



# Interaction of sound and sight during action perception: Evidence for shared modality-dependent action representations

Kaat Alaerts, Stephan P. Swinnen, Nicole Wenderoth\*

Motor Control Laboratory, Research Center for Movement Control and Neuroplasticity, Department of Biomedical Kinesiology, Group Biomedical Sciences, Katholieke Universiteit Leuven, Tervuursevest 101 B-3001 Heverlee, Belgium

## ARTICLE INFO

### Article history:

Received 20 January 2009

Received in revised form 6 April 2009

Accepted 11 May 2009

Available online 20 May 2009

### Keywords:

Mirror neurons

Action observation

Action sound listening

Transcranial magnetic stimulation

Primary motor cortex

## ABSTRACT

Seeing or hearing manual actions activates the mirror neuron system, i.e., specialized neurons within motor areas which fire not only when an action is performed but also when it is passively perceived. Although it has been shown that mirror neurons respond to either action-specific vision or sound, it remains a topic of debate whether and how vision and sound interact during action perception.

Here we used transcranial magnetic stimulation to explore multimodal interactions in the human motor system, namely at the level of the primary motor cortex (M1). Corticomotor excitability in M1 was measured while subjects perceived unimodal visual (V), unimodal auditory (A), or multimodal (V+A) stimuli of a simple hand action. In addition, incongruent multimodal stimuli were included, in which incongruent vision or sound was presented simultaneously with the auditory or visual action stimulus. A selective response increase was observed to the congruent multimodal stimulus as compared to the unimodal and incongruent multimodal stimuli.

These findings speak in favour of 'shared' action representations in the human motor system that are evoked in a 'modality-dependent' way, i.e., they are elicited most robustly by the simultaneous presentation of congruent auditory and visual stimuli. Multimodality in the perception of hand movements bears functional similarities to speech perception, suggesting that multimodal convergence is a generic feature of the mirror system which applies to action perception in general.

© 2009 Elsevier Ltd. All rights reserved.

## 1. Introduction

Mirror neurons discharge not only when an action is performed but also when the same action is observed (Di Pellegrino, Fadiga, Fogassi, Gallese, & Rizzolatti, 1992). This mechanism is thought to play an important role in action recognition by matching visual representations of observed actions to motor plans (Rizzolatti & Craighero, 2004). Only recently, Kohler et al. (2002) reported that mirror neurons in monkey's inferior frontal cortex (region F5), discharge also when listening to action-related sounds, such as breaking of a peanut. Similarly in humans, activity of the motor system is modulated in response to both visual and acoustic action-related stimuli (Aziz-Zadeh, Iacoboni, Zaidel, Wilson, & Mazziotta, 2004; Fadiga, Fogassi, Pavesi, & Rizzolatti, 1995; Gazzola, Aziz-Zadeh, & Keysers, 2006; Iacoboni, 2005). However, it remains a topic of debate to which extent vision and sound of a perceived action may interact to retrieve/activate the corresponding motor plan in the observer's motor system. At the cell level, it was shown that

half of monkeys' F5 mirror neurons (11 out of 22) responds equally to either multi- or unimodal stimuli, indicating that activation of these neurons is 'modality-independent' (Keysers et al., 2003). This lead to the notion that perception-induced activation of motor plans may obey a 'whole-or-nothing' principle, e.g., perception of 'tearing paper' evokes the action event 'tearing paper', irrespective of whether it is heard, seen or both heard and seen ('modality-independent' action representations). However, the other half of the explored F5 neurons (8 out of 22) exhibited more vigorous responses to congruent audio-visual stimuli, such that the multimodal response was roughly equal to the sum of the unimodal responses (Keysers et al., 2003). As such, the presence of these neurons suggested an alternative mechanism of perception-induced 'action retrieval' that benefits from the simultaneous input of vision and sound describing the same movement (shared 'modality-dependent' action representations).

At the systems level, audio-visual interactions during action perception have only rarely been addressed experimentally. A functional imaging study by Kaplan and Iacoboni (2007), showed that a region of the ventral premotor cortex responded specifically to the combination of visual and auditory action-related stimuli, however, no true foci of multisensory convergence were demonstrated (Kaplan & Iacoboni, 2007).

\* Corresponding author. Tel.: +32 16 32 91 57; fax: +32 16 32 91 97.

E-mail addresses: [Nici.Wenderoth@faber.kuleuven.be](mailto:Nici.Wenderoth@faber.kuleuven.be),  
[Nicole.Wenderoth@faber.kuleuven.be](mailto:Nicole.Wenderoth@faber.kuleuven.be) (N. Wenderoth).

Here we used transcranial magnetic stimulation (TMS) to explore responses of the human primary motor cortex (M1) to visual and auditory stimuli with particular interest in the extent to which the two modalities interact. Corticomotor excitability in M1 was measured while subjects perceived a unimodal visual (V), a unimodal auditory (A), or a multimodal (V+A) stimulus of a simple hand action. In addition, two incongruent multimodal stimuli were included, in which incongruent vision or sound was presented simultaneously to the auditory or visual action stimulus.

Different hypotheses on the nature of action representations within the motor system lead to different predictions for this experiment. First, actions may be represented in a 'modality-independent way', such that seeing, hearing, or both seeing and hearing an action event, are equally salient in eliciting activity in the corresponding action representation. Following this hypothesis, excitability responses in M1 are expected to be comparable during the presentation of unimodal (either auditory or visual) or multimodal stimuli (either congruent or incongruent). On the other hand, movement perception might activate 'shared modality-dependent action representations', which are evoked more robustly from congruent visual and auditory input about the corresponding action. This hypothesis predicts a selective response increase to congruent multimodal audio-visual stimuli as compared to unimodal and incongruent multimodal stimuli.

