The Role of Synchrony and Ambiguity in Speech–Gesture Integration during Comprehension

Boukje Habets[1], Sotaro Kita[2], Zeshu Shao[2], Asli Özyurek[3,4], and Peter Hagoort[3,5]

Abstract

■ During face-to-face communication, one does not only hear speech but also see a speaker's communicative hand movements. It has been shown that such hand gestures play an important role in communication where the two modalities influence each other's interpretation. A gesture typically temporally overlaps with coexpressive speech, but the gesture is often initiated before (but not after) the coexpressive speech. The present ERP study investigated what degree of asynchrony in the speech and gesture onsets are optimal for semantic integration of the concurrent gesture and speech. Videos of a person gesturing were combined with speech segments that were either semantically congruent or incongruent with the gesture. Although gesture and speech always overlapped in time, gesture and speech were presented with three different degrees of asynchrony. In the SOA 0 condition, the gesture onset and the speech onset were simultaneous. In the SOA 160 and 360 conditions, speech was delayed by 160 and 360 msec, respectively. ERPs time locked to speech onset showed a significant difference between semantically congruent versus incongruent gesture–speech combinations on the N400 for the SOA 0 and 160 conditions. No significant difference was found for the SOA 360 condition. These results imply that speech and gesture are integrated most efficiently when the differences in onsets do not exceed a certain time span because of the fact that iconic gestures need speech to be disambiguated in a way relevant to the speech context. ■

INTRODUCTION

Linguistic communication is one of the most important backbones of human society. Communication with speech is often accompanied by spontaneous hand gestures regardless of the cultural background and age of speakers (Kita, 2009; McNeill, 1992, 2000, 2005; Goldin-Meadow, 2003). Gestures make a crucial contribution to communication, and the listener integrates information from speech and gesture to form a unified representation of the speaker's message (Cassell, McNeill, & McCullough, 1999). One important feature of the speech–gesture relationship is that their relative timing varies greatly in natural conversation (Morrel-Samuels & Krauss, 1992). The speech–gesture integration process is likely to be affected by synchronization of the two modalities as other types of multimodal integration processes are also known to be sensitive to synchronization (Munhall, Gribble, Sacco, & Ward, 1996). Thus, this study investigates how the synchronization of speech and gesture influences the speech–gesture integration process at the semantic level.

Among different types of speech-accompanying gestures that have been identified (McNeill, 1992), this study focuses on the so-called iconic gestures, in which the gestural form resembles its referent (e.g., a downward hand movement representing an object falling). The integration between speech and iconic gesture is interesting because the two modalities have distinct semiotic properties (McNeill, 1992, 2000). Although a spoken message is encoded in a discrete and sequential manner, with each word adding meaning to the message, a gesture depicts an event as a whole. For example, in the sentence "The car slid sideways," the verb conveys the movement and the verb particle conveys the direction. However, the accompanying gesture would combine both movement and direction in one movement with a flat hand making a sliding movement to the side. Thus, the listener/viewer in conversation is often presented with semantically related information in different representational formats in the visual and auditory modalities.

The listener/viewer pays attention to iconic gestures and the information they convey when they accompany speech (Beattie & Shovelton, 1999; McNeill, Cassell, & McCullough, 1994; Graham & Argyle, 1975). For example, Graham and Argyle (1975) presented descriptions of line drawings with and without gestures to listeners. The listeners were more accurate in reproducing the line drawings that were described with gestures than the ones described without

[1] University of Hamburg, Germany, [2] University of Birmingham, United Kingdom, [3] Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands, [4] Radboud University Nijmegen, The Netherlands, [5] Donders Institute for Brain, Cognition and Behaviour, Nijmegen, The Netherlands

gestures. Beattie and Shovelton (1999) showed that listeners were more accurate in recounting the relative position and size of objects when the description of events involving these objects was accompanied by gestures than when the description was presented without gestures.

Although iconic gestures accompanying speech contribute to communication, the meaning of iconic gestures is vague when they are presented without any accompanying speech (Beattie & Shovelton, 1999; Krauss, Morrel-Samuels, & Colasante, 1991; Feyereisen, van de Wiele, & Dubois, 1988). For example, in Beattie and Shovelton's (1999) study, the participants were presented with descriptions of events involving moving objects and then asked to recount various physical features of the event. They were far less accurate in this task when presented with the gesture part of the description without the accompanying speech than when presented with only the speech part or with both the speech and the gesture parts of the description. The inaccuracy in the gesture only presentation probably stems from the fact that a wide range of interpretations can be drawn from an iconic gesture in the absence of accompanying speech. Thus, concurrent speech and gesture enrich each other's interpretation (Kelly, Özyürek, & Maris, 2010).

Integration processes for various types of semantic information have been investigated by electrophysiological methods, using the N400, a negative deflection between 200 and 500 msec, as a marker of semantic integration processes. Originally, the N400 effect was used to investigate semantic integration in language. For example, a greater N400 was observed when a word was semantically incongruent to the preceding sentence context (e.g., "socks" in "She spread her bread with socks") than when it was congruent ("butter" in "She spread her bread with butter"; Kutas & Hillyard, 1980, 1984). More recently, the N400 effect has also shown to be sensitive to integration of linguistic information with extralinguistic contexts (Hagoort, Hald, Bastiaansen, & Petersson, 2004; Van Berkum, Zwitserlood, Brown, & Hagoort, 2003) and to integration of pictorial information (West & Holcomb, 2002; Federmeyer & Kutas, 2001; McPherson & Holcomb, 1999; Ganis, Kutas, & Sereno, 1996; Barrett & Rugg, 1990). For example, Hagoort et al. (2004) found an N400 effect when people's world knowledge, for example, their beliefs of what is false or true in their world, was violated. Studies looking at the effect of incongruent pictures with regard to a sentence context also revealed an N400 effect (Willems, Ozyurek, & Hagoort, 2008). The distribution of this effect appears in some studies more frontally in comparison to studies that used language stimuli (Ganis et al., 1996; Nigam, Hoffman, & Simons, 1992).

