Multisensory Identification of Natural Objects in a Two-Way Crossmodal Priming Paradigm

Till R. Schneider[1], Andreas K. Engel[1], and Stefan Debener[1,2]

[1]Department of Neurophysiology and Pathophysiology, Center of Experimental Medicine, University Medical Center Hamburg-Eppendorf, Hamburg, Germany, [2]MRC Institute of Hearing Research Southampton, Royal South Hants Hospital, Southampton, UK

Abstract. The question of how vision and audition interact in natural object identification is currently a matter of debate. We developed a large set of auditory and visual stimuli representing natural objects in order to facilitate research in the field of multisensory processing. Normative data was obtained for 270 brief environmental sounds and 320 visual object stimuli. Each stimulus was named, categorized, and rated with regard to familiarity and emotional valence by $N = 56$ participants (Study 1). This multimodal stimulus set was employed in two subsequent crossmodal priming experiments that used semantically congruent and incongruent stimulus pairs in a S1–S2 paradigm. Task-relevant targets were either auditory (Study 2) or visual stimuli (Study 3). The behavioral data of both experiments expressed a crossmodal priming effect with shorter reaction times for congruent as compared to incongruent stimulus pairs. The observed facilitation effect suggests that object identification in one modality is influenced by input from another modality. This result implicates that congruent visual and auditory stimulus pairs were perceived as the same object and demonstrates a first validation of the multimodal stimulus set.

Keywords: auditory, visual, stimulus set, crossmodal priming, object identification

Over the past years, increasing research effort has been devoted to the study of multisensory interactions and their role for attention, perception, memory and behavior (Calvert, 2001; Calvert, Spence, & Stein, 2004). Object identification in the real world usually requires that information from multiple sensory modalities is integrated and utilized (Amedi, von Kriegstein, van Atteveldt, Beauchamp, & Naumer, 2005; Beauchamp, Lee, Argall, & Martin, 2004; Woods & Newell, 2004). Ideally, tasks used to study object identification in multisensory paradigms should have high ecological validity which can be achieved by using complex stimulus materials reflecting, if possible, real world objects (Newell, 2004).

In order to test whether auditory perception of objects is influenced by visual cues of natural objects and vice versa, at first we developed a new multimodal stimulus series and subsequently applied some of these stimuli in a crossmodal priming paradigm. Priming is commonly referred to as a change in the speed or accuracy of the identification of an object following repeated experience with the same or a related stimulus (Henson, 2003; Tulving & Schacter, 1990). A large number of priming studies examined verbal priming within a single modality, whereas only a few studies investigated priming across sensory modalities, mainly by using verbal tasks (McClelland & Pring, 1991). For instance, the identification of environmental sounds is facilitated by prior presentation of the same sound, but not by the corresponding verbal label (Stuart & Jones, 1995, 1996). Crossmodal priming effects have also been reported between the visual and the haptic system with a magnitude comparable to within-modal priming effects (Easton, Greene, & Srinivas, 1997). Greene, Easton, and LaShell (2001) examined crossmodal priming effects using visual-auditory events. The visual trace of short video sequences facilitated the identification of the sounds of the same sequences, but not vice versa.

It is common practice to study only a relatively small set of stimulus pairs in studies of natural object identification, and the same stimuli are presented repeatedly. This approach is suboptimal because the repetition of identical stimuli poses a confounding factor in various fields such as novelty processing, memory encoding or repetition priming. A large set of auditory and visual stimuli representing natural objects suitable for multisensory research is lacking so far. In the field of emotion research, the development and free distribution of normative data for a large set of visual stimuli, the international affective picture system (IAPS; Lang, Öhman, & Vaitl, 1988) and a separate database of affective auditory stimuli (Bradley & Lang, 1999) has led to clear advances (Lang, Bradley, & Cuthbert, 1998). For example, the utilization of the same stimulus set in different laboratories and experiments allows more valid comparisons across studies.

Most existing databases, however, contain object stim-

DOI 10.1027/1618-3169.55.2.121

uli for either the visual or the auditory modality (Ballas, 1993; Fabiani, Kazmerski, Cycowicz, & Friedman, 1996; Snodgrass & Vanderwart, 1980). The only published set of combined visual and auditory stimuli of objects consists of black-and-white line drawings and nonverbal sounds in different lengths ranging between hundreds of milliseconds and over four seconds (Saygin, Dick, & Bates, 2005). In the visual domain a frequently used standardized set of objects has been provided by Snodgrass and Vanderwart (1980). This set consists of black-and-white line drawings and provides normative data on several variables of relevance for the study of cognitive processing. Similar visual stimuli in a gray-level and a color version with stimulus norms are provided by Rossion and Pourtois (2004). In the auditory domain a stimulus set consisting of 41 environmental sounds published by Ballas (1993) provides information regarding the accuracy of identification, familiarity, and mean identification time. Another available set of auditory stimuli consists of 96 environmental sounds (Fabiani et al., 1996). Due to their short duration these stimuli are suitable for psychophysiological research and have been used to study stimulus-driven attention (Debener, Herrmann, Kranczioch, Gembris, & Engel, 2003; Debener, Kranczioch, Herrmann, & Engel, 2002; Debener, Makeig, Delorme, & Engel, 2005; Gaeta, Friedman, & Hunt, 2003).

In the present study we developed a new multimodal stimulus set (MULTIMOST) containing semantically congruent visual and auditory stimuli representing natural objects and collected stimulus information on several psychological variables. We aimed at obtaining a set of stimuli with high ecological validity, for instance, by using color photographs of objects instead of black-and-white drawings. All stimuli were tested in high quality, allowing the subsequent selection and adjustment of stimuli tailored to address different research questions, which for example require normalization or degradation of the material. Physical equalization techniques adjusting brightness, contrast or visual frequency often degrade the natural appearance of stimuli. This adjustment can be easily applied to original high quality stimuli, whereas the reverse processing from normalized to the original stimulus is more complicated. Thus, if stimuli representing natural objects are being used, they inevitably differ in some physical properties. Accordingly, information about psychological attributes is of the utmost importance here. We therefore collected data on familiarity, emotional valence, identification, categorization, and name agreement of each stimulus. Methods and results of the stimulus norming procedure are described in Study 1 below. In Study 2 and 3 semantically congruent and incongruent stimulus pairs were presented and participants were asked to make a fast categorical decision based on the identity of the stimulus. Based on the assumption that the selected stimulus pairs were perceived as reflecting one object, we predicted the identification of an object to be facilitated by the presentation of a semantically congruent stimulus in the other modality.