## 2. Methods

### 2.1. Preliminary work: measurements of muscle activity during action execution

#### 2.1.1. Subjects

Ten subjects (age range 23–30; 9 females, 1 male) participated in the preliminary experiment. All participants were right-handed, as assessed with the Edinburgh Handedness Questionnaire (Oldfield, 1971) and were naive about the purpose of the experiment for which written informed consent was obtained.

#### 2.1.2. Task

Participants observed a video showing how a plastic bottle is slowly crushed by an actor's right hand (Fig. 1A) and were instructed to actually perform the same action in synchrony with the video.

### 2.1.3. EMG

During execution, a surface electromyogram (EMG) was simultaneously recorded from the right Opponens Pollicis (OP) thumb muscle and wrist Flexor (FCR) and Extensor Carpi Radialis (ECR) muscles by means of Ag–AgCl surface electrodes (Blue Sensor SP) placed over the muscle belly and aligned with the longitudinal axis of the muscle.

### 2.1.4. Data analysis

Each subject performed the action 16 times. In four additional trials, the EMG was recorded during maximal voluntary contraction (MVC) of each muscle. EMG changes (amplitudes) were calculated for a short time-interval of 40 ms during the actual crushing of the drinking bottle (Fig. 1A). EMG changes were expressed as the percentage of subjects' muscle-specific MVC-scores.

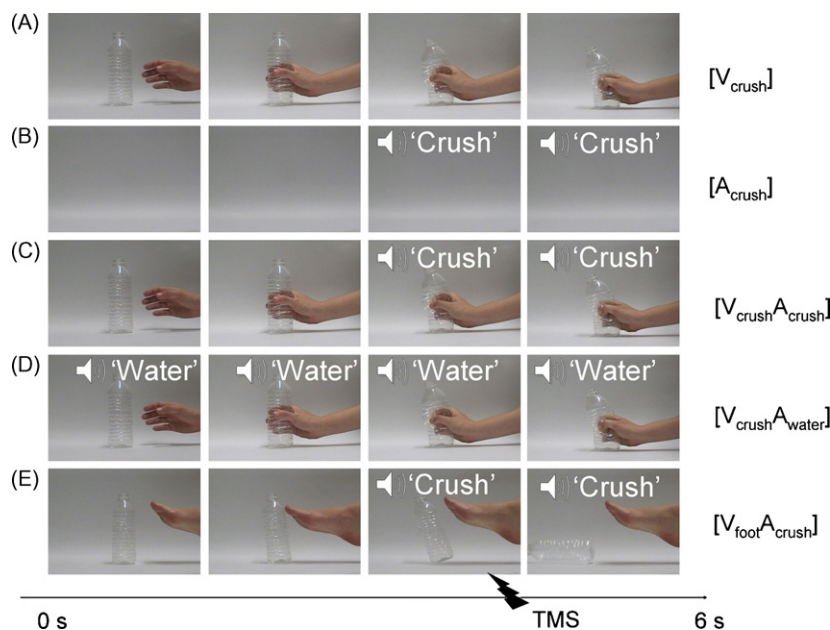
## 2.2. Main study: excitability measurements in primary motor cortex during action perception

### 2.2.1. Subjects

Thirteen subjects (age range 22–35; 8 females, 5 male) participated in the action perception experiment. All participants were right-handed, as assessed with the Edinburgh Handedness Questionnaire (Oldfield, 1971) and were naive about the purpose of the experiment. Written informed consent was obtained before the experiment and participants were screened for potential risk of adverse effects during TMS. The experimental procedure as well as the informed consent were approved by the local Ethics Committee for Biomedical Research at the Katholieke Universiteit Leuven in accordance to The Code of Ethics of the World Medical Association (Declaration of Helsinki) (Rickham, 1964).

### 2.2.2. EMG and TMS

Based on the muscle activity pattern during active crushing, the OP and ECR muscles were chosen to measure corticomotor excitability during action perception. To do so, EMG activity was recorded from these muscles while focal transcranial magnetic stimulation (TMS) was applied by means of a 70 mm figure of eight coil connected to a Magstim 200 stimulator (Magstim, Whitland, Dyfed, UK). The coil was positioned over the left hemisphere, tangentially to the scalp with the handle pointing backward and laterally at 45° away from the mid-sagittal line, such that the induced current flowed in posterior–anterior direction, i.e., approximately perpendicular to the central sulcus. The optimal scalp position was defined as the position from which motor evoked potentials (MEPs) with maximal amplitude were recorded in the right OP muscle. The rest motor threshold (rMT) was defined as the lowest stimulus intensity evoking MEPs in the OP with an amplitude of at least 50  $\mu$ V in 5 out of 10 consecutive stimuli (Rossini et al., 1994). Subjects' rest motor thresholds, expressed as a percentage of the maximum stimulator output, varied from 38% to 66% (mean  $52 \pm 9\%$ ). For all experimental trials, stimulation intensity was set at 130% of the subjects' rMT. Parameter setting procedures were prioritised for the OP muscle but MEPs were simultaneously obtained for the ECR muscle. Stimulation parameters were assumed to be satisfactorily similar, due to the overlapping representations of



**Fig. 1.** Illustration of the digital video clips presented to the subjects. All video clips lasted for 6 s, with a TMS pulse delivered at a random time point between 4.1 and 4.8 s. (A) Visual stimulus [ $V_{\text{crush}}$ ], (B) auditory stimulus [ $A_{\text{crush}}$ ], (C) congruent audio-visual stimulus [ $V_{\text{crush}}A_{\text{crush}}$ ], (D) incongruent sound [ $V_{\text{crush}}A_{\text{water}}$ ], and (E) incongruent vision [ $V_{\text{foot}}A_{\text{crush}}$ ].

finger and forearm extensor muscles (Scheiber, 1990). EMG recordings were sampled at 5000 Hz (CED Power 1401, Cambridge Electronic Design, UK) amplified, band-pass filtered (30–1500 Hz), and stored on a PC for off-line analysis. Signal Software (2.02 Version, Cambridge Electronic Design, UK) was used for TMS triggering and EMG recordings.

