# Resolving gesture-speech contradictions and response modality

by

Murat Dikmen

Koç University

A Thesis Submitted to the

Graduate School of Social Sciences and Humanities

in Partial Fulfillment of the Requirements for

the Degree of Master of Arts

in Psychology

Koç University

November 2014

**STATEMENT OF AUTHORSHIP**

This thesis contains no material which has been accepted for any award or any other degree or diploma in any university or other institution. It is affirmed by the candidate that, to the best of her knowledge, the thesis contains no material previously published or written by another person, except where due reference is made in the text of the thesis.

Signed Murat Dikmen

## ABSTRACT

This study examines the effects of conflicting bimodal instructions and the response medium on decision-making in a visio-spatial task. Participants (*N*=100), in an hypothetical navigation scenario, watched audio-visual clips with congruent and incongruent hand gesture and vocal directions, and had to pass the direction on either manually or vocally. Results showed that participants were faster when gesture and speech were congruent than incongruent. Also, regardless of the response medium, the information presented via gestures was the dominant information participants used in their responses.

**Keywords:** gesture-speech conflict, multimodal communication, deictic gestures, response modality.

## ÖZET

Bu çalışma, iki farklı modalitede sunulan çelişkili yönergelerin ve cevap mecrasının, görsel-mekansal bir görevde karar verme üzerindeki etkisini incelemektedir. 100 katılımcı, varsayımsal bir yön bulma senaryosunda, yön belirten uyumlu ve uyumsuz söz ve el hareketleri barındıran görsel-işitsel video klipleri izleyip tarif edilen yönleri el ile veya ses ile belirtmişlerdir. Sonuçlar, katılımcıların, söz ve jestler uyumlu iken, uyumsuz olmalarına göre daha hızlı olduğunu göstermiştir. Ayrıca, cevap mecrasından bağımsız olarak, katılımcıların cevaplarında, jestler ile sunulan bilginin, sözlerle sunulan bilgiye oranla daha baskın olduğu gözlemlenmiştir.

**Anahtar Sözcükler:** söz-mimik çatışması, çoklu modalite iletişimi, gösterici mimikler, cevap modalitesi.

**DEDICATION**

*To My Family*

**ACKNOWLEDGEMENTS**

First, I would like to express my deepest gratitude to Professor Gün Şemin for providing guidance and support during this process. His wisdom and guidance allowed me to overcome the difficulties along the way and his warm support provided me with the courage and motivation to complete my thesis. He introduced me the tools and knowledge that will help me to deal with challenges during my career. I am very grateful to have a chance to do research with him and write this thesis under his supervision.

I would like to thank Assoc. Prof. Zeynep Cemalcılar for her support throughout my graduate education. I am also thankful to Asst. Prof. Lemi Baruh for devoting his time to contribute to my thesis and sharing his insights.

I especially would like to thank Linda Timmerman and Evi-Anne van Dis for their invaluable help with data collection. Furthermore I would like to express my thanks to my friends, Ecem Esgin, Selen Küçükarslan and Demet Kara, for providing support and motivation during my graduate education. Special thanks to Neriman Saner, who has always been with me through tough times.

Last, I would like to thank my family, for their endless support and for being the source of inspiration and motivation during my journey.

## Resolving gesture-speech contradictions and response modality

On January 31, 2001, two airplanes operated by Japan Airlines had a near collision in mid-air when the pilots received conflicting instructions from their aircraft's Traffic Collision Avoidance System (TCAS) and from the flight controller in Tokyo Area Control Center. One of the pilots had followed the order issued by the TCAS and descended and the other pilot followed the order issued by the flight controller and also descended. In another similar situation, two airplanes, one of them operated by DHL and the other operated by Bashkirian Airlines, were not as lucky as the Japanese airplanes, and crashed in mid-air over Germany close to the town Überlingen on July 1, 2002, killing all 71 people. Had both pilots followed the order issued by TCAS, the accident would have been prevented. Among other problems, these examples illustrate a common problem: dealing with conflicting instructions, and specifically when conflicting information is received in multiple modalities (visual and auditory).

In the current study, we examine how people deal with conflicting information coming from two different modalities. The question we investigate is; how do people perceive, respond and take action when presented with conflicting information coming from auditory (verbal) and visual (gestural) modalities. We choose gestural and verbal information because these two modalities are the most prominent means of social communication[1]. Any type of communication problem is likely to result in adverse situations (Heath and Luff, 2000; Goodwin, 1995; Goodwin & Goodwin, 1996) as is particularly evident in the examples referred to above. Accurate and smooth information flow is an important part of successful work environments, and smooth communication is even more important in environments that are critical for safety, such as healthcare (Leonard et al., 2004) and aviation contexts (Krivonos, 2007). Also, in an emergency context such as a rescue operation or evacuation

[1] We exclude in this context computer based communication (e.g., email, twitter, blogs, etc.)

after a disaster, unambiguous communication is critical to success with communication problems having disastrous consequences. One of the ways in which the information flow can fail is when a message received from different modalities conflict. Our focus is how people perceive and make decisions in the presence of conflicting information presented in auditory (verbal) and visual (gestural) channels, and, in a second step, how people convey this information to other people down a communication chain.

In the following, we will first present an overview of the relevant literature on gestures as the visual medium, speech as the auditory medium and the relationship between speech and gestures. Next, we turn to multimodal perception. Following that, we will provide an overview of the current research.

## Gesture and Speech: Bimodal Communication

Gestures are important vehicles of communication. They often accompany and improve speech; although they can convey information on their own (Goldin-Meadow, 1999). Gestures have two main functions. They enhance communication (e.g. Bavelas et al. 2002) and help the speaker in speech production (e.g. Hostetter, Alibali, & Kita, 2007). There are different perspectives on the roles of gestures in the production and comprehension of language, such as the lexical gesture process model (Krauss, 1998), the interface model (Kita & Özyürek, 2003), and gestures-as-simulated-action framework (Hostetter & Alibali 2008, 2010). Past research provides evidence that gestures play an important role in conveying information (see, Goldin-Meadow & Alibali, 2013, for a review). For example, gestures convey critical information regarding content when depictions of cartoon stories are narrated using gestures. People can extract more information and consequently answer questions related to semantic meaning of the story more accurately when the story is presented with gestures accompanying speech compared to speech only (Beattie & Shovelton, 1999b). A

meta-analysis (Hostetter, 2011) of studies comparing messages presented only verbally (speech) with messages that are presented in both verbally and with gestures, found that gestures have a significant effect and improved communication efficacy. Gestures accompanying speech also increase memory for verbal content, as demonstrated by Galati and Samuel (2011).