Study 1

Methods

Participants

Fifty-six volunteers (30 women and 26 men, mean age = 23.9, range: 21 to 33 years) were recruited at the University Medical Center Hamburg and received monetary compensation for participation. All participants were native German speakers, had normal or corrected-to-normal vision and normal hearing, and reported no history of neurological or psychic illness.

Materials and Procedure

A set of visual object stimuli was constructed by selecting 320 color photographs from a pool of 50.000 pictures of a digital photo database (Hemera Photo Objects, Vol. 1, Hemera, Hull, Canada). Visual objects were selected for which a characteristic environmental sound could be found and which would likely be recognized. Each selected picture stimulus was processed such that only the object on a black background was visible in the center of the image, and saved in JPEG format. The size of the pictures was adjusted, so that each picture covered approximately the same space on the screen resulting in a mean size of 413 × 511 pixels with a range of 202 to 826 pixels (height) and 239 to 974 pixels (width). All visual stimuli were allocated in a consensual decision (S.D., T.R.S.) to one of the ten following categories: animals, computer & communicative devices, kitchen utensils, musical instruments, sport equipment, machines, vehicles, weapons, tools, and everyday objects. The categories chosen widely overlapped with those used in previous studies on natural objects (e.g., Fabiani et al., 1996).

A set of auditory stimuli was created, selecting characteristic sounds of natural objects out of 1200 environmental sound files, which were derived from 12 sound effect CDs (100 Spectacular Sound FX, Mediaphon, Leinfelden-Echterdingen, Germany). Only those environmental sounds were selected which would likely be recognized in a very short time period. Each auditory stimulus was created by selecting the most characteristic 400 ms epoch from the respective sound file. The epoch containing the distinctive sound of one object was saved in a digital sound file with a sampling rate of 22 kHz (16-bit, mono, WAV-format). For some objects two or more sound files were created, in order to provide the opportunity to empirically find those sounds fitting best to a visual object. As a result, 180 out of the total of 270 sound files referred to different sound objects. The sound intensities were adjusted by equalizing the root mean square power of all sound files. In order to avoid on- and offset clicking noises, stimulus intensity at the beginning and the end of each file was decreased by a filter, resulting in a 10 ms rise and fall time. Note that the standardized duration, the equalization of the loudness, and the abolishment of clicking noises make the

sound files suitable for electrophysiological experiments. All selected auditory object stimuli were allocated to the same ten categories as the visual objects. Selection and processing of the sounds was motivated by suggestions according to Shafiro and Gygi (2004).

Participants took part individually in a visual and an auditory session on two separate days. Half of the participants started with the auditory session, the other half started with the visual session. Each session lasted approximately 150 minutes. Subjects sat in a comfortable chair in a dimly lit, sound attenuated chamber, facing a computer screen with a distance of 80 cm. Visual stimuli were presented centrally for 400 ms on a 21-inch monitor. Mean image size subtended 4.5° visual angle vertical and 5.5° horizontal (range: 2.15° to 9° vertical; 2.5° to 10.5° horizontal). Auditory stimuli (400 ms) were delivered binaurally via Eartone foam protected air-tube earphones (Aero Company, Indianapolis, IN, USA) at approximately 70 dB SPL. All stimuli were presented in an individually randomized order to each subject using Presentation software (Version 0.80, www.neuro-bs.com, NeuroBehavioral Systems, Albany, CA, USA).

Participants were instructed to attend to each stimulus. Immediately after each stimulus presentation, participants were requested to respond to the questions appearing on the screen. The following judgments had to be made after each stimulus in this order of appearance:

Familiarity

For the familiarity rating of the stimuli, participants were instructed to indicate how familiar they were with the object presented on a scale ranging from 1 (familiar) to 4 (unfamiliar). They were asked to respond to the object itself and not to the way it was presented.

Emotional Valence

For the emotional valence rating of the stimuli, participants rated the pleasantness of the object represented by the stimulus. The scale ranged from 1 (*pleasant*) via 3 (*neutral*) to 5 (*unpleasant*).

Categorization

Participants allocated each sound to one of the ten given categories which were displayed on the screen.

Identification

Participants were instructed to name each sound silently and type the name on the keyboard. If they were not sure which object was presented, they were allowed to type a code for *don't know*.

Confidence

Participants judged how confident they felt about their decisions in the identification and the categorization task on a scale ranging from 1 (*confident*) to 4 (*unconfident*).

Data Analysis

For each stimulus, means and standard deviations of familiarity, emotional valence and confidence ratings were calculated. In order to measure the identification rate of each stimulus the percentage of subjects correctly naming each object according to the physical sound source was determined. Two different expressions, which were referring to the same concept were accepted as correct identification (e.g., Computer & PC), and typing errors were ignored.

The information measure H was computed for each item as it reflects naming agreement among subjects. H depends on the number of alternative names which are given to the same stimulus. The greater the number of alternative names given for one object, the larger H. For example, $H = 0$ indicates that all subjects named the sound identically, whereas $H = 1$ indicates that two different names were given with equal frequency to the item. H was calculated as follows according to Snodgrass and Vanderwart (1980):

$$H = \sum_{i=1}^{k} p_i \log_2(1/p_i),$$

where i indexes the name given, k denotes the number of different names given to each stimulus and p_i is the proportion of subjects giving each name. Note the logarithmic dependency of the measure, $H = 2$ for example indicates that four different names were given on average to the item. The *don't know* answers are not included in the calculation.