### 2.2.3. Procedure

Participants were seated in a comfortable chair in front of a Dell P992 monitor (resolution,  $1024 \times 768$  pixels; refresh frequency 60 Hz) on which video clips (audio–video interleaved (AVI)) were displayed with a frame rate of 25 Hz (or frames/s). Video presentation timing was controlled by Blaxton Video Capture software (South Yorkshire, UK). Before each experiment, video clips were presented to the subjects to familiarize them with the experimental stimuli. During the session, they were instructed to keep their hands and forearms as relaxed as possible and to pay full attention to the video presented. Vision of their own hand and forearm was never allowed. Muscle relaxation was monitored, and whenever increased EMG activity became apparent during data collection, the trial was discarded and repeated.

### 2.2.4. Stimuli

Audio–visual video clips of a simple hand action were presented to the subjects. A single object manipulation was used, i.e., the crushing of a small drinking bottle by abducting the thumb of the right hand. This action has an easily recognizable and powerful sound that is part of the motor repertoire of all subjects. In total, six different video clips (duration = 6 s) were presented to the subjects (Fig. 1): (A) Visual stimulus [ $V_{\text{crush}}$ ], i.e., only seeing the crushing action (Fig. 1A), (B) auditory stimulus [ $A_{\text{crush}}$ ] i.e., only hearing the sound related to the crushing action, while looking at a white background (Fig. 1B), and (C) congruent audio–visual stimulus [ $V_{\text{crush}}A_{\text{crush}}$ ] i.e., seeing and hearing the crushing action (Fig. 1C). Additionally, two audio–visual video clips were created in which the sound and sight were incongruent: (D) incongruent sound [ $V_{\text{crush}}A_{\text{water}}$ ], i.e., seeing the crushing action while hearing the sound of floating water (Fig. 1D). (E) Incongruent vision [ $V_{\text{foot}}A_{\text{crush}}$ ], i.e., hearing the (manual) crushing action while seeing a foot action, i.e., a right foot gently pushing over the drinking bottle (Fig. 1E). The foot action normally produces hardly any sound. (F) Additionally, a video displaying a 'white background + no sound' (BASELINE) was included to measure baseline corticomotor activity.

Sounds were controlled for intensity (recorded at similar amplitudes) and played at a volume which was comfortable to the subject. Each of the aforementioned audio–video clips was presented 20 times in blocks of four, with the order of the blocks randomized within and across subjects. Between blocked video clips, a black screen was shown for an interval of 3 s. The interval between blocked trials was approximately 2 min. During the presentation of each audio–video clip, a single TMS pulse was delivered at a random time point (between 4.1 and 4.8 s) during the actual contraction-phase of the right OP muscle (i.e., to crush the drinking bottle) (Fig. 1). In total, 120 MEPs were recorded for each subject.

### 2.2.5. Data analysis

From the EMG data, peak-to-peak amplitudes of the MEPs were determined. Additionally, the background EMG was quantified by calculating the root-mean-square error (RMSE) for the 50 ms interval prior to TMS stimulation to ensure that subjects were completely relaxed during the stimulation. Trials were removed from the analysis when EMG RMSE scores were larger than  $Q3 + 1.5 (Q3 - Q1)$  with  $Q1$ ,  $Q3$  being the first and third quartile considering all trials of one observation condition and subject. After data acquisition, one male and female subject were rejected from the analyses due to extreme background EMG recorded from respectively, the ECR (130.8% of baseline RMSE scores measurements) and OP muscle (115.3% of baseline RMSE). For the remaining subjects, 4% of all trials were discarded from further analyses because of high background EMG.

For OP and ECR muscles, MEP amplitudes were averaged separately for each of the six observation conditions. Since MEP size usually exhibits large inter-individual variability, MEP amplitudes were normalized for each muscle relative to the mean BASELINE measure ( $\text{MEP}/\text{MEP}_{\text{BASELINE}} \times 100$ ) in order to make them comparable across subjects. RMSE scores of the background EMG were normalized accordingly ( $\text{RMSE}/\text{RMSE}_{\text{BASELINE}} \times 100$ ).

### 2.2.6. Statistics

Analyses of variance (ANOVA) with repeated measures were conducted on the normalized peak-to-peak MEP amplitudes using Statistica 7.0 (StatSoft, Inc., Tulsa, USA). The level of significance was set to  $\alpha = .05$ . Since MEP responses from the OP and ECR muscles displayed similar modulations, all analyses were conducted on the pooled responses of the two muscles.