More recently, the N400 has been used to investigate semantic integration between speech and gesture. To be more specific, there is growing evidence that gestures and language are semantically integrated in the way words are integrated into the preceding linguistic context. For example, iconic gestures that did not semantically match the preceding sentence elicited a larger N400 than those that matched (Özyurek, Willems, Kita, & Hagoort, 2007). Gestures can also contribute to building a context into which subsequent words can be semantically integrated. Words that did not semantically match the preceding gesture elicited a larger N400 than those that semantically matched (Bernardis, Salillas, & Caramelli, 2008; Kelly, Kravitz, & Hopkins, 2004). In addition, words that did not semantically match the preceding gesture–speech combination elicited a larger N400 than those that matched (Holle & Gunter, 2007).

The previous studies on semantic processing of gestures were limited in their scope in that they did not manipulate the temporal relationship between gesture and speech. In naturally occurring discourse, the temporal relationship between coexpressive gesture and speech varies greatly. Gestures tend to be initiated before or with (but rarely after) the coexpressive words and the degree of gesture–speech asynchrony varies (Morrel-Samuels & Krauss, 1992; Butterworth & Beattie, 1978), although the gesture stroke, the meaning–bearing part of the gestural hand movement, tends to overlap with the coexpressive word (McNeill, 1992).

The present study investigated this yet to be explored aspect of semantic integration of gesture and speech. More specifically, it investigated the semantic integration of temporally overlapping speech and gesture and asked the question in what time-frame gesture and speech information are integrated best. To this end, we manipulated the asynchrony of gesture and speech onsets and the semantic congruity of the gesture and speech (i.e., match vs. mismatch stimuli). This investigation is important for two reasons. First, it helps us understand processing implications of the natural variation in gesture–speech asynchrony that occurs during spontaneous use. Second, it allows us to observe how speech and gesture influence each other's meaning interpretation on-line in different degrees of synchrony, given that gesture is inherently ambiguous without the help of speech context (Krauss et al., 1991). It is possible that, when gesture precedes the coexpressive word by a relatively large margin, the upcoming speech cannot influence the interpretation of gesture. Thus, an ambiguous interpretation of the gesture is finalized and stabilized before the word onset. This stable interpretation is then integrated with the meaning of the upcoming word. However, when a gesture starts simultaneously or overlaps earlier with a coexpressive word, the unfolding interpretation of the gesture that is less ambiguous in relation to the overlapping speech is integrated with the meaning of the word.

In the current study, the stimuli consisted of gestures and verbs that referred to concrete events such as rotating, connecting, and so forth. In the SOA 0 condition, the gesture stroke and the speech onset were presented simultaneously. This condition was chosen because the onset of the stroke phase often coincides with the onset of the relevant speech segment in natural discourse (McNeill, 1992). In the SOA 160 and 360 conditions, the gesture onset preceded the speech onset by 160 and 360 msec, respectively. In all three conditions, the semantic relations between speech and gesture were manipulated, that is, match versus mismatch.

We predict that the three SOA conditions will show different integration processes. The pretest of the materials by a "gating" method (Marslen-Wilson, 1987) showed that when gestures are presented without speech, participants finalize the interpretation of the gestures after about 360 msec of the gesture onsets. Furthermore, the interpretations varied greatly across participants, in line with the previous finding that gestures without accompanying speech are ambiguous (Krauss et al., 1991). Thus, in the SOA 360 condition, participants would already have reached a final interpretation of the gestures at speech onset but importantly the interpretation would vary across participants. On the other hand, in the SOA 160 and the SOA 0 conditions, the interpretation of the gestures would still be ongoing at speech onset.

This manipulation could lead to two possible outcomes if one considers either an on-line integration between speech gesture where the two determine each other's meaning interpretation or an off-line one where the meaning of the word is integrated after the meaning of gesture has been determined. According to the first one, when the interpretation of gesture is still ongoing, the concurrent speech should play an important role in narrowing down gestural interpretations (e.g., on-line integration of gesture and speech information). Because the speech information is being used to derive a final interpretation of the gesture, the interpretations of the gesture and the speech should semantically converge in match stimuli, leading to a clear difference between match and mismatch stimuli, reflected in an N400 effect. In case of the SOA 360 condition, the interpretation of the gesture should have stabilized before the speech onset, and because of the ambiguity of the gesture, the interpretations of the gesture and the speech may not always semantically converge even in match stimuli. This would lead to no difference between match and mismatch stimuli in this condition. However, in the case of the second possible outcome, an N400 effect will be found for the condition where gesture interpretation is finalized around the moment of speech onset because it is easier to compare the additional speech information to a gesture interpretation that is stable (e.g., off-line integration of gesture and speech information). In that case, a clear distinction between match and mismatch stimuli, as reflected in an N400 effect, will only be found for the SOA 360 condition because the interpretation of the gesture has already been finalized around speech onset. The N400 effect would then be absent for the SOA 0 and SOA 160 conditions because the interpretation of the gesture is still ongoing and has not been finalized at the moment speech is presented.

METHODS

Participants

Twenty-three subjects (15 women, mean age = 20 years, range = 18–23 years) with normal or corrected-to-normal vision took part in this study. All subjects were right-handed and had English as their native language. They gave written informed consent and were paid for their participation. Subjects with more than 25% loss of trials caused by blinking or movement artifacts were excluded from further analysis. Nine subjects were excluded, leaving 14 subjects for the final analysis (7 women).