Results

A grand mean of 25% (SD = 28.13) of the auditory and of 84% (SD = 24.14) of the visual stimuli were identified correctly. Examples for auditory stimuli with high values of correct identification are the items *Sheep* (100% correct naming) and *Guitar* (98%). Auditory items with poor distinct temporal information such as *Hairdryer* and *Laser Printer* were named correctly by none of the participants. Numerous visual stimuli were correctly named by all participants, such as *Car* or *Dog*. Visual stimuli with low values of correct identification such as *Cello* (5%) and *Lemur* (10%) are items which require more specific knowledge than the majority of the objects. The distribution of the correct identification values of all stimuli is presented in Figure 1A. As can be seen, the fraction of items with high correct identification rates was higher among visual stimuli than among auditory stimuli. However, the number of identification rates, which range between 25% correct and 75% correct is 70 for the auditory and 54 for the visual stimuli. Similar results were apparent for the distribution of the categorization values (Figure 1B). Here, a grand mean of 40%

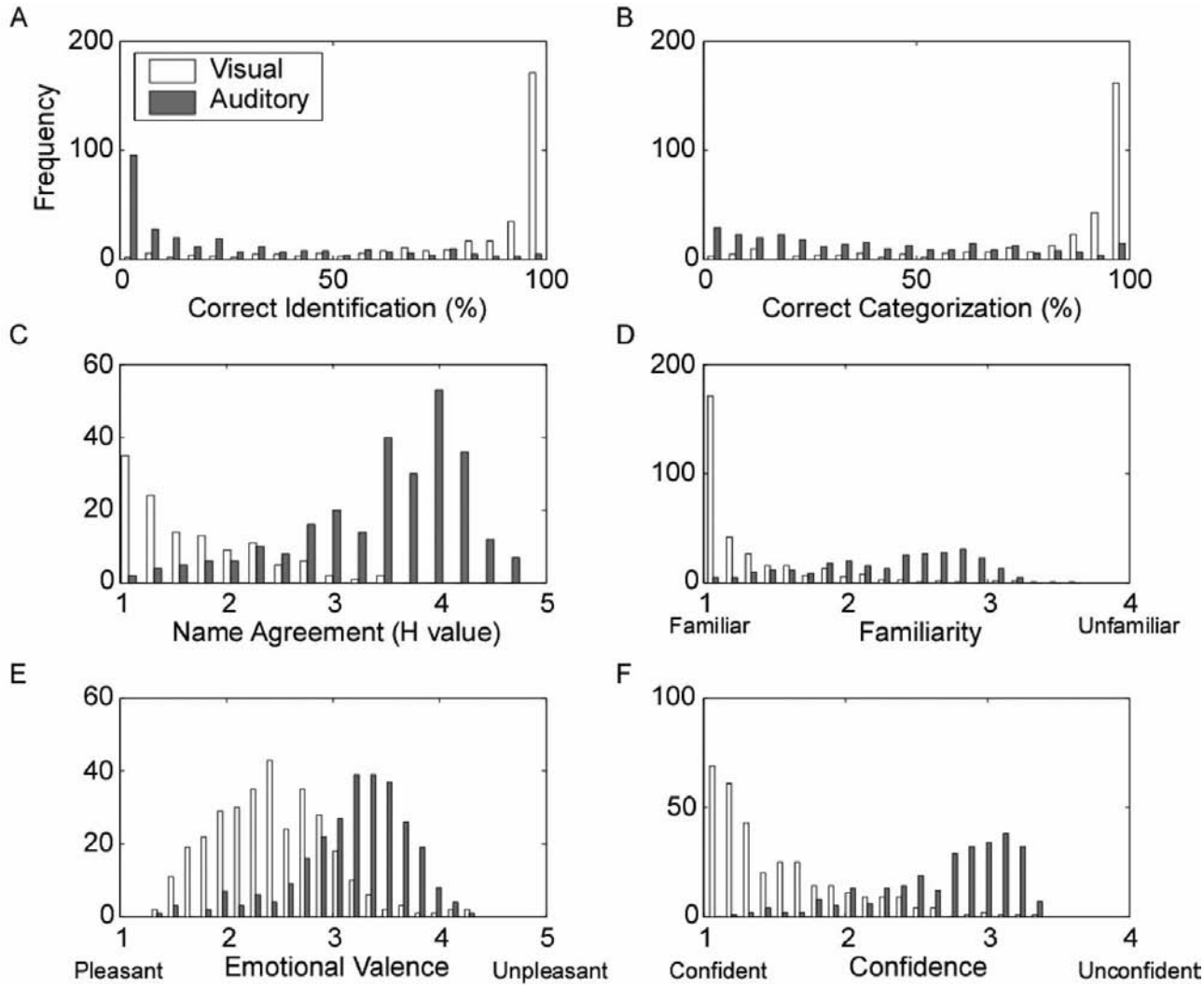

Figure 1. Frequency distribution of visual and auditory stimuli in six variables: correct identification (A), correct categorization (B), name agreement (C), familiarity (D), emotional valence (E) and confidence of judgment (F). (Note the different overall numbers of auditory (270) and visual (320) stimuli.)

Table 1. Mean ratings and standard deviations for the familiarity, emotional valence, and confidence of judgment values for all stimuli ($N = 56$)

Stimuli	Familiarity		Emotional valence		Confidence	
	M	*SD*	*M*	*SD*	*M*	*SD*
Auditory	2.31	0.54	3.19	0.52	2.71	0.47
Visual	1.33	0.47	2.43	0.54	1.51	0.46

Note. Familiarity: 1 = very familiar, 4 = very unfamiliar, Emotional Valence: 1 = very pleasant, 5 = very unpleasant, Confidence: 1 = very confident, 4 = very unconfident.

(*SD* = 29.48) of the auditory stimuli and of 83% (*SD* = 25.53) of the visual stimuli were allocated to the correct category. The naming agreement of auditory stimuli as reflected in average *H* ranged from 0.6 to 4.8 (Figure 1C) with a grand mean of 3.24 (*SD* = 0.85). The average *H* of visual stimuli ranged from 0 to 3.5 with a grand mean of 0.80 (*SD* = 0.81). Examples of visual stimuli with high naming agreement (*H* = 0) are, for instance, *Computer* and *Lion* and among auditory items *Guitar* (*H* = 0.63) and *Cat* (*H* = 0.88). Auditory stimuli with low naming agreement were *Bees* (*H* = 4.78) and *Zipper* (*H* = 4.56) and *Lemur* (*H* = 3.5) and *Cembalo* (*H* = 3.29) for the visual domain.