For each sensory input (visual and auditory) a one-way ANOVA model was designed to test the effect of 'Condition' (unimodal, multimodal congruent, multimodal incongruent). As such, the ANOVA for the visual modality contained the observation conditions: [ $V_{\text{crush}}$ ], [ $V_{\text{crush}}A_{\text{crush}}$ ] and [ $V_{\text{crush}}A_{\text{water}}$ ], whereas the model for the auditory modality contained the conditions: [ $A_{\text{crush}}$ ], [ $V_{\text{crush}}A_{\text{crush}}$ ] [ $V_{\text{foot}}A_{\text{crush}}$ ]. With this analysis we can specifically test the 'modality-independent action representation' hypothesis versus the 'multisensory action representation' hypothesis in the following way: absence of any effect would speak in favour of 'modality-independent action representations', as in this case, no differences in MEP responses

are found for the presentation of unimodal or multimodal stimuli (either congruent or incongruent). The hypothesis of 'shared modality-dependent action representations', on the other hand, predicts a significant response increase for the congruent multimodal stimulus as compared to the unimodal and incongruent multimodal stimuli.

To address whether peak-to-peak MEP amplitude scores were confounded by modulations in background muscle activity, all analyses performed on the peak-to-peak amplitude data were also conducted for the corresponding background EMG data (normalized RMSE scores).

## 3. Results

### 3.1. Action execution

Crushing a small plastic bottle with the right hand led to a substantial increase of EMG activity in the OP ( $63 \pm 13\%$  MVC) and ECR muscle ( $60 \pm 8\%$  MVC), but not in the FCR ( $26 \pm 4\%$  MVC). Accordingly, a one-way ANOVA design with the within factor 'Muscle' (OP, ECR, FCR) revealed a main effect of 'Muscle' [ $F(2,18) = 5.05$ ,  $p = .018$ ], indicating significantly higher muscle activation in the OP and ECR muscles compared to the FCR. Based on these results, effects of action perception were quantified for the OP and ECR.

### 3.2. Perception of unimodal visual input versus multimodal input

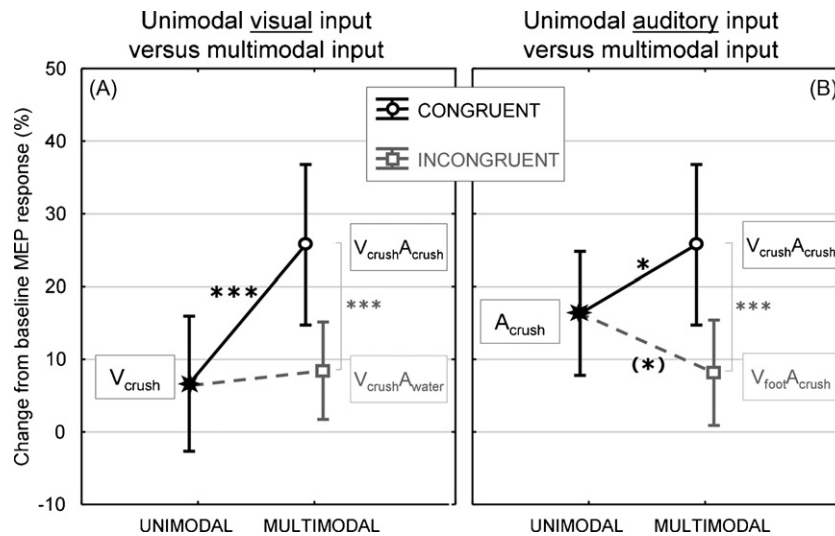
Compared to the unimodal visual stimulus [ $V_{\text{crush}}$ ], MEP responses were significantly higher for the multimodal congruent condition [ $V_{\text{crush}}A_{\text{crush}}$ ] [ $p < .001$ ], but not for the multimodal incongruent condition [ $V_{\text{crush}}A_{\text{water}}$ ] [ $p = .66$ ] (Fig. 2A). This was revealed by post hoc analysis (Fisher) of the significant main effect of 'Condition' [ $F(2,20) = 10.18$ ,  $p = .001$ ]. More specifically, MEP responses were significantly larger when subjects simultaneously heard and saw the crushing action [ $V_{\text{crush}}A_{\text{crush}}$ ] than when they only saw the action [ $V_{\text{crush}}$ ]. On the other hand, hearing the incongruent (non-action-related) sound of floating water while viewing the crushing action [ $V_{\text{crush}}A_{\text{water}}$ ], did not have this augmenting effect (Fig. 2A).