There are different types of gestures. The categorization provided by McNeill (1992) is the traditional taxonomy of gestures. The categories in this taxonomy are iconic, metaphoric, deictic, and beat gestures. Iconic gestures refer to concrete objects or events, such as the gesture produced when defining a particular shaped box. Metaphoric gestures refer to abstract concepts such as producing backward and forward hand movements when referring past and future. Beat gestures are repetitive small movements that do not carry a meaning. Deictic gestures are gestures that are pointing behaviors that carry spatial information such as pointing forward and then right when telling someone to "go forward and turn right".

Information from speech and gestures seem to be processed in parallel in the brain (Willems, Ozyurek, & Hagoort, 2006; Ozyurek, Willems, Kita, & Hagoort, 2007). The motor system is involved in understanding action-related language, and neural processing of gestures is similar to neural processing in understanding words (see Willems & Hagoort, 2007, for a review), indicating a strong relationship between gestures and speech. Recent ERP and fMRI studies (e.g. Holle & Gunter, 2007; Özyürek et al. 2007) have shown that semantic information received in gestural and verbal modalities is integrated, sometimes very quickly, to form a semantic representation. Although this process is automatic, some studies (Kelly, Ward, Creigh, & Bartolotti, 2007; Kelly, Creigh, & Bartolotti, 2009) suggest that some degree of control is involved in the integration process. Kelly, Özyürek, and Maris (2010) further proposed an "integrated-systems hypothesis" to account for different processes through which this integration occurs. According to this hypothesis, the two modalities, gesture and speech

influence each other (bidirectional influence) and integration of gesture and speech is obligatory. In other words, speech influences the way gesture is perceived, gesture influences the way speech is perceived, and these together form an integrated semantic representation.

Previous work on gestures and speech interaction can be classified into two streams. One line of research examined primarily the role of gestures in communication with the focus on how gestures influence communication and cognition. For example, Louwerse and Bangerter (2010), in a face identification task, showed that ambiguous but informative pointing gestures facilitate target identification, even in the presence of clear verbal descriptions of target features. Gestures also help when other sources of information are ambiguous. Pine, Reeves, Howlett, and Fletcher (2013) showed participants degraded images of objects and animals on a screen, which gradually became clearer. Before seeing the pictures, participants made either a congruent gesture, or an incongruent gesture, or no gesture at all. Gesture-picture congruence was based on the resemblance of the gesture to the image. Results showed that congruent gestures lead to facilitation of naming (faster RTs) whereas incongruent gestures were detrimental in naming, and no gestures falling in-between. Similarly, Holler, Shovelton and Beattie (2009) compared information uptake about size and relative position about objects obtained from watching a person telling a story using either gesture only, speech only, or both. When both speech and gesture was present, participants obtained more information. Interestingly, gesture only and speech only presentation did not differ in conveying size and position information, suggesting that gestures can be as powerful as speech in certain contexts.

Most of the previous studies used deictic gestures, which are critical for both thinking and communicating spatial information (see, Alibali 2005, for a review). Allen (2003) examined the role of deictic gestures in verbal route directions by analyzing video clips of participants whose job was to give verbal route directions to different locations on campus.

He found that deictic gestures are indeed a useful source of information when communicating spatial information. They were more frequently used than iconic or beat gestures when giving verbal route directions, and the number of gestures and speech rate were positively correlated. Gestures can also convey information that is not present in speech. For example, an analysis of pain communication (Rowbotham, Holler, & Lloyd, 2012) showed that information about size and location of pain is represented better in gestures than speech. Overall, almost half of the information that was conveyed in pain talk was through gestures.

Gestures not only enhance comprehension, but also they affect the actions listeners take. Cook and Tanenhaus (2009) designed a naturalistic communication experiment in which a speaker participant first solved a Tower of Hanoi problem either with real objects or on a computer, and then described and explained it to a listener participant who then solved the problem on a computer. When explaining the problem to the listener-participant there was no difference in spoken information (frequency of particular words) between speakers who solved the problem with real objects and speakers who solved the problem on a computer. However, speakers who solved the problem with real objects showed more grasping gestures and more curved trajectories when explaining the problem to listener participants. More importantly, there was a positive relationship between the height of the speakers' gestures when explaining the problem and the height of the mouse trajectories of listeners' mouse movements when solving the problem later on computer, indicating that the way the speakers used gestures affects the actions the listeners engage in. The authors concluded that procedural information is conveyed in gesture but not speech and the procedural information extracted from gesture influences subsequent behavior of the listener.

Another line of research examined how gestures and speech interact with each other. Such studies compared information uptake or responding to speech and gestures that carry compatible or incompatible information. For example, Cassell, McNeill, and McCullough

(1998) asked participants to retell a narrative after they watched a video clip in which a person was narrating while at the same time making hand gestures, whose meanings were either matched or mismatched to the speech. While retelling the narrative, participants made significantly more mistakes (i.e. the content did not match the original spoken narrative) in gesture-speech mismatch phrases compared to gesture-speech match phrases, implying that listeners consider the gestural information when retelling the stories and mismatches produce more errors. In a similar study, Beattie and Sale (2012) showed that metaphoric gestures accompanying speech affect the interpretation of the message. Participants first watched a video in which a narrative included speech and hand gestures either matched or mismatched, and they made semantic judgments regarding the content. The results showed that gesture matches and mismatches with speech influenced participants' semantic judgments. For example, when participants saw a message like "I set my goals high" along with a gesture that a hand, palm facing down, moves up from stomach to the shoulder (match) or move horizontally out to the side (mismatch), those who watched the matched clip thought the person in the video clip seemed more driven to achieve goals than participants who watched the mismatch.