Means and standard deviations of the familiarity, emotional valence, and confidence of judgment ratings are presented in Table 1. The distribution of familiarity ratings (Figure 1D) ranged from 1 to 3.68 for the visual stimuli and

Table 2. Twenty exemplars of 270 auditory stimuli with mean values in familiarity, emotional valence, categorization, identification, and confidence of judgment ($N = 56$)

	Fam	Emo	Cat (%)	Id (%)	*H*	Con
Airplane	2.07	3.55	86	70	2.47	2.39
Ambulance	1.54	3.48	63	79	3.85	1.98
Bell	1.66	3.14	32	79	2.91	2.30
Bike Bell	1.50	3.00	4	59	3.51	2.09
Car	2.13	3.41	71	64	3.52	2.55
Cat	1.61	2.59	79	77	0.88	1.93
Cow	1.68	2.50	64	57	2.64	2.11
Dog	1.16	2.55	96	96	1.30	1.30
Frog	1.36	2.27	100	80	1.90	1.68
Goat	1.20	1.98	98	95	1.54	1.57
Guitar	1.13	1.45	100	98	0.63	1.48
Gun	2.02	3.88	71	61	2.63	2.38
Horse	1.59	2.30	80	66	1.88	1.82
Motorbike	1.93	3.11	80	55	2.18	2.41
Percussion	1.38	2.09	86	79	2.36	2.04
Pipe	1.48	3.21	20	70	2.19	2.04
Saw	1.91	3.11	73	68	1.69	2.29
Sheep	1.11	2.04	98	100	1.10	1.29
Tiger	1.84	2.86	82	71	1.64	2.39
Trumpet	2.13	3.20	39	32	3.32	2.67

Note. Fam = Familiarity (1 = very familiar, 4 = very unfamiliar), Emo = Emotional Valence (1 = very pleasant, 5 = very unpleasant), Cat = Categorization, Id = Identification, *H* = *H*-Value, Con = Confidence (1 = very confident, 4 = very unconfident).

Table 3. Twenty exemplars of 320 visual stimuli with mean values in familiarity, emotional valence, categorization, identification, and confidence of judgment ($N = 56$)

	Fam	Emo	Cat (%)	Id (%)	H	Con
Airplane	1.09	1.93	82	100	1.33	1.27
Ambulance	1.27	3.07	100	75	1.13	1.46
Bell	1.48	2.57	68	96	0.63	1.79
Bike Bell	1.77	2.91	14	79	0.87	1.86
Car	1.09	1.98	100	98	1.18	1.23
Cat	1.04	1.79	100	100	0.00	1.04
Cow	1.00	1.71	100	98	0.00	1.05
Dog	1.00	1.98	100	100	0.75	1.04
Frog	1.02	2.50	98	100	0.00	1.07
Goat	1.36	2.04	95	84	0.42	1.54
Guitar	1.02	1.46	100	96	0.00	1.05
Gun	1.09	3.95	98	100	0.00	1.21
Horse	1.00	1.71	98	100	0.00	1.05
Motorbike	1.05	2.50	98	98	0.00	1.18
Percussion	1.02	2.21	94	96	0.22	1.05
Pipe	1.11	2.78	96	100	0.00	1.32
Saw	1.13	2.96	100	100	0.13	1.36
Sheep	1.05	1.63	100	100	0.13	1.13
Tiger	1.02	1.88	98	98	0.13	1.05
Trumpet	1.07	2.20	95	93	0.31	1.27

Note. Fam = Familiarity (1 = very familiar, 4 = very unfamiliar), Emo = Emotional Valence (1 = very pleasant, 5 = very unpleasant), Cat = Categorization, Id = Identification, *H* = *H*-Value, Con = Confidence (1 = very confident, 4 = very unconfident).

Table 4. Mean values of stimulus norms for the congruent and incongruent stimulus sets used in Study 2

	Identification		Categorization		Familiarity		Emotional valence	
Crossmodal	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Congruent	27.08	28.48	42.29	28.67	2.23	2.3	3.13	3.18
Incongruent	26.06	31.37	40.41	31.49	2.32	2.48	3.15	3.31
Unimodal	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Congruent	26.13	28.13	38.61	28.97	2.26	0.55	3.14	0.5
Incongruent	28.08	30.5	42.69	30.31	2.25	0.53	3.12	0.55

from 1.05 to 3.2 for the auditory stimuli. Ratings on the emotional valence of the stimuli ranged from 1.34 to 4.36 for the visual stimuli and from 1.27 to 4.36 for the auditory stimuli. Visual and auditory stimuli differ in their mean emotional valence ratings (see Table 1). Familiar stimuli are usually perceived more positive than unfamiliar stimuli, an effect known as *mere exposure effect* (Zajonc, 1968). The observed differences in emotional valence between auditory and visual stimuli can be explained by this familiarity effect, as the valence ratings are highly correlated with the familiarity ratings ($r = .70$). However, the frequency of stimuli with emotional valence values around the neutral score (between 2.5 and 3.5) was 168 for the auditory and 123 for the visual stimuli (Figure 1E). Thus, it is possible to select a set of neutral stimuli for experiments in which controlling the emotional valence of the stimuli is important. Note that the stimulus set is not suitable for emotion research, which requires stimuli eliciting strong emotional reactions. Normative data in the collected six dependent variables are listed for 20 exemplary auditory (Table 2) and 20 visual (Table 3) stimuli. The complete normative data as well as the multimodal stimulus set are available from the authors upon request (www.multimost.com).

Study 2

Methods

Participants

A new group of 26 students from the University Medical Center Hamburg volunteered and received monetary compensation. The data of four participants had to be excluded due to technical malfunction. None of the selected individuals (21 women and 1 man, mean age: 23.82, range: 19 to 31 years) had taken part in study 1. All individuals were native German speakers, had normal or corrected-to-normal vision and normal hearing, and reported no history of neurological or psychic illness. The data of two subjects were discarded because in these more than 25% of the trials showed too long (> 2400 ms) or too short (< 400 ms) reaction times.

Materials and Procedure

Visual and auditory stimuli with the highest identification values in the norming study were selected for this experiment. For the unimodal condition 140 auditory-auditory stimulus pairs and for the crossmodal condition 170 visual-auditory stimulus pairs were selected to build similar stimulus sets. In each condition 50% of the stimulus pairs were semantically congruent, i.e., representing conceptually the same object, and 50% were semantically incongruent. Incongruent stimulus pairs represented always different objects and mostly different categories, 7% of the stimulus pairs in the crossmodal and 17% in the unimodal condition belonged to the same category. Congruent and incongruent stimulus sets were matched by adjusting the mean values in the variables familiarity, emotional valence, identification, and name agreement of each set (Table 4). The same matched relation of congruent and incongruent stimulus pairs was accomplished for the unimodal set, with the difference that the congruent stimulus pairs in this set were physically identical. This difference between the two conditions (concerning semantic vs. physical congruence) was intentional, as on the one hand the number of semantically but not physically congruent auditory stimulus pairs is limited and on the other hand results are better comparable to previous studies of repetition priming with physically identical stimulus pairs (e.g., Stuart & Jones, 1995). No significant differences emerged between the congruent and incongruent stimulus sets, a 2 × 2 ANOVA with factors Congruency (congruent vs. incongruent) and Modality (crossmodal vs. unimodal) revealed no main effects and no interactions in each of the four rating scales (all F values < 1).

