prominence in Language Production. *Cognition 106(3),* 1548–1557.