Materials

Sixteen target gestures iconically represented concrete events (e.g., connecting, bouncing) (see Table 1 for an overview of all the target gestures). For example, in the connecting gesture, the right and left hand were located at the same height in front of the chest with the extended index fingers pointing at each other, and the two hands moved horizontally toward each other until the tips of the index fingers touched each other. In the bouncing gesture, a flat hand with the palm facing down moved up and down. Thirty similar filler gestures were used as well. Each gesture video captured a frontal view of a woman (see Figure 1). It started with 1000 msec of still picture taken from the frame immediately before the onset of the preparation phase, and this was followed by the onset of the stroke phase of the gesture (Kita, Van Gijn, & Van der Hulst, 1998; McNeill, 1992; the mean duration of the stroke was 490 msec, within

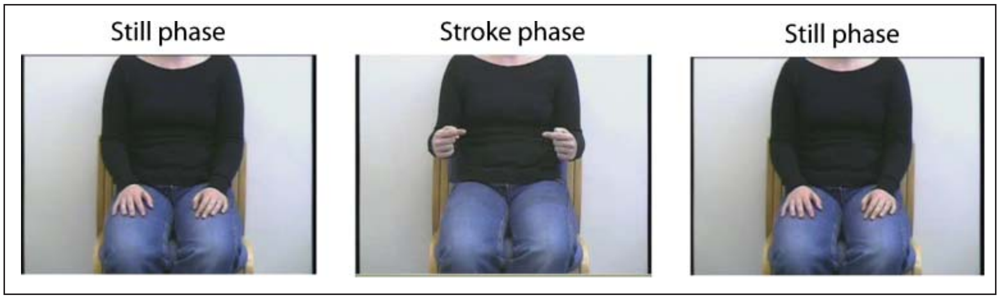

Figure 1. Three frames from a movie used as a stimulus. Left, a frame from the initial still phase, followed by a frame of the stimulus "connecting" (the gesture and the verb "connecting" constituted a semantically congruent combination). On the right, a frame of the end phase of the video (the last frame of the retraction phase).

the range 400–720 msec). The stroke phase is the main meaning bearing part of the gesture (McNeill, 1992) and is usually preceded by the preparation phase that brings the hand from the resting position to the beginning location of the gesture stroke. In the gesture videos, the preparation phase was removed by video editing to make the point at which gestural information became available fully controlled (the same editing procedure as in Özyurek et al., 2007). After the stroke phase, the hands went back to the original starting position (the retraction phase). Then the last frame of the retraction was shown as a still frame to make the total duration of the video 3000 msec. To eliminate information from lips and head posture, only the torso of the speaker was visible in the video. To make sure that the gestures appeared as natural as possible, the model produced semantically congruent speech while performing the gestures. All videos were filmed against the same background with a digital video camcorder (MCX4i) and edited with *iMovie* HD 6.0.3. During editing, the audio was removed from the video.

The speech stimuli were verbs that were semantically congruent to one of the gestures. There were 16 target verbs, listed with a short description in Table 1, and 30 filler verbs. Verbs were spoken by a female native English speaker and digitized at a sample frequency of 16 bit (44.100 kHz) with a mean duration of 413 msec within the range 230–600 msec (*SD* = 103 msec). The filler verbs had a similar mean duration as the target verbs (553 msec). The CELEX database was used to check for frequency for all the target verbs (mean = 13.7, range = 0–109). Each verb was extracted from the recording with Praat 4.5.02 (www.praat.org).

The gestures were then combined with the speech segments into congruent and incongruent pairs, with three different degrees of gesture–speech asynchrony (with *iMovie* HD 6.0.3). For example, the gesture "connecting" was combined with the speech segment "connecting" with three different degrees of gesture–speech asynchrony to create the congruent stimuli. Incongruent stimuli were created by combining the gesture "connecting" with the speech segment "falling," again with three different speech onsets. Thus, 96 target gesture–speech combinations were created from 16 target gestures and verbs.

The same was done for the 30 filler gestures and verbs, resulting in 180 gesture–speech combinations. The filler combinations were added not only to minimize recognition effects on the target gestures but also to make sure that similar target gesture–speech combinations were not presented in close succession. The filler gestures were not taken into the analysis.

Pretesting of the material was done before the current study to test the degree of matching between gesture and speech. Twenty-two subjects, who did not participate in the EEG study, were asked to check for the degree of matching in the congruent and incongruent gesture–speech combinations (with a 1–7 matching scale, 1 = *perfect match* to 7 = *not matching*). The preparation phase was removed in these videos as well, and as this pretesting was solely done to test whether gesture and speech were found similar in meaning with an off-line judgment, no SOA was created for these videos. The congruent combinations had a significantly higher match rating (median = 6.0, interquartile range = 1.02) than the incongruent combinations (median = 1.04, interquartile range = 0.77), $t(21) = 23.2$, $p < .001$.

In the three SOA conditions, the gesture onset preceded verb onset or gesture and verb onset were presented simultaneously. In other words, the three SOA conditions varied with respect to the amount of gestural information available before the onset of speech.

In the SOA 0 condition, the gesture onset and the speech onset were simultaneous. This condition was included as the onset of the stroke phase often coincides with the onset of the relevant speech segment in natural discourse (McNeill, 1992). In this condition, no gestural information was available before the speech onset.

The other two SOAs were selected on the basis of results from a gating experiment that determined the amount of gesture information necessary to reach a stable interpretation of gestural information. A group of 20 native English speakers, who did not participate in the following EEG study, watched the 16 target gestures, without accompanying speech. These videos were divided into 12 gates (the first gate showed the first 120 msec of the gesture). Each of the subsequent gates added 40 msec (one video frame) on top of the preceding gate. The last gate then showed the entire gesture. The participants' task was to write down the interpretation (meaning) of the gesture. The relevant information was the stability point gate, in other words, the gate at which the participants gave the final interpretation of the gesture (i.e., the interpretation of the gesture did not change anymore in the subsequent gates). The participants reached a stability point on average at 7.4 gates (*SD* = 1.48). In other words, the participants needed approximately 376 msec to come to a final interpretation. On the basis of this result, for the last SOA condition, it was decided that gesture onset preceded speech onset by 360 msec. In this condition, a sufficient amount of gesture was presented before speech onset so that a stable interpretation of the gesture was available at speech onset.

However, this stable interpretation of the gesture did not necessarily match the semantically matching verb used in the current experiment. In fact, the stable interpretation included the semantically matching verb on average only 11.2% of the time (*SD* = 25.2). Each gesture was given on average 4.56 different stable interpretations (*SD* = 1.55). This is consistent with the finding that iconic gestures are inherently ambiguous without accompanying speech (Krauss et al., 1991).