All subjects were tested individually in a dimly lit, sound attenuated chamber, facing a computer screen (21-inch; 80 cm distance) and participated in both the unimodal and the crossmodal condition. The auditory setup was the same as in study 1. Unimodal and crossmodal trials were presented in blocks; half of the subjects started with the unimodal, the other half started with the crossmodal block. The stimuli in both blocks were presented in a pseudo-randomized order, never presenting the same object-category in two consecutive trials.

Each trial started with a fixation cross in the center of the screen for 500 ms, followed by the prime stimulus (S1) presented for 400 ms, either an auditory stimulus in the unimodal condition or a visual stimulus in the crossmodal con-

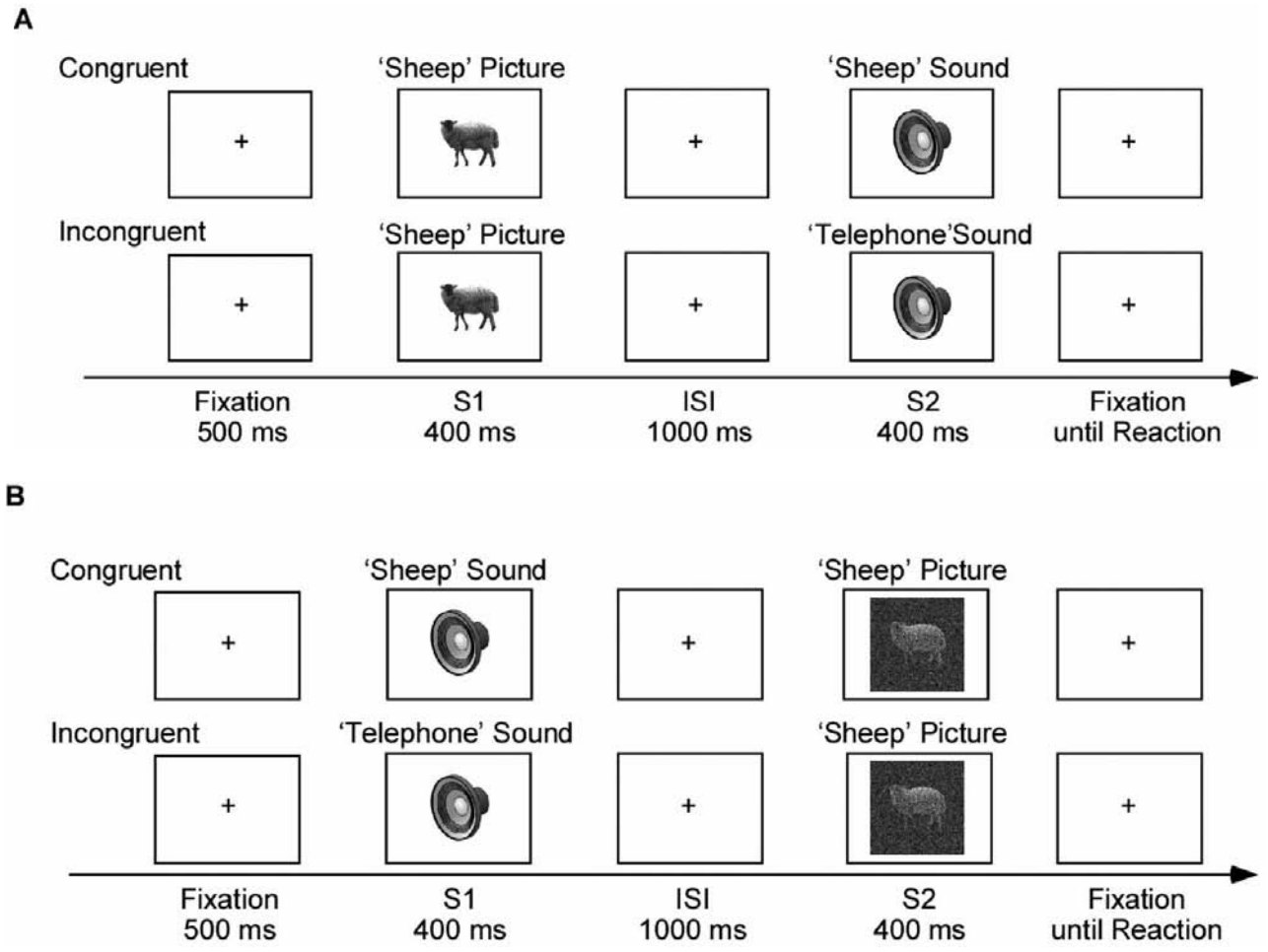

Figure 2. Schematic representation of the crossmodal priming paradigm. The trial order of the visual-auditory experiment (Study 2) is depicted in A and of the auditory-visual experiment (Study 3) is depicted in B. Each of the upper images represent a semantically congruent trial and each of the lower images a semantically incongruent trial. Subjects in both studies were asked to respond as fast and accurately as possible only after the second stimulus. Note that random noise was added to the visual stimuli of Study 3. (S1 = prime, S2 = target, ISI = interstimulus interval.)

dition (Figure 2A). Following the interstimulus interval of 1000 ms, the target stimulus (S2, auditory in both conditions) was presented for 400 ms. The task of the subjects in both conditions was to decide as quickly and accurately as possible after presentation of S2, whether the object represented by the second stimulus would fit into a shoebox or not. Subjects had to indicate their decision by pressing one of two buttons on a serial response pad (Model RB-420, Cedrus Corp., San Pedro, CA, USA) with their left or right thumb respectively. In order to counterbalance the use of the dominant hand, button labels were switched for 50% of the participants. The fixation cross remained until one of the two buttons was pressed or 2000 ms elapsed. In the latter case a screen appeared which reminded the subjects to respond faster. The entire trial lasted for about 4 seconds. Prior to the start of the experiment six practice trials were presented in order to familiarize subjects with the experimental procedure. The duration of the complete experimental procedure was approximately 30 minutes.

