Abstract: Part of the justification for an integrated view of speech and gestures is their temporal coordination. Gestures generally coincide with or precede, but rarely follow lexical affiliate (McNeill, 1992). How synchrony impacts listeners remains less explored, despite potential relevance for video communication and virtual conversational agents. ERP studies suggest that temporal alignment affects how words and gestures are integrated (Obermeier & Gunter, 2015) (Habets et al., 2011). Explicit perception of asynchrony is less sensitive and shifts longer than 1s can be tolerated (Kirchhof, 2014). However, gestures that are preceded by their lexical affiliates deviate from the expected pattern given regular exposure to speech which might implicitly affect listeners. We investigated whether the asymmetry of timing observed in production was reflected in differential effects of gestures shifted in either direction on how listeners perceive the speaker’s behavior as natural (Exp1) and/or impairing their processing and subsequent recall of words. (Exp2) Using motion capture to animate virtual speakers (giving explanations) allowed shifting specific gesture strokes within longer segments while preserving synchronized lip movements. For 16 short segments we produced videos in 3 conditions defined by the timing of a target gesture stroke relative a target word; either overlapping (SYNC) or shifted 500ms earlier (G-BEFORE) or later (G-AFTER). We classified the verbal content overlapping with shifted strokes by (unequally frequent) categories: “congruent”, “incongruent” or “filled/unfilled pauses”. In Exp1, 32 participants saw a composition of 4 videos from each of the 3 mentioned conditions plus a variation of SYNC with distorted pitch during a few non-target words (AUDIO). After each video the participants rated their impression that it was based on a capture of natural or was artificially generated (by an undefined algorithm). We transformed each participant’s responses to the range between 0 (most artificial) and 1 (most natural). Results revealed no significant differences between conditions. However, comparing the ratings between the categories of overlap revealed that strokes shifted to “filled/unfilled pauses” were rated as more artificial. In Exp2, 79 participants saw all 16 videos in one of four
conditions. SYNC, G-BEFORE and G-AFTER were contrasted by a condition with seamlessly extinguished target gestures. Following each video and a distraction task, participants attempted to repeat what they heard in the video. Results revealed impaired recall of target words with extinguished or delayed gestures. In summary, asynchronous gestures were not perceived as less natural if overlapping with any words. Synchronous- and preceding- but not following- gestures facilitated recall, as expected if the processing of speech and gestures (involved in this particular task) would be tuned to temporal patterns common in natural speech.

Keywords: Co-speech gestures, multimodal integration, timing, animation, recall, memory, comprehension