In the SOA 160 condition, the degree of asynchrony was roughly the midway between the SOA 0 condition and the SOA 360 condition. Namely, gesture onset preceded speech onset by 160 msec.

Procedure

The stimuli were presented using Presentation 11.0 (www.neurobs.com). The subjects sat 50 cm from the computer screen, and the dimensions of the videos were 10 × 10 cm. The movies were played at a rate of 25 frames per second. The verbs were presented over two speakers that were placed left and right from the computer screen. A trial started with a 500-msec black screen, followed by the presentation of the video. The video was followed by a 1000-msec fixation cross.

Participants were instructed to sit as still as possible and to carefully pay attention to the videos and the verb phrases. Further, they were instructed that blinking was

Table 1. The Description of the Target Gestures and Corresponding Words in the Match and Mismatch Conditions

	Target Words	
Target Gestures	*Match*	*Mismatch*
(1) The two fists are placed on top of each other, as if to hold a club, and they move away from the body twice.	Battering	Hurdling
(2) The right fist moves downward forcefully twice, as if to stab something.	Bludgeoning	Oscillating
(3) The left extended index finger pointing upward stays motionless. The right extended index finger pointing downward makes a half circle away from the body and around the left index finger.	Bypassing	Swivelling
(4) The two extended index fingers pointing away from the body start at the body midline and moves in symmetrical arcs laterally and back, as if to conduct an orchestra.	Conducting	Juddering
(5) The two extended index fingers pointing at each other move toward the body midline until they touch.	Connecting	Flattening
(6) The left fist stays motionless. The right loose open hand moves across and land on the left fist to cover it.	Covering	Rotating
(7) The two hands are put together with the palms facing each other and the fingers interlocked, as if to pray, and the hands are twisted about 90°.	Entwining	Wobbling
(8) The two flat open hands with the palms facing downward move downward twice.	Flattening	Connecting
(9) The left extended index finger pointing to the right stays motionless. The right hand with the index and middle fingers extended downward jumps away from the body over the left index finger.	Hurdling	Battering
(10) The two loosely open hands face each other. They forcefully move away from the body twice.	Juddering	Conducting
(11) Both hands form a loose cup facing upward. They alternately move sharply upward once.	Juggling	Rewinding
(12) The right extended index finger facing away from the body traces two peaks and troughs laterally.	Oscillating	Bludgeoning
(13) The right extended index finger pointing away from the body makes two circles.	Rewinding	Juggling
(14) The right extended index finger facing downward makes a circle twice.	Rotating	Covering
(15) The right open flat hand with the palm facing downward rotates to the direction of the little finger twice.	Swivelling	Bypassing
(16) The right loose open hand with the palm facing left moves to the left and back while the wrist and fingers flexing lithely.	Wobbling	Entwining

only allowed at times where a fixation cross was visible. Target and filler speech–gesture combinations were presented in a pseudorandom order. Each target speech–gesture combination was always followed by one, two, or three filler speech–gesture combinations. To make sure that participants kept attentive throughout the experiment, they had to fill in a questionnaire about both videos and verb phrases at the end of the experiment.

A practice session containing 10 videos, which differed from the ones used in the main experiment, preceded the start of the experiment. The whole session lasted approximately one hour.

EEG Recording and Analysis

The EEG was recorded from 128 electrode sites across the scalp, relative to an (off-line) averaged left and right ear reference. The electrodes were placed according to the 10-5 electrode system (Oostenveld & Praamstra, 2001) using a nylon electrode cap. Vertical eye movements were measured with bipolar montages from an electrode above the left eyebrow and an electrode placed below the left orbital ridge. Two electrodes placed at the left and right external canthus measured horizontal eye movements. EEG data were recorded continuously using a band-pass filter of 0.01–30 Hz with a sampling rate of 250 Hz. The EEG recordings were filtered off-line with a 20-Hz low-pass filter. Epochs were time locked to speech onset and baseline corrected with a prestimulus interval of 100 msec preceding gesture onset. To make sure that baseline correction was done with "clean" 100 msec where no stimuli were presented, this 100 msec was chosen before gesture onset for all three conditions. This resulted in a prestimulus interval of −100 to 0 msec for the SOA 0 condition, −260 to 160 msec for the SOA 160 condition, and −460 to 360 msec for the SOA 360-msec condition.

Trials contaminated by eye movement artifacts and drifts were rejected off-line using a maximal allowed absolute voltage difference between 80 and 140 μV. No more than 25% of the trials in each particular condition of a given participant were rejected because of eye-movement artifacts.

Repeated measures ANOVAs with the factors SOA (0, 160, 360), match (match, mismatch), Hemisphere (left, right), and Regions (anterior, central, posterior) were conducted. Significant interactions were followed up by another ANOVA or *t* tests (when two conditions were compared). As the N400 effect reveals itself as a large deflection in the EEG signal, the 128 electrodes were divided into six subgroups, with each group containing six electrodes of the 10/20 position system. The electrodes clustered for hemisphere and regions are as follows: left anterior: F1, F3, F5, FC1, FC3, FC5; left central: C1, C3, C5, CP1, CP3, CP5; left posterior: P1, P3, P5, PO3, PO5, O1; right anterior: F2, F4, F6, FC2, FC4, FC6; right central C2, C4, C6, CP2, CP4, CP6; and right posterior: P2, P4, P6, PO4, PO6, O2.

A separate ANOVA with factors SOA, match, and Electrodes was performed for the midline electrodes (Fz, FCz, Cz, CPz, Pz, Oz). An alpha level of .05 was used for all statistical tests. Huynh–Feldt correction for violation of sphericity assumption was applied when appropriate (Huynh & Feldt, 1976).