Data Analysis

For the analysis of the reaction times (RT) the averages of the medians were computed for each single subject in each condition. Only correct responses within a time window starting 400 ms and ending 2400 ms after stimulus onset were included in the calculation of the RTs and error rates. Error rates were computed as the overall percentage of incorrect decisions in each condition. The RT and the error rate data were statistically evaluated by a 2 × 2 repeated measurements ANOVA with the factors Modality (unimodal/crossmodal) and Congruency (congruent/incongruent).

Results

Reaction Times

Analysis of the reaction times revealed a priming effect in the unimodal as well as in the crossmodal condition, as can

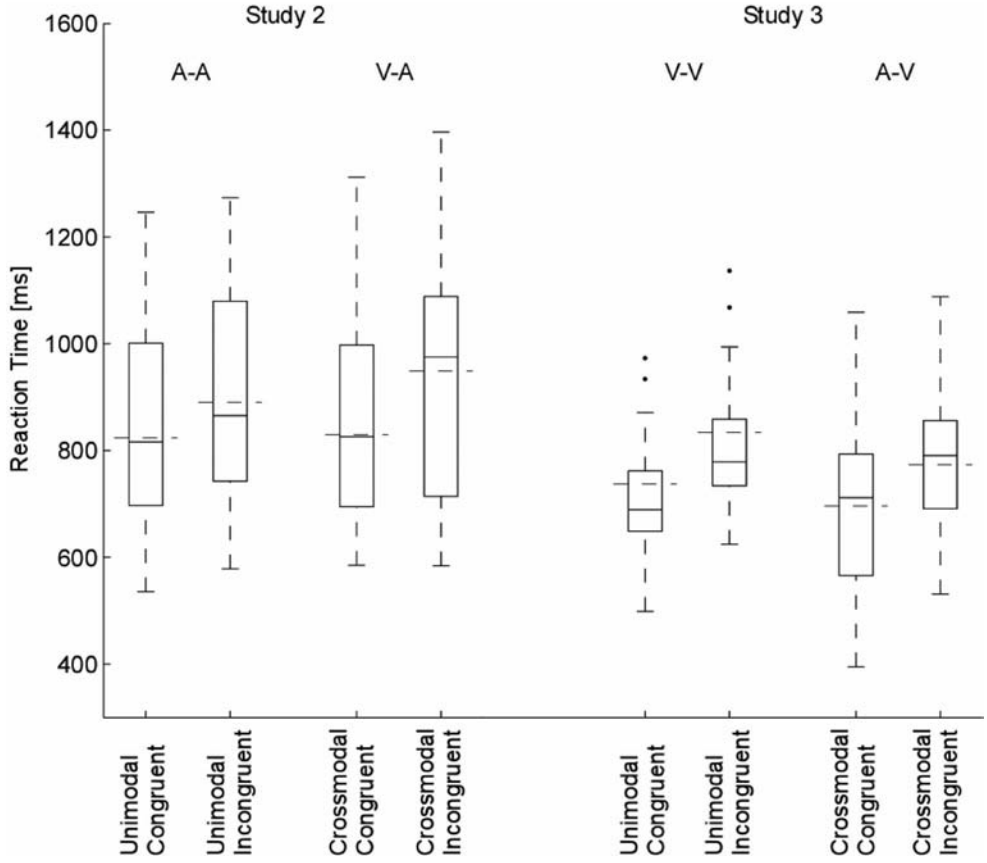

Figure 3. Boxplots of the reaction times in Study 2 and 3 showing median, lower, and upper quartile. The horizontal dashed line indicates the mean, the whiskers the range of values. In Study 2 targets were auditory stimuli (A) and in Study 3 targets were visual stimuli (V) preceded by either auditory or visual primes. A crossmodal behavioral priming effect was present in all four conditions (AA, VA, VV, AV) showing shorter reaction times for congruent compared to incongruent stimulus pairs.

be seen in Figure 3. The ANOVA revealed a significant main effect of Congruency, $F(1, 19) = 21.53$, $p < .001$, $\varepsilon = .53$, but no statistically significant main effect of Modality, $F(1, 19) = 1.63$, $p = .22$, $\varepsilon = .08$. No significant interaction between the two factors was found, $F(1, 19) = 3.73$, $p = .07$. In the unimodal condition the following contrast revealed that reaction times in congruent trials ($M = 824$ ms, $SD = 189$) were significantly shorter than in incongruent trials ($M = 890$ ms, $SD = 194$), $t(19) = 3.64$, $p < .01$. In the crossmodal condition the analysis revealed a similar result, showing shorter reaction times for congruent ($M = 830$ ms, $SD = 177$) than for incongruent trials ($M = 949$ ms, $SD = 278$), $t(19) = 4.09$, $p < .01$. In the unimodal condition an average of 0.95% of the trials were discarded due to too slow and 0.36% due to too fast responses, in the crossmodal condition 2.29% of the trials were discarded due to too slow and 1.56% due to too fast responses.

Error Rates

Analysis of error rates revealed a significant main effect of Congruency, $F(1, 19) = 132.06$, $p < .001$, $\varepsilon = .87$, and a significant interaction between Congruency × Modality, $F(1, 19) = 153.47$, $p < .001$, $\varepsilon = .89$, but no significant main effect of Modality was found, $F(1, 19) < 1$, $p = .36$. A paired t-test showed in the unimodal condition that error rates did not differ significantly between congruent ($M = 36.28\%$, $SD = 5.34$) and incongruent trials ($M = 34.15\%$, $SD = 7.36$), $t(19) = 1.26$; $p = .22$, whereas in the crossmodal condition a significantly lower error rate for congruent trials ($M = 20.18\%$, $SD = 5.12$) compared to incongruent trials ($M = 47.42\%$, $SD = 4.99$), $t(19) = 17.78$; $p < .001$, was present.

Discussion

A crossmodal priming paradigm was used in order to investigate, whether pairs of semantically congruent stimuli presented in two modalities activate the same semantic knowledge. We found facilitated auditory object identification expressed in shorter reaction times if target stimuli were preceded by semantically congruent visual objects. Additionally, error rates were lower in semantically congruent compared to semantically incongruent trials. In the

unimodal control condition, in which also congruent and incongruent stimulus pairs were presented, a similar facilitation effect was present in the reaction times, but not in the error rates.