We need to draw attention to a point that requires clarification and more precise definition, of the mismatching and conflicting nature of *bimodally* presented information that is conveyed. Although conflicting speech and gesture can be considered as a gesture-speech mismatch, not all mismatches are necessarily conflicts. Information in one modality can be irrelevant to the information in the other modality, but not necessarily signaling a contrast (e.g. verbal description of a building accompanied by irrelevant hand movements). In addition, two modalities can convey mismatching information yet have the same goals (e.g. "saying very high" accompanied by gestures that indicate moderate levels of height). Mismatches can also be classified as strong or weak depending on how much information in

one modality resembles information in another modality (e.g. “chop” and “cut” as a weak mismatch, “chop” and “twist” as a strong mismatch; Kelly et al. 2010). On the other hand, information in gesture-speech *conflict* is contradictory in nature in the sense that they signal a clear contrast (e.g. saying “left” while pointing “right”), and information represented in both modalities cannot be combined in a meaningful way. Regarding gesture-speech conflicts, Langton, O’Malley, and Bruce (1996), in five reaction time (RT) experiments, showed that there is a symmetrical interference across deictic (directional) gestures and corresponding spoken and written words. In these experiments, participants watched static images of a person making a deictic gesture (right, left, top and bottom) and simultaneously listened to congruent or incongruent spoken directional words. The task was to respond to the information given in gesture or speech by pressing the corresponding button. Results showed that reaction times were faster on congruent trials (gesture and word matching) then incongruent trials both in responding to gesture trials and in responding to speech trials. They replicated the findings with written directional words instead of spoken words, and pointing arrows instead of pointing gestures, and verbal response instead of button pressing. They concluded that there is a symmetrical interference between speech and gesture and these two stimuli are analyzed in parallel and integrated into a single response, which is affected by relative discriminability of verbal and visual information. In subsequent research, Langton and Bruce (2000) showed participants pointing gestures (up or down) with spoken words (congruent or incongruent). This time, head movements were also manipulated so that the person in the picture oriented his head either up, down, or looked straight ahead. Reaction times were faster in gesture-speech congruent trials (e.g. “up” gesture with the word “up”) compared to gesture-incongruent trials (e.g. “down” gesture with the word “up”). The difference between incongruent and congruent trial RTs was largest when the head

movements were congruent with gestures. The authors concluded that head movements and eye gaze moderate the interaction of gesture and speech in communication.

Based on the literature reviewed above, several conclusions can be drawn. First, gestures are an important aspect of communication, and gestures accompanying speech almost always lead to better comprehension than only speech. Second, gestures can be as informative as speech, if not more so in some cases. This is especially true when the information is visio-spatial. Third, information received from gesture and speech are integrated to form a multimodal, unified semantic representation and this process is both automatic and involves control to some degree. Fourth, gesture-speech mismatches result in slower processing, and potentially more errors.

Although previous research revealed the importance of gestures and gesture-speech mismatches in communication, a few issues remained unclear. First, most studies focused on interference effects where participants were forced to attend to a particular modality and the second modality was used as a distractor. Although this method helps to understand how speech and gesture interact, it fails to recognize when and how a particular modality is a more valuable information source. Second, previous research almost exclusively relied on manual response (Langton et al. 1996; Langton et al. 2000; Kelly et al. 2007; Kelly et al. 2009; Kelly et al. 2010). Thus, the relative importance of response modality has not been considered. It remains unclear how people respond to bimodal information (congruent/incongruent) and specifically, which factors influence the interpretation of a conflicting bimodal message. Interpretation of a message is different from processing and perception because while processing of information is mostly automatic and emphasizes semantic integration, interpretation of a message is more about which information people choose to rely upon to use or convey information to another person. Interpretation, as we use it here, emphasizes what the person takes from the message, and how he or she makes a decision when the information

they received in different modalities conflict. Studies to date have examined how information in different modalities is perceived, and how semantic meanings of different modalities are integrated. That is, the question of how semantic integration is affected by congruent or incongruent stimuli has been the focus of interest. What remains unclear is how people interpret a given message, when the modality they have to use to pass on the message they received is controlled, and the information they received in two different modalities conflict. Therefore, our aim is to examine how input and output modalities influence the course of communication. Before we proceed with an overview of the current research, we turn to the issue of bimodal conflict in communication chains.

## Gesture and Speech: The Case of Bimodal Conflict in a Communication Chain

A communication chain consists of a source person, a recipient who has to instruct to a third person, and the different media he or she can use to convey the information. First, the recipient receives the information and has to interpret the message. In this phase, information can be presented only in speech (auditory), only in gesture (visual), or in both modalities. When information is presented in two modalities, gesture and speech can be conflicting or non-conflicting.

When information presented in different modalities conflict (e.g. saying “push” while producing “pull” gesture), the recipient has to make a choice about which information he or she will pass to another person or execute an action. There are several ways of dealing with a conflicting message. First, the recipient can attend to information presented in one modality and ignore the other. However, studies (e.g. Langton et al. 1996; Kelly et al. 2009) have shown that even when participants are instructed to ignore one modality, they were still influenced by information presented in that modality, and this process was primarily automatic. Second, the recipient can integrate information presented in different modalities

into a single meaningful representation (e.g. Cassell et al. 1999; Beattie & Sale, 2012). For example, if a speaker tells a recipient that he or she has "high" standards while producing a hand gesture that signals "low", the recipient can interpret this message as "medium". However, this is unlikely to occur if information in different modalities *conflict*. For example, receiving "push" instructions in speech while seeing "pull" instructions in gesture is unlikely to result in an averaged interpretation. Third, depending on relative strengths, information in one modality can drive the interpretation of the message. Here, a factor may be relative ambiguity of different modalities. For example, if a gesture is a general movement of hand instead of specific finger positions that is required, individuals may rely on information that is presented in speech. Similarly, if the spoken words are degraded, gestural information may drive the interpretation. Another factor may be relative informativeness of a particular modality compared to the other. Some abstract concepts are hard to express in gestures such as emotional states and come concrete objects are hard to express in speech such as the description of a complex building. Therefore, contextual factors that change the strength a particular modality may have significant influence on interpretation of the message.

It is also possible that the conflict is resolved at the second stage. Here, the recipient, who was previously presented information bimodally and has to instruct another person or take an action, needs to make a decision regarding which of the conflicting information she has to choose (Second phase of communication chain). These modalities can be visual (e.g. camera), auditory (e.g. microphone), or both. Under these conditions, we propose that which communication tool a person has access to may be a key factor in the decision-making process. We base our arguments on the situated nature of cognition and stimulus-response congruency research.