RESULTS

Figure 2 displays the grand average waveforms time locked to the onset of the speech segment. First, visual inspection of the grand averages (Figure 2) revealed a more pronounced N1–P2 complex for the SOA 0 condition than for the SOA 160 and 360 conditions. This is due to the fact that only at SOA 0 the N1–P2 complex reflects coincided gesture and speech processing. In the other two SOA conditions, the waveforms show processes triggered by speech onset only. A clear comparison between the three SOA conditions for these early components is therefore difficult because the amount of visual/gestural information preceding speech onset differs over the three conditions. Furthermore, because these early processes are not known to reflect semantic integration, they are not crucial for our results and are therefore not discussed any further.

In the SOA 0 condition, the N1–P2 complex was followed by a negative deflection between 300 and 900 msec. This large latency window has been found in previous gesture–speech integration studies with regard to the N400 (Bernardis et al., 2008; Wu & Coulson, 2005). Match and mismatch for this condition are similar on the N1–P2 complex, but they start to diverge at the N400 component, with mismatch stimuli being more negative in amplitude compared with match stimuli (see also Figure 2 for the mismatch–match difference wave). The SOA 160 condition reveals a similar N400 component, starting around 300 msec. Match and mismatch for SOA 160 start to diverge around 130 msec, with mismatch stimuli being more negative in amplitude compared to match stimuli. The SOA 360 condition does not show an N400 effect for match versus mismatch. Visually, it appears that the N400 effect is peaking between 400 and 600 msec in the SOA 0 condition. In the SOA 160 condition, a peak is more difficult to distinguish. To make a best possible comparison between all three conditions, the time window 300–900 msec was chosen for analysis.

For this time window, mean amplitudes were submitted to a repeated measures ANOVA with the factors SOA, Match, Hemisphere, and Region. The analysis revealed a main effect of Region, $F(2, 26) = 12.4, p = .002$, and interactions between SOA and match, $F(2, 26) = 4.7, p = .030$, and between SOA and Regions, $F(4, 52) = 5.7, p = .001$. A separate ANOVA with factors SOA and Match for the Midline revealed a main effect of Match, $F(1, 13) = 7.5, p = .017$, and an interaction between SOA and Match, $F(2, 26) = 3.9, p = .047$. Because the Hemisphere factor did not show any significant results, it was taken out of the subsequent analysis.

To probe the interaction between SOA and match found in the ANOVA described above, planned t tests comparing the match and mismatch conditions were carried out separately for each SOA condition. The mismatch condition was significantly more negative than the match condition for the SOA 0 condition (non-midline regions, $t(13) = 2.11$, $p = .049$; midline, $t(13) = 2.46$, $p = .028$) and the SOA 160 condition (non-midline regions, $t(13) = 3.8$, $p = .004$; midline, $t(13) = 3.58$, $p = .003$) but not for the SOA 360 condition (non-midline regions, $t(13) = -0.5$, $p = .628$; midline, $t(13) = -0.035$, $p = .972$).

Furthermore, a separate ANOVA was carried out for the SOA × Region interaction, revealing a main effect of Region, $F(2, 26) = 12.44$, $p = .002$, and an SOA × Region interaction, $F(4, 52) = 5.8$, $p = .001$. To probe the interaction between SOA and Region, planned t tests compared the three regions within each SOA. The amplitudes were significantly higher in central and posterior regions for the SOA 0 and SOA 360 condition (SOA 0: anterior vs. central, $t(13) = -7.1$, $p = .00$; anterior vs. posterior, $t(13) = -3.8$, $p = .002$; central vs. posterior, $t(13) = 1.2$, $p = 2.3$; SOA 360: anterior vs. central, $t(13) = -4.1$, $p = .001$; anterior vs. posterior, $t(13) = -1.6$, $p = .1$; central vs. posterior, $t(13) = 1.1$, $p = .3$). For the SOA 160 condition, the highest amplitudes were found on central electrodes (anterior vs. central, $t(13) = -5.1$, $p = .00$; anterior vs. posterior, $t(13) = -1.4$, $p = .176$; central vs. posterior, $t(13) = 2.8$, $p = .15$).

These results show that the mismatch condition was significantly more negative than the match condition for SOA 0 and SOA 160 (for the topographical distribution of