Study 3

Methods

Participants

A new group of 32 students from the University Medical Center Hamburg volunteered and received monetary compensation. The data of four participants had to be excluded due to technical malfunction. None of the selected individuals (13 female and 15 male, mean age: 24.86, range: 20 to 31 years) had taken part in study 1 or 2. For 15 participants the EEG was recorded while running the experiment. All individuals were native German speakers, had normal or corrected-to-normal vision and normal hearing, and reported no history of neurological or psychic illness. The data of four subjects were discarded because in these more than 25% of the trials showed too long (> 2400 ms) reaction times.

Stimuli and Experimental Paradigm

Different levels of pixel noise were added to the original pictures, in order to create a visual stimulus set, which is comparable to the auditory stimuli in terms of identification rates. The original color photographs were transferred to gray level versions and the intensities of single pixels were randomly shifted with standard deviations of 200, 300, 400, 500 and 600, resulting in pictures with five different levels of pixel noise. In a pilot study these stimuli were presented successively starting with the highest level of pixel noise (600) and ending with the original version. Participants had to indicate by button press, whether they were able to identify each depicted object. Pictures identified by approximately 50% of the participants were selected for the following experiment.

The same experimental paradigm as in Study 1 was used (Figure 2B), with the exception that the modality in which S1 and S2 were presented was changed. That is, in the crossmodal condition S1 was presented acoustically and S2 was presented visually and in the unimodal condition both stimuli were presented visually. As in Study 2 incongruent stimulus pairs represented always different objects and mostly different categories, 7% of the stimulus pairs in the crossmodal and 20% in the unimodal condition belonged to the same category.

Results

Reaction Times

The ANOVA revealed a main effect of Congruency, $F(1, 23) = 65.77$, $p < .001$, $\varepsilon = .74$, but no main effect of Modality, $F(1, 23) = 2.65$, $p = .12$. There was no interaction between the two factors, $F(1, 23) < 1$; $p = .40$. In the unimodal condition (Figure 3), responses to congruent trials ($M = 737$ ms, $SD = 108$) were shorter compared to incongruent trials ($M = 834$ ms, $SD = 123$), $t(23) = 10.54$, $p < .001$. The same pattern appeared in the crossmodal condition (Figure 3), where reaction times were significantly shorter in congruent ($M = 696$ ms, $SD = 157$) compared to incongruent trials ($M = 774$ ms, $SD = 123$), $t(23) = 4.08$, $p < .001$. In the unimodal condition an average of 3.2% and the crossmodal condition an average of 7.7% of outlier trials were discarded.

Error Rates

Analysis of the error rates showed a main effect of Modality, $F(1, 23) = 12.89$, $p < .01$, $\varepsilon = .36$, and an interaction between the factors Modality and Congruency, $F(1, 23) = 6.29$, $p < .05$, $\varepsilon = .22$. No main effect of Congruency, $F(1, 23) = 2.51$, $p = .13$ was present in the data. In the crossmodal condition error rates were lower in response to congruent trials ($M = 27.81\%$, $SD = 9.86$) compared to incongruent trials ($M = 32.7\%$, $SD = 12.7$), $t(23) = 2.49$, $p < .05$. In the unimodal condition error rates were similar in response to congruent ($M = 34.97\%$, $SD = 8.32$) and incongruent trials ($M = 34.0\%$, $SD = 8.71$).

Discussion

A crossmodal priming effect was observed in the auditory-visual priming experiment. Visual object identification was facilitated by preceding semantically congruent auditory stimuli expressed in reaction times and error rates. This advantage in reaction times was of similar size in the unimodal control condition, in which a visual stimulus was presented as prime and target. As in Study 2 no facilitation effect was present in the unimodal condition regarding error rates.

General Discussion

In a natural environment, perception is fundamentally a multisensory process (Calvert et al., 2004). The parallel influx of information through several sensory pathways jointly contributes to the detection and identification of objects and to the selection of appropriate behavioral responses. One of the main goals of contemporary research on multi-

sensory processing is to better understand how the different modalities contribute to perception in humans (Calvert & Thesen, 2004). Accordingly, the study of cognitive processing subserving object perception will benefit from an approach where dynamic interactions are studied not only within a single sensory system, but between perceptual systems. In order to advance research in this field, we developed a multisensory stimulus series, including 270 auditory and 320 visual stimuli representing natural objects. Normative ratings for each of these stimuli were obtained for several psychological variables. We employed a crossmodal priming paradigm to investigate whether pairs of semantically congruent stimuli presented in two different sensory modalities are activating the same semantic knowledge. We found facilitated processing in auditory and visual objects expressed in reaction times in semantically congruent compared to incongruent stimulus pairs. That is, auditory and visual object identification was facilitated only if the stimuli were preceded by their semantically congruent counterpart in the other modality. The crossmodal facilitation effects expressed in reaction times were of similar size as the effects in the unimodal control conditions, which revealed facilitated identification of congruent stimulus pairs. Accordingly, the findings on the crossmodal priming effects strongly suggest that congruent visual and auditory items from the MULTIMOST series were indeed mostly perceived as reflecting the same object, demonstrating that these stimuli can be efficiently used to address current issues in multisensory processing.

The descriptive comparison between normative ratings for visual and auditory stimuli reveals some important characteristics of the stimuli. Firstly, it was clearly easier to identify visual as compared to auditory objects. In the present case, this may be partly explained by the natural differences between visual and auditory stimuli. Whereas objects in complex visual scenes can be clearly identified within less than 150 ms (Thorpe, Fize, & Marlot, 1996), the recognition of environmental sounds requires more time (Ballas, 1993; Debener et al., 2005). As a result, the reduction of auditory stimuli to a length of 400 ms limits the ability to correctly identify auditory objects. It is well known, from visual (Busch, Debener, Kranczioch, Engel, & Herrmann, 2004) and auditory (Michalewski, Starr, Nguyen, Kong, & Zeng, 2005; Hine & Debener, 2007) electrophysiological research, that both the onset and offset of stimulation evoke separate event-related neural responses. Accordingly, it was necessary to trim the environmental sounds to a constant duration. Secondly, the considerable differences observed between auditory and visual object identification seem to reflect inherent features of the respective sensory systems (Welch & Warren, 1980). In auditory perception, information on the object is conveyed over time. Environmental sounds clearly differ in the amount of information which is conveyed in 400 ms. Notwithstanding our efforts to capture the most informative interval of the environmental sounds, some objects evidently require longer stimulus duration in order to be correctly identified.

