Gesture and speech are linked in time and meaning. The linkage is limited to the lexico-semantic integration windows of theme-rheme pairs, which by itself already indicates that the temporal synchrony of speech and gesture is more flexible than previously thought. While we know that listeners do perceive both hand and mouth movements (e.g. Gullberg & Kita 2009; Vatakis et al. 2008), audiovisual integration (AVI) research has so far mainly focused on lip-synchrony. The temporal window of successful AVI in the latter is optimal between ~0-250ms of asynchrony (Wassenhove et al. 2007; cf. Massaro et al. 1996). Recent ERP studies show that co-occurring gesture strokes and words are integrated by the listener at least up to an auditory delay of 160ms (Habets et al. 2011; Özyürek et al. 2007). But how large can the asynchrony between gestures and their conceptual affiliates be before it is perceived as unnatural?

In the studies by Kirchhof & de Ruiter (2012), subjects were shown sentence-long clips of narrators in frontal view in an online interface. The videos were desynchronized at six levels between -600ms and +600ms, based, among others, on Massaro et al. (1996) with 0ms as the (natural) control. Audio gaps were filled with silences, video gaps with stills. A second and third condition had blurred faces/ a box covering all head features. 618 native speakers of German rated the perceived naturalness of 9327 stimuli on a 4-point Likert scale. In condition 1 all results are around chance (SD=~10) except for +200ms and -600ms (~73%), which is consistent with Wassenhove et al. (2007, v.s.). In the two obscured-head conditions, subjects rated all stimuli as ~68% natural (SD=~5.5). These findings suggest that the AVI window of gesture is rather large. In a follow-up study, 5 stimuli with asynchronies of -600ms, +200ms, and the control in each condition were rated against each other for naturalness. While lip-visibility resulted in a 50/50 preference of 0ms and +200ms, the head-obscured stimuli again had more random ratings across asynchronies, with a lead of +200ms.

The second study let subjects re-synchronize a selection of the clips used in study one in a slider interface, and additional stimuli of two physical events were added as controls. In contrast to a judgment task, the subjective window of audiovisual integration, or rather preferred synchrony of bimodal signals was tested here. The results show a classic Gaussian distribution for the physical stimuli, i.e. an acceptable range of stimulus onset asynchronies (SOA) of -978ms (video first) to +442ms (audio first), with an SOA mean of +14 ms (stddev 246). The gesture stimuli, on the other hand, were synchronized by the subjects reaching from SOA of -1778 ms (gesture first) to +754 ms (speech first), with a mean of -72 ms (stddev. 422).

From these results it can be concluded that the language perception system is very tolerant of both semantic and temporal synchrony, and that the observed synchronies between gesture and speech are possibly "accidental" side effects of the architecture of the speech production system. However, only one type of gestures has been analyzed here: iconic ones. Furthermore, the number of physical stimuli has to be expanded in order to harden the results. We are currently conducting a follow-up study, which we would like to present at the DGfS Workshop next year. The subjects are again using the slider interface to re-synchronize these stimuli as well as speech and gesture. This time, the study includes emblems and deictic gestures in addition to regular and metaphoric iconics in order to research whether the range of the AVI window is valid across gesture categories.
Selected References