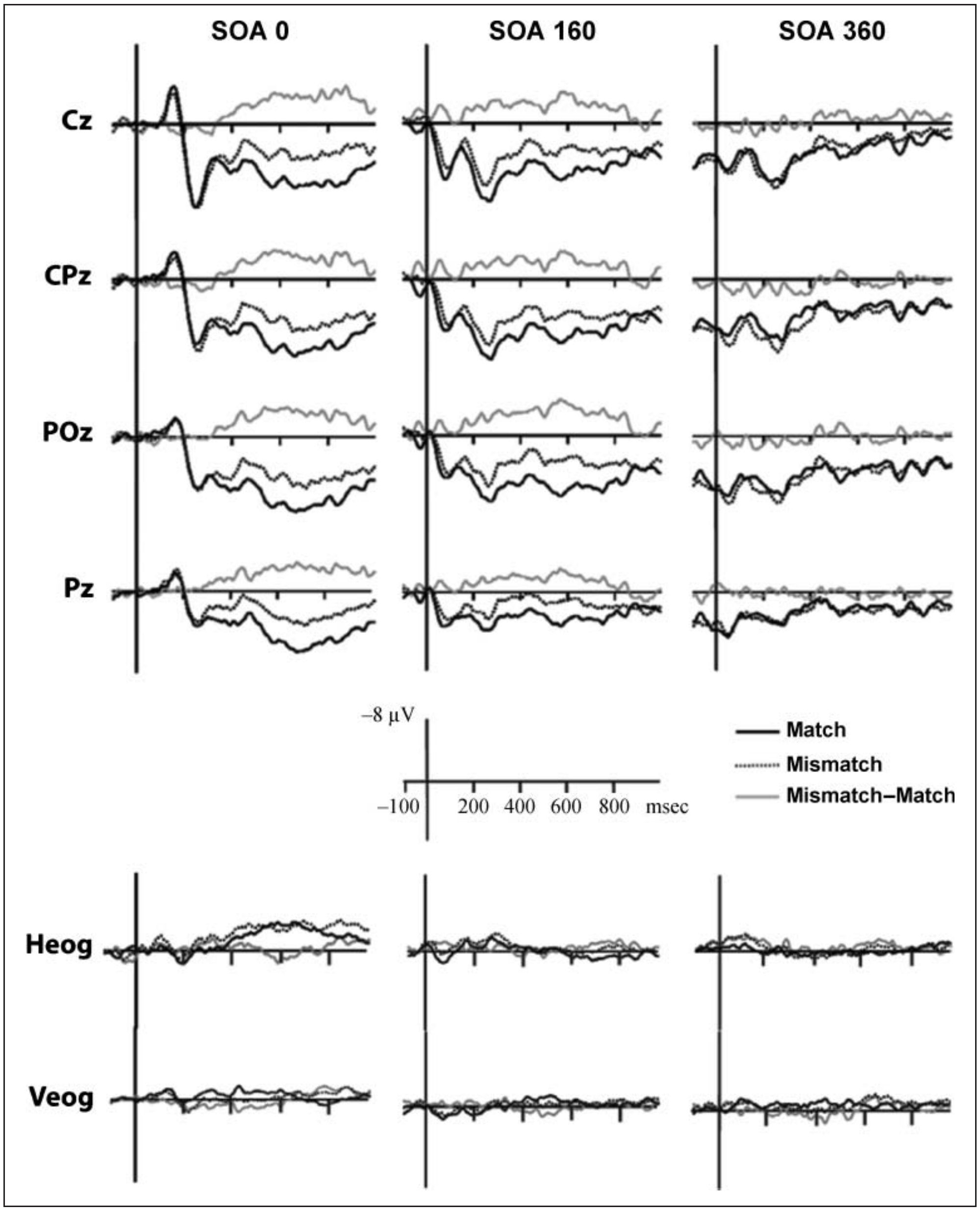

Figure 2. Grand-average waveforms for ERPs elicited in the match and mismatch pairs and the difference wave (mismatch–match) at four midline electrode sites (Cz, CPz, Pz, and POz) and a horizontal (Heog) and vertical (Veog) eye electrode for each SOA condition. Negativity is plotted upward. Waveforms are time locked to the onset of the speech (0 msec) and displayed with a −100 prestimulus interval (Note that only SOA 0 shows gesture and speech onset simultaneously. Speech onset in SOA 160 and SOA 360 is preceded by gesture onset by 160 and 360 msec, respectively).

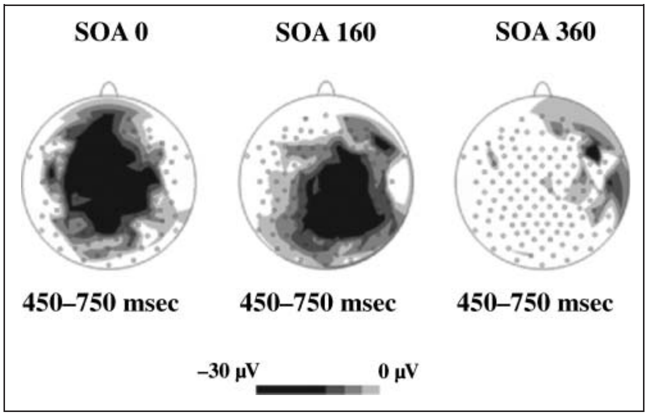

Figure 3. Spline-interpolated isovoltage maps displaying the topographic distributions of the mean difference between mismatch versus match from 450 to 700 msec for the three SOA conditions on all electrodes (Note that this map shows the most visually prominent part of the 300- to 900-msec time window. Statistical analysis on this smaller time window shows identical results as the analysis on the 300- to 900-msec time window).

the effects, see Figure 3[1]). No significant difference between match and mismatch was found for the SOA 360 condition (see Figures 2 and 3). Furthermore, greatest amplitudes were found on centro-parietal sites for the SOA 0 and SOA 360 condition, whereas the SOA 160 condition revealed highest amplitudes in central regions.

DISCUSSION

This study investigated semantic integration of concurrent gesture and speech by manipulating the asynchrony time window between gesture and speech onsets as well as the semantic congruency of gesture–speech. Three different asynchrony conditions were used; the SOA 0 condition in which gesture and speech were presented simultaneously, the SOA 160 condition in which gesture onset preceded speech onset by 160 msec, and the SOA 360 condition in which gesture onset preceded speech onset by 360 msec.

ERP analysis revealed several results. Most importantly, the semantic congruency manipulation yielded the following pattern of results. Mismatching gesture–speech combinations lead to a greater negativity on the N400 component in comparison with matching combinations. This indicates that semantic integration of speech and gesture information was more complex for incongruent gesture–speech pairs, a finding in line with results of previous ERP studies on gesture–speech integration (Holle & Gunter, 2007; Özyurek et al., 2007; Kelly et al., 2004). However, this difference between mismatching and matching stimuli was found for the SOA 0 and the SOA 160 condition but was absent for the SOA 360 condition. In other words, incongruency in gesture–speech information affected gesture–speech integration only within a certain time window of gesture and speech onsets.

Why was the N400 effect observed in the SOA 0 and SOA 160 conditions but not in the SOA 360 condition? There are two possible explanations. First, as mentioned in the Materials section, gesture interpretation in the SOA 360 condition stabilized around speech onset. Consequently, gesture interpretation might not be influenced by the information carried by speech. Second, because of the ambiguousness of the gestures, their interpretation at the moment of speech onset may not have been semantically congruent to the speech information, even in the match condition. This explanation was corroborated by the pretest of the materials, in which participants gave an interpretation of the gesture consistent with the meaning of the target verb in only 11% of the cases in the match condition. Thus, the ambiguity of the gestures may have reduced the difference between congruent and incongruent gesture–speech combinations, leading to the absence of an N400 effect in the SOA 360 condition. In contrast, in the SOA 0 and 160 conditions, the interpretation of the gestures was still ongoing at speech onset, and the speech could thus shape the gesture interpretation. In other words, the ambiguous gesture could be interpreted in the context of the concomitant speech, provided that a congruent interpretation was possible (i.e., in the match condition). To summarize, in the SOA 0 and 160 conditions, semantically matching speech and gesture were integrated in an optimal way because the interpretation of the gesture could be, to some extent, guided by the speech to increase the semantic congruity between the two modalities. In contrast, in the SOA 360 condition, the semantic congruity between speech and gesture was not maximized. That is, the interpretation of the gesture was fixed before the speech onset, and because of the ambiguity of the gesture (i.e., as evidenced by multiple meaning interpretations across participants) in the absence of speech, the interpretation of the gesture often did not semantically match the speech even in the match condition where the two modalities could have been semantically congruent.