The idea that cognition emerges through interaction with environment, in other words, situated is not new. Information processing and actions will inevitably depend on the

interaction with tools that serve as a medium for communication. In other words, accessibility and availability of communication tools (visual and auditory devices) can be regarded as part of human communication capabilities. These tools extend cognition beyond human body and brain and may extend or limit his or her cognition regarding a message. Therefore, in a communication scenario, the person may perceive, process, and act according to the instruments that are available in the environment.

Availability of verbal or visual medium can influence relative accessibility of corresponding information and may result in "readiness" in that modality which may drive the decision. We argue that having to use a particular medium to convey information may initiate a process similar to priming effect, resulting in more accessibility to information that was previously received in the same modality. When the recipient has access to only auditory or only visual modality, then the information received in corresponding modalities (verbal and visual, respectively) may activate relevant representations. For example, being instructed to respond using microphone may activate representations on the previous speech that the person produced or listened. When responding, these previous experiences in the relevant modality can be more accessible and shape the response. Thus, in a bimodal information setting, the person may resolve the conflict by extracting information that was received in a primed modality.

Research on multimodal perception also supports this view by demonstrating that individuals tend to respond in a particular fashion if it the response option matches a previously presented stimulus, namely stimulus-response congruency effect (SRC; Kornblum et al. 1990). In studies examining SRC, participants are shown a set of multimodal stimuli (e.g. dots with an irrelevant sound) and their response options (e.g. key positions) either correspond to or differ from stimuli in terms of spatial location. When the stimulus shown in the visual modality matches the modality (spatial position) of the response options,

individuals tend to use the matching response option and they are faster in responding. Although SRC research is mainly concerned with perception of arbitrary stimuli (e.g. dots) in a spatial context, the effects of modality match may still be relevant in multimodal communication. If the recipient has to response in a particular modality, then the information presented in the corresponding modality may be easier to access, similar to the SRC effect.

Both situated cognition approach and the work on SRC suggest that when a bimodal input is received, having access to different output options (e.g. microphone, camera) may result in different flow of information. Specifically, we expect that if an auditory medium has to be used, information in auditory modality (spoken information) is more likely to be conveyed; if a visual medium is available, information in visual modality (gestural information) is more likely to be conveyed.

Studies that use match-mismatch paradigms consistently showed that gesture-speech congruence leads to superior performance compared to gesture-speech incongruence, demonstrated in faster processing and fewer errors (e.g. Langton et al. 1996; Langton et al. 2000; Cassell et al. 1998; Galati & Samuel 2011; Chieffi, Secchi, & Gentilucci, 2009; Kelly et al. 2007; Kelly et al. 2009). Accordingly, we also expect that gesture and speech matches will result in faster responses compared to gesture speech mismatches.

**The Current Study**

The question we investigate here is what happens when a person receives conflicting information from two modalities and the modality in which they have to pass on information is controlled? So, if the visual message (gesture) depicts a right turn and the auditory message (verbal instruction) states a left turn with the communicator having to pass on information to a third person verbally (or manually): what does the communicator do? To this end, we control for the modality in which they are expected to pass the information while presenting bimodal

information that is equally salient but conflicting. This will allow us to examine how the modality they have to use influences their choice of the bimodally received conflicting information. In other words, do they choose the information from the auditory modality of the conflicting information when they have to communicate auditorily or do they choose the information from the visual modality? The same dilemma and its resolution exist in the case of having to pass on the information they receive when they have to communicate visually (manual execution). These are issues that become relevant in the type of situation that we depicted in the very introduction to this research. What information do I use when I receive conflicting information coming from an auditory source (control tower) and onboard instruments (visual information)? What is the modality of the incoming information that I rely on when the modality in which I have to change the flight course can be performed manually or auditorily?

A further issue that can be raised about earlier studies is the frame of reference for the dependent variable, namely how is error assessed in cases of mismatches or conflict between auditory and visual information? In previous studies, participants, after being exposed to congruent or incongruent information, were asked to respond to the auditory information (verbal instructions) they listened. Therefore, an error is defined as failure to respond correctly to information presented in acoustic modality. Since in this study verbal and visual stimuli will be equally important for participants, there were no errors, only choices.

## Method

### Participants

100 students (mean age = 21.73; SD=3.18; 50 women) from a Dutch university participated as paid volunteers (€5) in this study. Gender distribution was 43.7% females in

the manual condition and 56.2% in the vocal condition. 84.5% of the participants were right-handed and 92.8% were native Dutch speakers.

**Design**

The study used a 2x2x2x2 mixed design. Congruency (congruent, incongruent) and gender of the person giving instructions (female vs. male) were the within participants variables. Response type (manual, vocal) and gender of participant (female vs. male) were the between participants variables. Since both gender variables did not produce any effects, they were collapsed and are not reported any further in the analyses below.

Half of the participants responded their choice of direction by pressing the appropriate key on a response box (manual condition), and half of the participants responded by giving their answers aloud (vocal condition). Dependent variables were reaction times and actual responses.

**Materials**

The experiment consisted of 100 experimental trials and a questionnaire. Of the 100 experimental trials, 30 were congruent, 30 were incongruent, 20 were forward, 10 were gesture-only and 10 were voice-only trials. The trials were presented in a random order. All video clips had a male and female version. Half of the participants were presented with the male versions, and half of the participants were presented with the female versions.

Each trial started with a white fixation cross on a black background for 1000 ms, followed by a 1600 ms long video clip. The video clip presented a continuous hand gesture, a spoken word, or both. When a gesture was presented, it lasted for the whole duration of the video clip. Spoken words were 600ms long and when presented with a gesture the onset of spoken words coincided with the onset of gesture stimulus. Participants could give their

responses when the video started playing. The video stopped playing when the participants pressed a key (manual condition) or when 1600 ms passed (vocal condition). After the video clip, the next trial started.

There were five different types of video clips, namely congruent, incongruent, forward, gesture-only and voice-only. Congruent and incongruent video clips consisted of a hand gesture indicating left or right along with a spoken word *left* or *right* (Figure 1). The direction indicated by gesture and voice were the same for congruent video clips and different for incongruent video clips. Forward video clips consisted of a hand producing forward movement gesture with an accompanying spoken word *forward* (Figure 2). Forward video clips were used as fillers. Gesture-only videos presented a hand producing left or right gesture without sound. Voice-only videos presented a black background with a spoken word *left* or *right*. Gesture-only and voice-only trials were used to ensure that participants attend to both visual and auditory information during the experiment. Forward, gesture-only and voice-only trials were not analyzed.