### 3.3. Perception of unimodal auditory input versus multimodal input

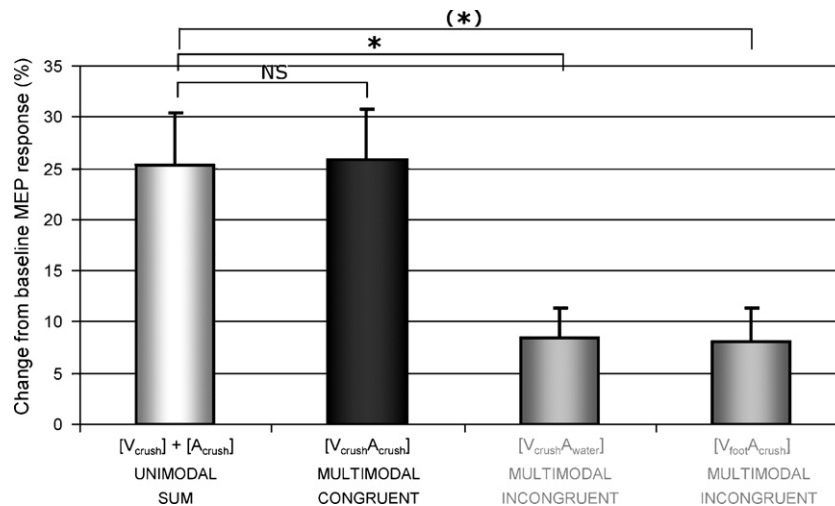
Compared to the unimodal auditory stimulus [ $A_{\text{crush}}$ ], MEP responses were significantly higher for the multimodal congruent condition [ $V_{\text{crush}}A_{\text{crush}}$ ] [ $p = .03$ ], whereas they were tentatively smaller for the multimodal incongruent condition [ $V_{\text{crush}}A_{\text{water}}$ ] [ $p = .05$ ] (Fig. 2B). This was revealed by post hoc analysis of the significant main effect of 'Condition' [ $F(2,20) = 7.06$ ,  $p = .005$ ]. More specifically, MEP responses were significantly larger when subjects simultaneously heard and saw the crushing action [ $V_{\text{crush}}A_{\text{crush}}$ ] than when they only heard the action [ $A_{\text{crush}}$ ]. However, when the sound of the crushing action was heard, but an incongruent action was viewed (foot action) [ $V_{\text{foot}}A_{\text{crush}}$ ], MEP responses were shown to be tentatively lower than responses evoked during the presentation of the action sound alone [ $A_{\text{crush}}$ ] (Fig. 2B).

### 3.4. Multimodal input versus the sum of unimodal stimuli

The sum of MEP responses yielded from the unimodal conditions [ $V_{\text{crush}}$ ] + [ $A_{\text{crush}}$ ] was compared to responses measured from the multimodal congruent and incongruent conditions. Interestingly, the MEP response to the multimodal stimulus [ $V_{\text{crush}}A_{\text{crush}}$ ] approximately equaled the sum of the unimodal responses [ $V_{\text{crush}}$ ] + [ $A_{\text{crush}}$ ] (Fig. 3) and, accordingly, a  $t$ -test for dependent samples revealed no significance when directly comparing [ $V_{\text{crush}}A_{\text{crush}}$ ] versus [ $V_{\text{crush}}$ ] + [ $A_{\text{crush}}$ ] [ $t = .07$ ,  $p = 2.5$ ]. Thus, MEP responses to congruent multimodal input were found to be additive as compared to matched unimodal stimuli (multimodal response equals the sum of unimodal responses). On the other



**Fig. 2.** (A) Normalized peak-to-peak MEP amplitudes recorded during the presentation of unimodal visual input [ $V_{crush}$ ] versus multimodal congruent [ $V_{crush}A_{crush}$ ] (black lines) or incongruent input [ $V_{crush}A_{water}$ ] (grey lines). (B) Normalized peak-to-peak MEP amplitudes recorded during the presentation of unimodal auditory input [ $V_{crush}$ ] versus multimodal congruent [ $V_{crush}A_{crush}$ ] (black lines) or incongruent input [ $V_{foot}A_{crush}$ ] (grey lines). Pooled MEP responses, recorded from the OP and ECR muscles, are presented as a change from the mean MEP response recorded during the baseline (control) observation condition (in %). Vertical bars denote  $\pm$  standard error. [Differences between conditions are indicated; \*\*\* $p < .001$ ; \* $p < .05$ ; (\*) $p = .05$ .]



**Fig. 3.** The sum of normalized peak-to-peak MEP amplitudes recorded from the unimodal conditions [ $V_{crush}$ ] + [ $A_{crush}$ ] (white bar) is compared to responses measured from the multimodal congruent [ $V_{crush}A_{crush}$ ] (black bar) and multimodal incongruent conditions [ $V_{crush}A_{water}$ ] [ $V_{foot}A_{crush}$ ] (grey bars). Pooled MEP responses, recorded from the OP and ECR muscles, are presented as a change from the mean MEP response recorded during the baseline (control) observation condition (in %). Vertical bars denote  $\pm$  standard error. [Differences between conditions are indicated; \* $p < .05$ ; (\*) $p = .05$ ; NS: not significant.]

hand, responses to the incongruent multimodal conditions were found to be *sub-additive* compared to the sum of unimodal stimuli [ $V_{crush}$ ] + [ $A_{crush}$ ] (Fig. 3), which was further confirmed by *t*-tests [both,  $t > .93$ ,  $p = .05$ ].

### 3.5. Background muscle activity during movement observation/sound listening

The background EMG was generally small and condition induced modulations were minimal, such that statistics on the RMSE scores did not reveal significant effects [all  $F(2,20) < 1.5$ ,  $p > .23$ ]. This indicates that the MEP peak-to-peak amplitude scores were not confounded by modulations in background EMG.

## 4. Discussion

With the present TMS study, we explored multimodal audio-visual interactions at the level of the human primary motor

cortex (M1) during action perception. The novel finding is that responses to TMS were additive when a hand action was simultaneously heard and seen, i.e., the multimodal response *equaled* the sum of unimodal responses. Substantially lower responses were obtained from the presentation of mismatched audio-visual stimuli. These findings speak in favour of 'shared modality-dependent' action representations at the level of the observer's primary motor cortex, which are evoked more robustly when congruent auditory and visual stimuli of hand actions are presented simultaneously.

### 4.1. Shared modality-dependent action representations in the human motor system

We have shown that congruent multimodal sensory input related to manipulative hand actions, activated the primary motor cortex (M1) in such a way that the obtained response equalled the sum of responses generated by the

visual and auditory input separately. The additive response depended strongly on stimulus congruency as mismatched combinations of audio–visual cues evoked substantially lower responses.