The fact that the SOA 360 condition did not show an N400 effect may, at first sight, seem at odds with the results from previous studies in which gesture stimuli preceded linguistic stimuli by a longer interval. In Kelly et al. (2004), the gesture onset preceded the speech onset by 800 msec. Similarly, in Bernardis et al. (2008) and Wu and Coulson (2005, 2007a), a gesture preceded a visual presentation of a word by a 1000-msec ISI. These three studies found an N400 effect because of the semantic congruity of gesture and linguistic stimuli, unlike the SOA 360 condition in the present study. However, in the previous studies, the interpretation of stimulus gestures may have been unambiguous without accompanying speech. For example, Kelly et al. used only four speech tokens, with four corresponding gestures, and repeatedly presented them as a matching, mismatching, or complementary combination. This made the interpretation of gestures highly predictable and unambiguous even without accompanying speech. In Wu and Coulson (2005), they presented a cartoon depicting the referent of the following gesture in a half of the critical trials. This may have made the interpretation of the gesture less ambiguous overall. In Wu and Coulson (2007a), it is possible that the gestures were less ambiguous than others. They selected materials that would be used in a task in which participants explicitly judged whether a gesture and a subsequently presented word semantically matched or not. This might have prompted them to select gestures that were relatively unambiguous on their own. In Bernardis et al. also, it appears that the authors specifically selected gestures that were unambiguously interpretable and linguistically namable to be used as stimuli (e.g., both hands in the O-shape in front of the eyes to refer to a binocular). This contrasts with the current study in which the pretest indicated that the gestures were ambiguous without speech, as in naturally occurring speech-accompanying gestures (Krauss et al., 1991). Thus, the seemingly conflicting findings from previous studies can be explained by the fact that the previous studies made gestures more unambiguous without speech in one way or another, unlike what happens in most everyday situations.

This study corroborates previous findings showing that listeners/viewers integrate information from speech and gesture by influencing each other's interpretation in an on-line (rather than off-line) fashion (Kelly et al., 2010). However, it goes beyond the findings in the literature by taking into account the variation in temporal synchrony and the inherent ambiguity of gesture in the absence of speech into account. Previous studies have shown that speech and gesture can be integrated into a stable (preexisting) and unambiguous context, similarly to how a visually presented word can be integrated into a preceding sentential context (Kutas & Hillyard, 1980). For example, they have shown that a word is integrated into a preceding gesture (Bernardis et al., 2008; Wu & Coulson, 2005; Kelly et al., 2004), that a gesture is integrated into the context created by a preceding sentence (Özyurek et al., 2007), and that a picture probe or a word was integrated with a preceding gesture–speech combination (Holle & Gunter, 2007; Wu & Coulson, 2007b). When previously the integration of concurrent speech and gesture was demonstrated (Kelly, Ward, Creigh, & Bartolotti, 2007; Kelly et al., 2004), a small number of predictable gestures and words were repeated a number of times, and thus the gestures were not as ambiguous as they usually are. Consequently, they could serve as a stable context for speech comprehension. However, in everyday multimodal communication, speech provides a context for the interpretation of inherently ambiguous gestures (the prior in Bayesian terms). Thus, speech and iconic gestures are most effectively integrated when they are fairly precisely coordinated in time.

Acknowledgments

This research was supported by a grant from the Biotechnology and Biological Sciences Research Council (BBSRC/08906) and the Netherlands Organization for Scientific Research (NWO/051.02.040).

Reprint requests should be sent to Boukje Habets, University of Hamburg, 20146 Hamburg, Germany, or via e-mail: Boukje.Habets@uni-hamburg.de.

Note

[1] Figure 3 displays the most prominent part of the 300- to 900-msec time window. This smaller 450- to 750-msec time window shows identical statistical results as the 300- to 900-msec time window analysis: GLM: SOA × Regions interaction, $F(2, 35) = 7.9$, $p = .001$. t Tests: SOA 0: anterior versus central, $t(13) = -7.2$, $p = .00$; anterior versus posterior, $t(13) = -4.5$, $p = .00$; central versus posterior, $t(13) = 1.18$, $p = .26$. SOA 160: anterior versus central, $t(13) = -5.5$, $p = .00$; anterior versus posterior, $t(13) = -1.8$, $p = .08$; central versus posterior, $t(13) = 2.4$, $p = .25$. SOA 360: anterior versus central, $t(13) = -4.1$, $p = .00$; anterior versus posterior, $t(13) = -1.2$, $p = .23$; central versus posterior, $t(13) = 1.7$, $p = .1$. GLM: SOA × Match interaction, $F(2, 26) = 5.7$, $p = .03$. t Tests: SOA 0: match versus mismatch, $t(13) = 2.2$, $p = .04$. SOA 160: match versus mismatch, $t(13) = 2.6$, $p = .01$. SOA 360: match versus mismatch, $t(13) = 0.95$, $p = .35$.

REFERENCES

Barrett, S. E., & Rugg, M. D. (1990). Event-related potentials and the semantic matching of pictures. *Brain and Cognition, 14,* 201–212.

Beattie, G., & Shovelton, H. (1999). Do iconic hand gestures really contribute anything to the semantic information conveyed by speech? An experimental investigation. *Semiotica, 123,* 1–30.

Bernardis, P., Salillas, E., & Caramelli, N. (2008). Behavioural and neuropsychological evidence of semantic interaction between iconic gestures and words. *Cognitive Neuropsychology, 25,* 1114–1128.