Before the end of the experiment, participants received a set of questions asking for demographic information, and included open-ended questions about the aim of the study, whether they recognized conflicts, and whether they used a particular strategy. Additionally, two five-point scales were used to assess how difficult they found the experiment and how difficult it was to make a decision on incongruent trials (1-very easy to 5-very difficult). Their handedness was also assessed.

*Figure 1.* “Right” gesture.

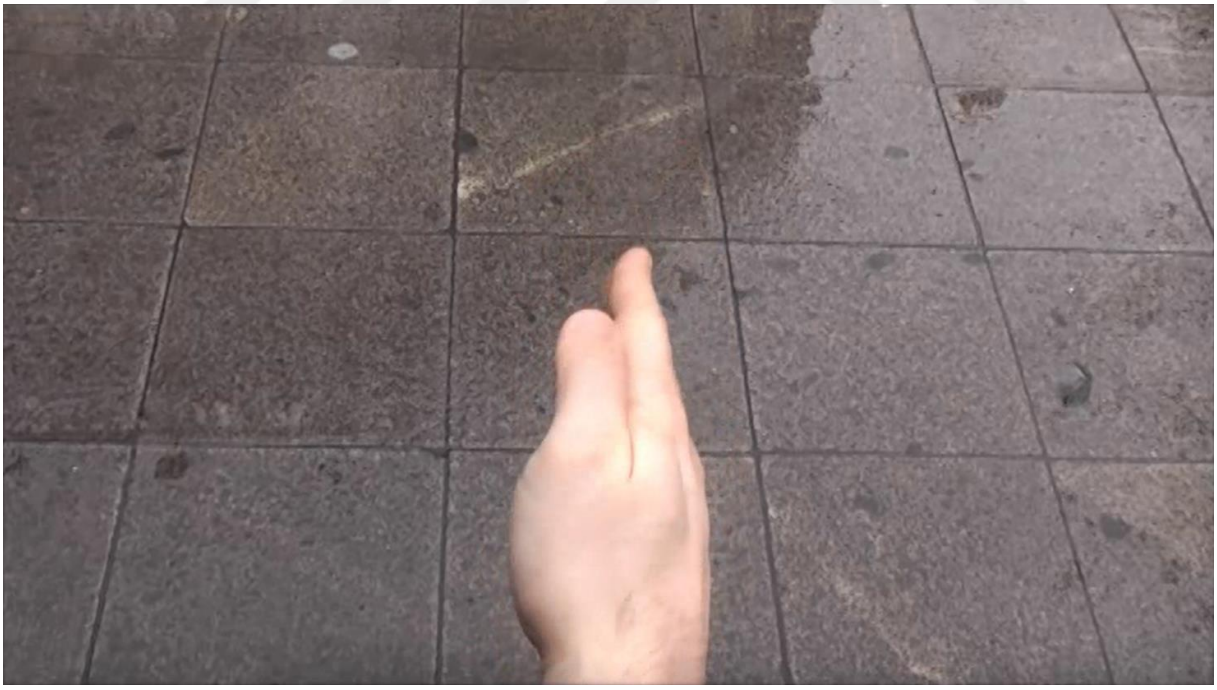

*Figure 2.* “Forward” gesture.

**Procedure**

Participants completed the study on a computer. Before the study, they were told that the study was about a hypothetical navigation scenario and their job was to identify the direction presented in each instruction. In the manual condition, they were asked to respond by pressing the appropriate key on the response box. In the vocal condition, they were asked to respond by giving their answers aloud. A microphone attached to the computer was used in the vocal condition. They were told to give their answers as fast as possible. They were also informed that during the experiment, there may sometimes be a discrepancy between what they see and what they hear.

First, they completed 10 practice trials. These trials were similar to congruent trials except they were longer (2000 ms) and included a feedback screen at the end of each trial. On the feedback screen, when participants were slow, they were informed that the time to give an answer had passed and they should be faster if they did not give an answer during the video clip. After the practice trials, participants completed the experiment and then filled out the questions on the computer. Then they were debriefed and thanked. The experiment took approximately 15 minutes to complete.

Reaction time and response data in manual condition were collected by E-Prime software. Responses in vocal condition were recorded by E-Prime and stored as a sound file. RT data was extracted using SayWhen software (Jansen & Watter, 2008), which analyzes a sound file and detects the speech onset when the volume level reaches a certain threshold, similar to a voice key hardware.

# Results

## Data Preparation

The data of 3 participants were removed due to technical problems in the recording. The data of the remaining 97 participants were used for the analyses. Only the 60 critical trials of each participant (30 congruent, 30 incongruent) were used in the analyses. Out of 5820 total trials, 27 (0.46%) were removed due to missing data.

Before analyzing the data, the reaction time data was log (natural) transformed in order to normalize the distribution of the data. Trials that had a reaction time above or below 2.5 standard deviation of the participant's median RT were eliminated. Trials shorter than 300 ms and longer than 1500 ms were also excluded. This process resulted in elimination of 3.19% of the trials. Of the eliminated data, 36.76% were congruent and 63.24% were incongruent trials.

Next, the percentage of errors in congruent trials was calculated. Error was defined as responses that were different from congruent gesture and speech instances. Overall, 0.67% of all congruent trials were errors with 15 participants making at least 1 error in congruent trials. 2 participants made errors in less than 11% of the congruent trials and the remaining 13 participants made errors in less than 5% of the congruent trials.

Finally, mean reaction time scores for congruent and incongruent trials for each participant were calculated.

## Analyses

Participant gender and video clip gender had no effect on any variable, therefore were excluded from the analyses and will not be discussed further.

**Reaction Speed as a Function of Information Congruence-Incongruence.** One of our predictions was that the participants will be slower on incongruent trials than congruent trials. The analysis confirmed our predictions. A 2x2 repeated measures ANOVA with congruency as the repeated factor (congruent vs. incongruent) and experimental condition as the between subjects factor (manual vs. vocal) yielded a significant main effect for congruency, $F(1, 95) = 97.54$, $p = .000$, $\eta^2 = .507$. As can be seen in Table 1, the participants were faster in congruent trials ($M = 674.78$ ms, $SD = 220.53$ ms) than in incongruent trials ($M = 734.04$ ms, $SD = 244.99$ ms).