These results do not support the hypothesis that perceived actions are represented in a ‘modality-independent’ way, as in this case, responses are expected to be similar irrespective of whether the action is only heard, only seen or both heard and seen. Instead, our data speak in favour of ‘shared modality-dependent action representations’, which are evoked more robustly from the simultaneous presentation of congruent auditory and visual stimuli.

It is important to note that TMS most likely measures the input of projections originating from regions “up-stream” of M1. In the context of our paradigm, the inferior frontal gyrus (IFG) is a very likely candidate, as it is a core area of the human mirror neuron system and it is known to have strong reciprocal cortico-cortical connections with M1 (Dum & Strick, 2005; Shimazu, Maier, Cerri, Kirkwood, & Lemon, 2004). However, in monkey’s F5 – the putative analogue of human IFG – evidence was found for both ‘modality-independent’ as well as ‘multisensory driven’ action representations (Keysers et al., 2003), whereas the present findings predominantly speak in favour of the latter type. A recent imaging study in humans also showed that a region of the ventral premotor cortex responded specifically to the combination of visual and auditory action-related stimuli (Kaplan & Iacoboni, 2007). Possibly this difference with monkeys can be explained by the development of communicative speech in humans which is known to rely strongly on the interaction between audio–visual inputs (see section below) (Summerfield, 1992).

However, the possibility remains that the convergence of auditory and visual input about the same action occurs even more up-stream of IFG. The superior temporal sulcus (STS) is a good candidate in this respect, as this region: (1) receives inputs from multiple senses (Hikosaka, Iwai, Saito, & Tanaka, 1988; Seltzer & Pandya, 1994), (2) is a major input structure to regions of the mirror neuron system (Allison, Puce, & McCarthy, 2000), and most importantly, (3) has repeatedly been implicated in cross-sensory integration. In the past, this has been studied most extensively in the domain of speech perception (Calvert, Campbell, & Brammer, 2000; Macaluso, George, Dolan, Spence, & Driver, 2004; Skipper, Nusbaum, & Small, 2005), although recent studies also report a role for the STS in multisensory integration of action-related stimuli. More specifically, monkey STS neurons displayed multisensory enhancements which were greater for congruent than for incongruent pairings of action-related input (Barracough, Xiao, Baker, Oram, & Perrett, 2005). Therefore, convergence of audio–visual input of a perceived action might already take place early during sensory-to-motor transformations, i.e., before this information is referenced to frontal mirror areas (Barracough et al., 2005).

Even though our data cannot clarify conclusively at which level of the neuroaxis ‘audio–visual–motor’ convergence was established, the finding of robust modality-dependent modulations at the systems level, more specifically, in M1, may contribute to the debate on how visual and auditory inputs interact to describe action events in the human motor system.

All together, the indication of perception-induced ‘activation’ of motor events strongly supports the main principle of ‘ideomotor theory’ first contended by James (1890), namely that “every representation of a movement awakens in some degree the actual movement which is its object” (James, 1890). In this context, it was suggested that these sensory-motor associations are not fixed, but arise primarily through correlated experience of executing and perceiving the same actions (Flach, Osman, Dickinson, & Heyes, 2006; Heyes, 2003; Heyes, Bird, Johnson, & Haggard, 2005). Indeed, through lifespan, people learn to associate certain

motor events with their perceptual consequences, and vice versa (Elsner & Hommel, 2004), and the action used in the present experiment (i.e., the manual crushing of a small drinking bottle) was certainly part of the motor repertoire of all the participating subjects.

However, aside from the notion that action representations are ‘integrated’ with respect to their perceptual and motor components, it was suggested that actions are represented as assemblies of codes that refer to the different features of the action (Hommel & Elsner, 2009). Within the latter framework (i.e., that multiple features represent the same action), different interpretations may account for the present findings.

On the one hand, audition and vision may interact to describe a similar set of action-specific features. This interpretation would suggest that the strong M1 response to congruent audio–visual stimuli reflects the binding of auditory and visual input in a process of true audio–visual integration for representing the perceived action.

Alternatively however, it can be assumed that congruent visual and auditory stimuli activated independent features from the same action. As such, the strong M1 responses for congruent audio–visual stimuli might have arisen from shared action representations without requiring true audio–visual integration at any level up-stream of M1. The fact that only *additive* responses (multimodal response equals the sum of unimodal responses) were found, and not *superadditive* responses (multimodal response exceeds the sum of unimodal responses) may speak in favour of the latter account (Stein & Stanford, 2008).

However, it is worth noting that to date, there is still considerable debate on how to interpret additive versus superadditive multisensory enhancements. On the one hand, it has been argued that regional brain responses (such as BOLD responses in imaging studies) must exceed the sum of responses to the modality-specific components if multisensory integration is to be conclusively established (Calvert, Hansen, Iversen, & Brammer, 2001; Meredith & Stein, 1986). Indeed, this superadditivity requirement recognizes that lesser enhancements (such as additive responses) could reflect the independent contributions of neighbouring unisensory neurons and not true multisensory integration. However, in the field of cellular multisensory integration, it is a well-known feature that the magnitude of multisensory enhancement is typically inversely related to the effectiveness of the individual cues that are being combined (i.e., the principle of inverse effectiveness) (Meredith & Stein, 1986). Therefore, also at the systems level, it is unlikely to find superadditive effects when highly salient unimodal cues are used (which was the case in the present study).

In summary, considering that the present study adopted rather salient action-related unimodal cues, it remains an open question whether the observed multisensory enhancement conclusively reflects ‘true multisensory integration’.