Butterworth, B. L., & Beattie, G. W. (1978). Gesture and silence as indicators of planning in speech. In R. N. Campbell & P. T. Smith (Eds.), *Recent advances in the psychology of language 4: Formal and experimental approaches* (pp. 347–360). London: Plenum.

Cassell, J., McNeill, D., & McCullough, K. E. (1999). Speech–gesture mismatches: Evidence for one underlying representation of linguistic and non-linguistic information. *Pragmatics & Cognition, 7,* 1–33.

Federmeyer, K. D., & Kutas, M. (2001). Meaning and modality: Influences of context, semantic memory, organization, and perceptual predictability on picture processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27,* 202–204.

Feyereisen, P., van de Wiele, M., & Dubois, F. (1988). The meaning of gestures: What can be understood without speech? *Cahiers de Psychologie Cognitive, 8,* 3–25.

Ganis, G., Kutas, M., & Sereno, M. (1996). The search for common sense: An electrophysiological study of the comprehension of words and pictures for reading. *Journal of Cognitive Neuroscience, 8,* 89–106.

Goldin-Meadow, S. (2003). *Hearing gesture: How our hands help us think*. Cambridge, MA: Harvard University Press.

Graham, J. A., & Argyle, M. (1975). A cross-cultural study of the communication of extra-verbal meaning by gestures. *International Journal of Psychology, 10,* 57–67.

Hagoort, P., Hald, L., Bastiaansen, M., & Petersson, K. M. (2004). Integration of word meaning and world knowledge in language comprehension. *Science, 304,* 438–441.

Holle, H., & Gunter, T. C. (2007). The role of iconic gestures in speech disambiguation: ERP evidence. *Journal of Cognitive Neuroscience, 19,* 1175–1192.

Huynh, H., & Feldt, L. S. (1976). Estimation of the box correction for degrees of freedom from sample data in randomized block and split-plot designs. *Journal of Educational and Behavioral Statistics, 1,* 69–82.

Kelly, S., Özyürek, A., & Maris, E. (2010). Two sides of the same coin: Speech and gesture mutually interact to enhance comprehension. *Psychological Science, 21,* 260–267.

Kelly, S. D., Kravitz, C., & Hopkins, M. (2004). Neural correlates of bimodal speech and gesture comprehension. *Brain and Language, 89,* 253–260.

Kelly, S. D., Ward, S., Creigh, P., & Bartolotti, J. (2007). An intentional stance modulates the integration of gesture and speech during comprehension. *Brain and Language, 101,* 222–233.

Kita, S. (2009). Cross-cultural variation of speech-accompanying gesture: A review. *Language and Cognitive Processes, 24,* 145–167.

Kita, S., Van Gijn, I., & Van der Hulst, H. (1998). Movement phases in signs and co-speech gestures, and their transcription by human coders. In I. Wachsmuth & M. Fröhlich (Eds.), *Gesture and sign language in human–computer interaction* (pp. 23–35). Berlin: Springer.

Krauss, R. M., Morrel-Samuels, P., & Colasante, C. (1991). Do conversational hand gestures communicate? *Journal of Personality and Social Psychology, 61,* 743–754.

Kutas, M., & Hillyard, S. A. (1980). Reading between the lines: Event-related brain potentials during natural sentence processing. *Brain and Language, 11,* 354–373.

Kutas, M., & Hillyard, S. A. (1984). Brain potentials during reading reflect word expectancy and semantic association. *Nature, 307,* 161–163.

Marslen-Wilson, W. D. (1987). Functional parallelism in spoken word-recognition. *Cognition, 25,* 71–102.

McNeill, D. (1992). *Hand and mind*. Chicago: University of Chicago Press.

McNeill, D. (2000). *Language and gesture*. Cambridge: Cambridge University Press.

McNeill, D. (2005). *Gesture and thought*. Chicago: University of Chicago Press.

McNeill, D., Cassell, J., & McCullough, K.-E. (1994). Communicative effects of speech-mismatched gestures. *Research on Language and Social Interaction, 27,* 223–238.

McPherson, W. B., & Holcomb, P. J. (1999). An electrophysiological investigation of semantic priming wit pictures of real objects. *Psychophysiology, 36,* 53–65.

Morrel-Samuels, P., & Krauss, R. M. (1992). Word familiarity predicts temporal asynchrony of hand gestures and speech. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18,* 615–623.

Munhall, K. G., Gribble, P., Sacco, L., & Ward, M. (1996). Temporal constraints on the McGurk effect. *Perception & Psychophysics, 58,* 351–362.

Nigam, A., Hoffman, J. E., & Simons, R. F. (1992). N400 to semantically anomalous pictures and words. *Journal of Cognitive Neuroscience, 4,* 15–22.

Oostenveld, R., & Praamstra, P. (2001). The five percent electrode system for high-resolution EEG and ERP measurements. *Clinical Neurophysiology, 112,* 713–719.

Özyurek, 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,* 605–616.

Van Berkum, J. J., Zwitserlood, P., Brown, C., & Hagoort, P. (2003). When and how do listeners relate a sentence to the wider discourse? Evidence from the N400 effect. *Cognitive Brain Research, 17,* 701–718.

West, W. C., & Holcomb, P. J. (2002). Event-related potentials during discourse-level semantic integration of complex pictures. *Cognitive Brain Research, 13,* 363–375.

Willems, R., Ozyurek, A., & Hagoort, P. (2008). Seeing and hearing meaning: ERP and fMRI evidence of word versus picture integration into a sentence context. *Journal of Cognitive Neuroscience, 20,* 1235–1249.

Wu, Y. C., & Coulson, S. (2005). Meaningful gestures: Electrophysiological indices of iconic gesture comprehension. *Psychophysiology, 42,* 654–667.

Wu, Y. C., & Coulson, S. (2007a). Iconic gestures prime related concepts: An ERP study. *Psychonomic Bulletin & Review, 14,* 57–63.

Wu, Y. C., & Coulson, S. (2007b). How iconic gestures enhance communication: An ERP study. *Brain and Language, 101,* 234–245.