**Reaction Speed as a Function of Response Medium.** There was also a main effect of condition, $F(1, 95) = 239.87$, $p = .000$ , $\eta^2 = .716$. Participants in the manual condition ($M = 510.41$ ms, $SD = 119.07$ ms) were faster than participants in the vocal condition ($M = 901.66$ ms, $SD = 125.09$ ms). Condition x Congruency interaction was not significant, $F(1,95) = .125$, ns. These results confirmed our expectations that both in manual and in vocal conditions, participants were slower in incongruent trials than in congruent trials. We should note that the RT differences between response mediums should be interpreted with caution because we had limited control over the devices that recorded manual and visual data. Moreover, the main dependent variable in this study was not RT, but rather the actual responses, which we will discuss next.

*Table 1. Descriptive Statistics for Reaction Times (ms)*

| Experimental Condition | Congruency | | | | | |
|---|---|---|---|---|---|---|
| | Congruent | | Incongruent | | Overall | |
| | M | SD | M | SD | M | SD |
| Manual | 488,20 | 101,76 | 533,32 | 140,80 | 510.41 | 119.07 |
| Vocal | 865,26 | 124,14 | 938,95 | 132,14 | 901.66 | 125.09 |
| Overall | 674.78 | 220.53 | 734.04 | 244.99 | | |

**Disambiguation of Incongruent Trials**. To examine how incongruent clips (visual/manual versus auditory) were disambiguated as a function of the response medium (manual versus vocal) a response index per participant was computed across all incongruent stimulus conditions. This resulted in a visual response index and an auditory response index. A response is considered as *a visual response* if it matched the visual information presented in the video. A response is considered as *an auditory response* if it matched the auditory information presented in the video. These labels should not be confused with the manual and vocal response modalities. Visual and auditory responses refer to the match between the response (e.g. left) and the visual or auditory information presented in the video clips, *regardless* of the response modality. Therefore, participants in the manual condition could have auditory responses, and participants in the vocal condition could have visual responses. The percentage of visual responses of a participant constituted his or her visual response index and the percentage of auditory responses of a participant constituted his or her auditory response index. For each participant, the sum of visual response index and auditory response index was 100; therefore only the visual response index was used in the following analyses. We predicted that participants in the manual condition would have more visual responses than participants in the vocal condition, and subsequently, participants in the vocal condition will have more auditory responses than participants in the manual condition. Since the distribution of percentages was not normal, we used the Mann-Whitney U test to compare the experimental conditions.

Overall median visual response index was 96.3%. Regardless of condition, participants followed visual information substantially more frequently than auditory information. For the manual condition, the median visual response index was 93.61% and for vocal condition, it was 96.61%. Contrary to our prediction, participants in the vocal condition had slightly more visual responses than the participants in the manual condition. This was

confirmed by a Mann-Whitney test, indicating a significant difference between manual and vocal conditions, $U = 867$, $p = .012$, $r = .25$. Participants in the manual condition had an average rank of 42.08, while participants in the vocal condition had an average rank of 56.06.

To elaborate the above findings, we looked at which modality participants found more informative during the study. 79.6% of the participants in the manual condition found visual modality more informative, and 89.6% of the participants in the vocal condition found visual modality more informative, similar to their response choices. However, a Chi-Square test yielded no significant difference between two conditions, $\chi^2(1, N = 97) = 1.85$, $p > .05$. Although participants in the vocal condition gave more visual responses than participants in the manual condition, the visual modality was found to be as informative for the participants in that condition as the participants in the manual condition.

To examine whether there was a speed benefit of relying on a particular modality, a 2x2 repeated-measures ANOVA with experimental condition as the between subjects factor and response type (visual response, auditory response) as the within subjects factor was conducted. There was a significant main effect of condition, $F(1, 59) = 86.45$, $p = .000$, $\eta 2 = .594$, with participants in the manual condition ($M = 569.68$ ms, $SD = 145.26$ ms) being faster than participants in the vocal condition ($M = 940.45$ ms, $SD = 131.6$ ms), similar to the overall RT results, which should be interpreted with caution, as we have mentioned before. There was also a significant main effect of response type, $F(1, 59) = 14.853$, $p = .000$, $\eta 2 = .201$. Participants were faster when giving visual responses ($M = 716.12$ ms, $SD = 243.01$ ms) than auditory responses ($M = 776.35$ ms, $SD = 235.59$ ms). More importantly, there was a significant Condition x Response type interaction, $F(1, 59) = 9.21$, $p = .004$, $\eta 2 = .135$. Simple effects analysis revealed that in manual condition, participants were faster when giving visual responses ($M = 538.88$ ms, $SD = 143.57$ ms) than auditory responses ($M = 637.62$ ms, $SD = 167.27$ ms), $p = .000$. However, in vocal condition, participants were as fast when giving

visual responses as auditory responses, $p>.05$. Relying on visual modality provided a speed benefit in the manual condition, but not in the vocal condition.

**Task Difficulty.** An independent samples t-test yielded a difference in perceived difficulty of the experiment between two experimental conditions, $t(95) = -2.95$, $p = .004$. Participants in the manual condition ($M = 2.45$, $SD = .65$) found the experiment more difficult than participants in the vocal condition ($M = 2.04$, $SD = .71$), although the scores suggest that both groups found the experiment between easy to medium on the scale. Perceived difficulty of the decisions in incongruent trials followed a similar pattern. An independent samples t-test showed a significant difference in perceived decision difficulty between conditions, $t(95) = -3.074$, $p = .003$. Participants in the manual condition ($M= 2.96$, $SD =.97$) found making decisions in incongruent trials more difficult than participants in the vocal condition ($M = 2.38$, $SD = .91$).

Overall, the data supported our predictions about reaction time differences between congruent and incongruent trials, however failed to support our predictions regarding the nature of responses. Next, we will discuss the findings, and address potential factors that may have led to current results.

## Discussion

In this study, our aim was to examine how people react to conflicting bimodal information and whether having access to different output options (manual and vocal) affects interpretation of this information. We examined the particular case of speech and gesture conflicts where we predicted that conflicts would results in slower responses, and the medium used to respond would shape responses that convey information in the corresponding modality. Our former prediction was supported; however our latter prediction was not supported.