#### 4.2. Functional relevance of shared modality-dependent action representations in the motor system

The recruitment of motor areas during action perception is broadly assumed to contribute to cognitive processes such as action recognition and understanding (Rizzolatti & Craighero, 2004). Following the “simulation theory”, action understanding is achieved by matching visual and/or auditory action-related sensory input onto the motor representation of the same action event (Jeannerod, 2001; Pazzaglia, Pizzamiglio, Pes, & Aglioti, 2008). Recently, two elegant studies (Pobric & Hamilton, 2006; Meister, Wilson, Deblieck, Wu, & Iacoboni, 2007) have established that the contribution of the mirror system, and more specifically IFG, is crucial during perceptual judgment and discrimination tasks relying on action observation. As such, it can be speculated that the convergence of

action-related 'audio-visual' cues at the level of the motor system is functionally relevant in creating a robust 'audio-visual-motor' percept. In our experiment, corticomotor excitability was highest when the auditory and visual input were congruent in source and related to the manual crushing action, whereas responses were substantially lower for 'mismatched' multimodal conditions in which only one sensory component corresponded to the hand action. Moreover, for the multimodal condition, in which only the auditory component matched the hand action and an 'irrelevant' foot action was seen, corticomotor responses were even tentatively smaller compared to only hearing the crushing action. On the other hand, no interference was found when 'irrelevant' audition (water sound) was presented on top of the visually perceived crushing action. These findings may relate to the well-established feature that people tend to rely more strongly on the visual modality to select afferent information from their surroundings—when audio-visual input is simultaneously available (Colavita, 1974; Colavita & Weisberg, 1979; Colavita, Tomko, & Weisberg, 1976). As such, the processing of 'irrelevant vision' may have dominated over the creation of an 'audio-motor' percept from the simultaneously heard manual action sound, whereas the creation of a visuo-motor percept from the sight of the hand action remained fairly unaffected by irrelevant sounds. It should be noted however, that the 'irrelevant vision' pertained to another motor act, namely to a foot action, whereas the presented irrelevant sound was entirely non-action-related (water sound). Nonetheless, we assume that the dominance of visual processing over auditory perception will persist, irrespective of whether it is action-related or not.

Interestingly, very similar audio-visual interactions have been demonstrated for speech perception which improves substantially when a speaker is both heard and seen (Grant & Seitz, 2000; Sumbly & Pollack, 1954). By contrast, when visual and auditory cues are incongruent, as when watching a dubbed movie, the viewing of the incongruent lip and mouth movements, may interfere substantially with the perception of an otherwise clear auditory speech signal. A well-described example of the influence of incongruent vision on the perception of speech sounds, is the McGurk effect, in which an auditory/ba/ combined with a visual/ga/ is typically perceived as da/ (McGurk & MacDonald, 1976). As such, the multimodal dynamics observed in our study seem to exhibit striking functional similarities to speech perception, even though manipulative hand actions were investigated.

In summary, our results suggest that multimodality is an important feature of the mirror neuron system which is not specific for human speech perception but applies to action perception in general.

## Acknowledgements

Support for this study was provided through grants from the Flanders Fund for Scientific Research (Projects G.0292.05, G.0577.06 and G.0749.09). This work was also supported by Grant P6/29 from the Interuniversity Attraction Poles program of the Belgian federal government.