**Reaction Speed Differences**

Reaction time data showed that, participants, regardless of the response medium, were slower when incongruent information was presented than when congruent information was presented. These results are in line with the previous literature on gesture-speech congruency showing that people attend to both visual and auditory information (Cassell et al. 1998), and processing takes longer when they do conflict (Langton et al. 1996; Langton et al. 2000; Kelly et al. 2007; Kelly et al. 2009). Remember that participants in this study were not forced to attend to or ignore a particular modality, although they could develop a strategy of focusing on a particular modality. Also, the majority of previous research (Langton et al. 1996; Langton et al. 2000; Kelly et al. 2007; Kelly et al. 2009; Kelly et al. 2010) used manual response as the only response option; this study demonstrated that the effect also occurs when responding verbally.

Manual responding produced faster reactions than vocal responding, however, as we have mentioned, these two response types may not be comparable from a technical point of view. The hardware equipment to record the two types of responses was based on very different mechanisms. A response box converts mechanical signals whereas a microphone converts auditory signals into digital data. Since we did not have control over these processes, reaction time comparisons should be taken with caution. Also, the method we utilized could not differentiate whether these differences are due to having different response media or due to differences in finger movements and speech production.

**Differences in Responses**

Having access to different response media had an influence, but not in the direction we predicted. Our argument was that the medium participants had to use would prime the relevant modality and consequently, the preference of the information in the corresponding

modality would increase. Results showed that the majority of participants preferred visual information over auditory information. More importantly, this preference was greater in the vocal condition than manual condition. A possible explanation might be the qualitatively different response situations across the experimental conditions. The task in the manual condition resembled a real world task better than the task in the vocal condition. Pressing a key, as opposed to saying the direction aloud, has two distinct features. First, the response creates a physical change in the environment where the participants actually push a button. Second, pressing the key produces auditory and tactile feedback. Thus, there is an identification element and an execution element present in the manual condition. In the vocal condition however, only the identification element is present and identification of auditory stimulus was simply a repetition, which participants may have found as not the real task or even a task at all, therefore ignoring auditory information completely. The fact that perceived difficulty was greater in manual condition implies a qualitative difference across response options. This can also be seen a limitation for the study, because in the manual response condition participants engaged in a more concrete task where in the vocal condition, they simply vocalized what they see or hear. A more similar task execution would be to design a system, where voice initiates a physical change in the environment, creating an experience similar to pressing the keys. However, even if this would provide a clearer picture, we would not expect a dramatic shift towards a preference for auditory information, which we will discuss next.

Although we had no specific prediction, the very large percentages of visual responses deserve more attention. Data showed that the information presented via gestures was the dominant response across experimental conditions. This supports the idea that gesture as an information source may be much more important than speech in certain contexts. A recent study (Fay, Lister, Ellison, & Goldin-Meadow, 2014) demonstrated that gestures are much

more effective than vocalization in creating a communication system within dyads also supports this idea. Previous studies used gestures mostly as an accompanying feature of speech (e.g. Cassell et al. 1999; Beattie & Shovelton, 1999b; Beattie& Sale 2012) or forced participants to attend to a particular modality (Langton et al. 1996; Langton et al. 2000); hence it was difficult to identify the value of gesture as an information source. This study shows that gestures can be very powerful in decision-making in multimodal communication and can dominate the information flow. One possible explanation for these findings is that gestures, in a spatial context, entail direct mapping between the gesture and the intended location. Showing a direction, as opposed to using words to describe it, is a more direct way of transmitting the relevant information, which might have increased participants' reliance on gestures as the correct source of information. Another possible explanation is that gestures used in this study are universal, whereas speech is language bound, and this might have lead participants to consider gestures as more convenient, especially if they based their decisions on which modality would be the most meaningful in daily life. This particular case is also reflected in settings where an individual is able to give directions to another person despite the possibility that they speak different languages.

The strong preference for gestures indicates that the content and the nature of task seem to be more important in driving decisions than the available tools. The match between input modality and output modality may still be important however, as following visual rather than auditory information provided a speed benefit in the manual condition, suggesting that stimulus and response modality match may improve performance. Manual responding might have increased participants' processing speed of information in the visual modality, resulting in faster responses. The reverse effect was absent in the vocal condition. Given the task related issues discussed above, it is difficult to draw conclusions on whether input-output modality match provides speed benefit in the vocal condition.

**Limitations and Future Research**

As we have mentioned, task execution in this study was limited to key pressing (manual condition) or vocalizing (vocal condition) and we did not used measures to ensure that these two options provide the same experience of task execution with only difference being the modality. This limitation also points to an interesting avenue for multimodality research where the execution of tasks in particular modalities and whether they are on the same level in terms of subjective experience should be considered. Also, it may be inappropriate to generalize the findings from this study to all audio-visual conflicts because (1) we employed a realistic scenario where we did not have strict control on the stimulus (e.g. precise movement directions, movement speed, and the frequency of the voice signals) and (2) we only examined deictic gestures in a visual-spatial task environment. Therefore, findings from this study should be enjoyed with caution, as the different task domains (e.g. abstract concepts, storytelling) and other types of gestures (e.g. iconic) may provide a richer environment and more complex speech-gesture interaction.

The fact that some participants had 100% visual responses whereas others had 100% auditory responses suggests that there may be individual differences when dealing with multimodal information. Some people tend to categorize themselves as a visual person or an auditory person, which was also reflected in participants' comments. Future studies should include assessment of individual differences, particularly the attitudes towards different modalities.

**Conclusion**

By using a gesture-speech congruency paradigm, we demonstrated that visual cues are regarded as a more powerful information source than auditory cues in a visio-spatial context and the medium plays a role in multimodal communication. We believe the research on multimodal communication can benefit from these findings by re-conceptualizing the role of

gestures in communication, and by giving more attention to how different modalities are reflected in an experimental setting and to what extend different modalities are comparable in terms of participant experience. Also, designers of multimodal communication technology can make more informed decision on how to include multimodal information within a communication system.

References

Allen, G. L. (2003). Gestures Accompanying Verbal Route Directions: Do They Point to a New Avenue for Examining Spatial Representations? *Spatial Cognition & Computation: An Interdisciplinary Journal, 3(4),* 259-268 .

Alibali, M. W. (2005). Gesture in Spatial Cognition: Expressing, Communicating, and Thinking About Spatial Information, *Spatial Cognition & Computation: An Interdisciplinary Journal, 5(4),* 307-331.