## References

- Allison, T., Puce, A., & McCarthy, G. (2000). Social perception from visual cues: Role of the STS region. *Trends in Cognitive Sciences*, 4, 267–278.
- Aziz-Zadeh, L., Iacoboni, M., Zaidel, E., Wilson, S., & Mazziotta, J. (2004). Left hemisphere motor facilitation in response to manual action sounds. *European Journal of Neuroscience*, 19, 2609–2612.
- Barracough, N. E., Xiao, D., Baker, C. I., Oram, M. W., & Perrett, D. I. (2005). Integration of visual and auditory information by superior temporal sulcus neurons responsive to the sight of actions. *Journal of Cognitive Neuroscience*, 17, 377–391.
- Calvert, G. A., Campbell, R., & Brammer, M. J. (2000). Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex. *Current Biology*, 10, 649–657.
- Calvert, G. A., Hansen, P. C., Iversen, S. D., & Brammer, M. J. (2001). Detection of audio-visual integration sites in humans by application of electrophysiological criteria to the BOLD effect. *Neuroimage*, 14, 427–438.
- Colavita, F. B. (1974). Human Sensory Dominance. *Perception & Psychophysics*, 16, 409–412.
- Colavita, F. B., Tomko, R., & Weisberg, D. (1976). Visual pre-potency and eye orientation. *Bulletin of the Psychonomic Society*, 8, 25–26.
- Colavita, F. B., & Weisberg, D. (1979). Further investigation of visual dominance. *Perception & Psychophysics*, 25, 345–347.
- Di Pellegrino, G., Fadiga, L., Fogassi, L., Gallese, V., & Rizzolatti, G. (1992). Understanding motor events: A neurophysiological study. *Experimental Brain Research*, 91, 176–180.
- Dum, R. P., & Strick, P. L. (2005). Frontal lobe inputs to the digit representations of the motor areas on the lateral surface of the hemisphere. *Journal of Neuroscience*, 25, 1375–1386.
- Elsner, B., & Hommel, B. (2004). Contiguity and contingency in action-effect learning. *Psychological Research*, 68, 138–154.
- Fadiga, L., Fogassi, L., Pavesi, G., & Rizzolatti, G. (1995). Motor facilitation during action observation: A magnetic stimulation study. *Journal of Neurophysiology*, 73, 2608–2611.
- Flach, R., Osman, M., Dickinson, A., & Heyes, C. (2006). The interaction between response effects during the acquisition of response priming. *ACTA Psychologica (Amsterdam)*, 122, 11–26.
- Gazzola, V., Aziz-Zadeh, L., & Keysers, C. (2006). Empathy and the somatotopic auditory mirror system in humans. *Current Biology*, 16, 1824–1829.
- Grant, K. W., & Seitz, P. F. (2000). The use of visible speech cues for improving auditory detection of spoken sentences. *The Journal of the Acoustical Society of America*, 108, 1197–1208.
- Heyes, C. (2003). Four routes of cognitive evolution. *Psychological Review*, 110, 713–727.
- Heyes, C., Bird, G., Johnson, H., & Haggard, P. (2005). Experience modulates automatic imitation. *Brain Research: Cognitive Brain Research*, 22, 233–240.
- Hikosaka, K., Iwai, E., Saito, H., & Tanaka, K. (1988). Polysensory properties of neurons in the anterior bank of the caudal superior temporal sulcus of the macaque monkey. *Journal of Neurophysiology*, 60, 1615–1637.
- Hommel, B., & Elsner, B. (2009). Acquisition, representation and control of action. In E. Morsella, J. A. Bargh, & P. M. Gollwitzer (Eds.), *Oxford handbook of human action* (pp. 371–398). New York: Oxford University Press.
- Iacoboni, M. (2005). Neural mechanisms of imitation. *Current Opinion in Neurobiology*, 15, 632–637.
- James, (1890). *The principles of psychology*. New York: Dover Publications.
- Jeannerod, M. (2001). Neural simulation of action: A unifying mechanism for motor cognition. *Neuroimage*, 14, S103–S109.
- Kaplan, J. T., & Iacoboni, M. (2007). Multimodal action representation in human left ventral premotor cortex. *Cognitive Processing*, 8, 103–113.
- Keysers, C., Kohler, E., Umiltà, M. A., Nanetti, L., Fogassi, L., & Gallese, V. (2003). Audiovisual mirror neurons and action recognition. *Experimental Brain Research*, 153, 628–636.
- Kohler, E., Keysers, C., Umiltà, M. A., Fogassi, L., Gallese, V., & Rizzolatti, G. (2002). Hearing sounds, understanding actions: Action representation in mirror neurons. *Science*, 297, 846–848.
- Macaluso, E., George, N., Dolan, R., Spence, C., & Driver, J. (2004). Spatial and temporal factors during processing of audiovisual speech: A PET study. *Neuroimage*, 21, 725–732.
- McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, 264, 746–748.
- Meister, I. G., Wilson, S. M., Deblieck, C., Wu, A. D., & Iacoboni, M. (2007). The essential role of premotor cortex in speech perception. *Current Biology*, 17, 1692–1696.
- Meredith, M. A., & Stein, B. E. (1986). Visual, auditory, and somatosensory convergence on cells in superior colliculus results in multisensory integration. *Journal of Neurophysiology*, 56, 640–662.
- Oldfield, R. C. (1971). The assessment and analysis of handedness: The Edinburgh inventory. *Neuropsychologia*, 9, 97–113.
- Pazzaglia, M., Pizzamiglio, L., Pes, E., & Aglioti, S. M. (2008). The sound of actions in apraxia. *Current Biology*, 18, 1766–1772.
- Pobric, G., & Hamilton, A. F. (2006). Action understanding requires the left inferior frontal cortex. *Current Biology*, 16, 524–529.
- Rickham, P. P. (1964). Human experimentation. Code of Ethics of the World Medical Association. Declaration of Helsinki. *British Medical Journal*, 2, 177.
- Rizzolatti, G., & Craighero, L. (2004). The mirror-neuron system. *Annual Review of Neuroscience*, 27, 169–192.
- Rossini, P. M., Barker, A. T., Berardelli, A., Caramia, M. D., Caruso, G., Cracco, R. Q., et al. (1994). Noninvasive electrical and magnetic stimulation of the brain, spinal-cord and roots—Basic principles and procedures for routine clinical-application—Report of an IFCN Committee. *Electroencephalography and Clinical Neurophysiology*, 91, 79–92.
- Scheiber, M. H. (1990). How might the motor cortex individuate movements. *Trends in Neurosciences*, 13, 440–445.
- Seltzer, B., & Pandya, D. N. (1994). Parietal, temporal, and occipital projections to cortex of the superior temporal sulcus in the rhesus monkey: A retrograde tracer study. *Journal of Comparative Neurology*, 343, 445–463.
- Shimazu, H., Maier, M. A., Cerri, G., Kirkwood, P. A., & Lemon, R. N. (2004). Macaque ventral premotor cortex exerts powerful facilitation of motor cortex outputs to upper limb motoneurons. *Journal of Neuroscience*, 24, 1200–1211.

- Skipper, J. I., Nusbaum, H. C., & Small, S. L. (2005). Listening to talking faces: Motor cortical activation during speech perception. *Neuroimage*, 25, 76–89.
- Stein, B. E., & Stanford, T. R. (2008). Multisensory integration: Current issues from the perspective of the single neuron. *Nature Reviews Neuroscience*, 9, 255–266.
- Sumby, W., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *The Journal of the Acoustical Society of America*, 26, 212–215.
- Summerfield, Q. (1992). Lipreading and audio–visual speech perception. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 335, 71–78.