Bavelas, J., Kenwood, C., Johnson, T., & Philips, B. (2002). An Experimental Study of When and How Speakers Use Gestures to Communicate, *Gesture,* 2(1) 1-17.

Beattie, G., & Sale, L. (2012). Do metaphoric gestures influence how a message is perceived? The effects of metaphoric gesture-speech matches and mismatches on semantic communication and social judgment.*Semiotica, 192,* 77-98.

Beattie, G., &Shovelton, H. (1999b). Mapping the range of information contained in the iconic hand gestures that accompany spontaneous speech. *J Lang Soc Psychol. 18(4),* 438- 462.

Cassell, J., McNeill, D., & McCullough, K. E. (1999). Speech-Gesture Mismatches: Evidence for One Underlying Representation of Linguistic and Nonlinguistic Information. *Pragmatics and Cognition, 7(1),* 1–33.

Chieffi, S., Secchi, C., & Gentilucci, M. (2009). Deictic word and gesture production: Their interaction. *Behav Brain Res, 203,* 200–206.

Cook, S.W., Tanenhaus, M.K. (2009). Embodied communication: Speakers' gestures affect listeners' actions. *Cognition, 113,* 98–104.

Fay N, Lister CJ, Ellison TM, Goldin-Meadow S. (2014). Creating a communication system from scratch: gesture beats vocalization hands down. *Front. Psychol. 5,* 354.

Galati, A., & Samuel, A. G. (2011). The role of speech-gesture congruency and delay in remembering action events, *Language and Cognitive Processes*, *26(3),*406-436.

Goldin-Meadow, S. (1999).The role of gesture in communication and thinking.*Trends in Cognitive Sciences*, 3, 419–429

Goldin-Meadow, S., &Alibali, M.W. (2012).Gesture's role in speaking, learning, and creating language.*Annu. Rev. Psychol*, *64(1),* 257.

Goodwin, C. (1995). Seeing in depth. *Social Studies of Science 25,* 237–274.

Goodwin, C., & Goodwin, M. H. (1996). Seeing as situated activity: formulating planes. In: Engestrom, Yrjo, Middleton, David (Eds.), *Cognition and Communication at Work.* Cambridge University Press, Cambridge, pp. 61–95.

Heath, C., & Luff, P. (2000). *Technology in Action*.Cambridge University Press, Cambridge.

Holler, J., Shovelton, H., & Beattie, G. (2009). Do iconic hand gestures really contribute to the communication of semantic information in a face-to-face context? *Journal of Nonverbal Behavior*, *33(2),* 73–88.

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

Hostetter A. B. (2011) When do gestures communicate? A metaanalysis. *Psychol Bull*, *137(2),* 297–315.

Hostetter, A. B., &Alibali, M. W. (2008). Visible embodiment: gestures as simulated action. *Psychon. Bull. Rev.15,* 495–514 .

Hostetter, A. B.,&Alibali, M. W. (2010). Language, gesture, action! A test of the Gesture as Simulated Action framework.*J. Mem. Lang*. *63,* 245–57.

Hostetter, A., Alibali M., & Kita, S. (2007). I see it in my hands' eye: Representational gestures reflect conceptual demands, *Language and Cogn. Processes 22,* 313-336.

Jansen, P., and Watter, S. (2008) SayWhen: An automated method for high-accuracy speech onset detection. *Behavior Research Methods, 40(3),* 744-751.

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.

Kelly, S. D., Creigh, P., Bartolotti, J. (2009). Integrating speech and iconic gestures in a stroop-like task: Evidence for automatic processing. *Journal of Cognitive Neuroscience, 22(4),* 683–694.

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., & Lee, A. L. (2011): When actions speak too much louder than words: Hand gestures disrupt word learning when phonetic demands are high. *Language and Cognitive Processes,* DOI:10.1080/01690965.2011.581125

Kita S., &Ozyurek A. (2003). What does cross-linguistic variation in semantic coordination of speech and gesture reveal? Evidence for an interface representation of spatial thinking and speaking. *J. Mem. Lang*. 48, 16–32.

Kornblum, S., Hasbroucq, T., & Osman, A. (1990). Dimensional overlap: Cognitive basis for stimulus-response compatibility: a model and taxonomy. *Psychological Review, 97,* 253-270.

Krauss, R. M. (1998). Why do we gesture when we speak? *Curr. Dir. Psychol. Sci. 7,* 54–60.

Krivonos, P. D. (2007). "Communication in Aviation in Safety: Lessons Learned and Lessons Required," *Regional Semi-nar of the Australia and New Zealand Societies of Air Safety Investigators,* 1-35.

Langton, S.R.H., O'Malley,C., & Bruce, V. (1996). Actions speak louder than words: Symmetrical cross-modal interference effects in the processing of verbal and gestural

information. *Journal of Experimental Psychology: Human Perception and Performance, 22,* 1357–1375.

Langton, S. R. H., & Bruce, V. (2000). You must see the point: automatic processing of cues to the direction of social attention. *J. Exp. Psychol. Hum. Percep. Perf,* 26, 747–757.

Leonard, M., Graham, S., &Bonacum, D. (2004). The human factor: the critical importance of effective teamwork and communication in providing safe care. *QualSaf Health Care*, *13,* 85–90.

Louwerse, M. M., &Bangerter, A. (2010). Effects of ambiguous gestures and language on the time course of reference resolution.*Cognitive Science, 34,* 1517-1529.

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

Pine, K. J., Reeves, L., Howlett, N., & Fletcher, B. (2013). Giving cognition a helping hand: The effect of congruent gestures on object name retrieval. *British Journal of Psychology, 104,* 57–68.

Rowbotham, S., Holler, J., Lloyd, D. (2012). How Do We Communicate About Pain? A Systematic Analysis of the Semantic Contribution of Co-speech Gestures in Pain-focused Conversations.*Journal of Nonverbal Behavior, 36,* 1-21.

Ozyurek 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. *J. Cogn Neurosc, 19,* 605–16.

Willems, R. M., Ozyurek, A., &Hagoort, P. (2006). When language meets action: The neural integration of gesture and speech. *Cerebral Cortex*. *17,* 2322—2333.

Willems R.M., &Hagoort P. (2007). Neural evidence for the interplay between language, gesture, and action: A review. *Brain Language, 101,* 278–289.