In Study 2, auditory object identification was facilitated following the presentation of congruent compared to incongruent visual objects in both performance measures. This result is in agreement with a recent study investigating visual-auditory priming using video sequences (Greene et al., 2001). The facilitation of visual object identification preceded by semantically congruent auditory objects in Study 3 contradicts the findings of Greene et al. (2001), where auditory stimuli did not influence the performance in visual object identification. This dissociation can be explained by the altered visual stimuli used in our experiment on auditory-visual priming. We degraded the quality of the visual stimuli and thus allowed information entered via the auditory channel to facilitate visual object identification.

Considerable differences between the two performance measures, reaction times and error rates, were evident. Interestingly, the analysis of error rates revealed an interaction between modality and congruency in both experiments. In both crossmodal conditions (visual-auditory, auditory-visual) facilitated processing of the target stimuli was observed following congruent stimuli, whereas in both unimodal conditions (auditory-auditory, visual-visual) no difference between congruent and incongruent trials was present. This result, however, is not surprising as compared to the crossmodal condition in which additional information is conveyed via the second sensory channel, in the unimodal condition information is simply repeated. Presenting the same stimulus twice does not convey much additional information, which could influence the accuracy of object identification. The dissociation between the results for reaction times and error rates in the unimodal condition should be interpreted keeping in mind the above mentioned differences between auditory and visual object stimuli.

In order to balance the identification difficulty of visual and auditory stimuli of the MULTIMOST series, we degraded the quality of the visual stimuli, by adding random noise to the visual stimuli. We obtained visual object stimuli which were on average as difficult to identify as the auditory stimuli. This approach was deliberately not considered an option for the initial normative study, as it is much easier to alter a high quality visual stimulus than to reverse physical normalization processes. Accordingly, we could show in Study 3 that auditory stimuli are effective primes for visual object stimuli, if the quality of the visual stimuli is degraded and identification rates of the visual stimuli are reduced.

At present it remains unclear at which level multisensory integration occurs. In the research of multisensory processing, three different views can be differentiated. The traditional view states that perceptual information is maintained exclusively in modality-specific perceptual systems, and integration takes place at higher cognitive processing stages. In contrast, an alternative account assumes that perceptual information is maintained in an amodal representation system independent of the input modality (Stoffregen & Bardy, 2001). A more intermediate view considers separate but interacting perceptual systems. Here, perceptual processing occurs in modality-specific areas, but information

is exchanged at very early stages of processing (Schroeder & Foxe, 2005; Stein & Meredith, 1993), possibly by binding of modality-specific neurons into multisensory representations by response synchronization (Engel, Fries, & Singer, 2001; Senkowski, Talsma, Herrmann, & Woldorff, 2005). The notion of early multisensory interaction is supported by numerous electrophysiological and neuroimaging studies reporting that perceptual systems influence each other selectively at early points of perceptual processing (Giard & Peronnet, 1999; Molholm, Ritter, Javitt, & Foxe, 2004; Schroeder & Foxe, 2004). The multimodal stimulus set may be instrumental for future research addressing on how multiple senses interact in concert to provide complex functions such as perception and memory.

Acknowledgments

We gratefully thank Kriemhild Saha for assistance in acquisition of participants, data recording and analysis and Peter Marquardt for help with the normalization of the auditory stimuli. This study was supported by grants from the Forschungsförderungsfond of the Medical Faculty, University Medical Center Hamburg-Eppendorf (S.D.), from the EU (A.K.E.) and the BMBF (A.K.E).

References

Amedi, A., von Kriegstein, K., van Atteveldt, N.M., Beauchamp, M.S., & Naumer, M.J. (2005). Functional imaging of human crossmodal identification and object recognition. *Experimental Brain Research, 166*, 559–571.

Ballas, J.A. (1993). Common factors in the identification of an assortment of brief everyday sounds. *Journal of Experimental Psychology: Human Perception and Performance, 19*, 250–267.

Beauchamp, M.S., Lee, K.E., Argall, B.D., & Martin, A. (2004). Integration of auditory and visual information about objects in superior temporal sulcus. *Neuron, 41*, 809–823.

Bradley, M.M., & Lang, P.J. (1999). *International affective digitized sounds (IADS): Stimuli, instruction manual and affective ratings* (Tech. Rep. No. B-2). Gainesville, FL: The Center for Research in Psychophysiology, University of Florida.

Busch, N.A., Debener, S., Kranczioch, C., Engel, A.K., & Herrmann, C.S. (2004). Size matters: effects of stimulus size, duration and eccentricity on the visual gamma-band response. *Clinical Neurophysiology, 115*, 1810–1820.

Calvert, G.A. (2001). Crossmodal processing in the human brain: Insights from functional neuroimaging studies. *Cerebral Cortex, 11*, 1110–1123.

Calvert, G.A., Spence, C., & Stein, B.E. (Eds.). (2004). *The handbook of multisensory processes*. Cambridge, MA: Massachusetts Institute of Technology.

Calvert, G.A., & Thesen, T. (2004). Multisensory integration: Methodological approaches and emerging principles in the human brain. *Journal of Physiology, Paris, 98*(1–3), 191–205.

Debener, S., Herrmann, C.S., Kranczioch, C., Gembris, D., & Engel, A.K. (2003). Top-down attentional processing enhances auditory evoked gamma band activity. *Neuroreport, 14*, 683–686.

Debener, S., Kranczioch, C., Herrmann, C.S., & Engel, A.K. (2002). Auditory novelty oddball allows reliable distinction of top-down and bottom-up processes of attention. *International Journal of Psychophysiology, 46*, 77–84.

Debener, S., Makeig, S., Delorme, A., & Engel, A.K. (2005). What is novel in the novelty oddball paradigm? Functional significance of the novelty P3 event-related potential as revealed by independent component analysis. *Brain Research. Cognitive Brain Research, 22*, 309–321.

Easton, R.D., Greene, A.J., & Srinivas, K. (1997). Transfer between vision and haptics: Memory for 2-D patterns and 3-D objects. *Psychonomic Bulletin and Review, 4*, 403–410.

Engel, A.K., Fries, P., & Singer, W. (2001). Dynamic predictions: Oscillations and synchrony in top-down processing. *Nature Reviews Neuroscience, 2*, 704–716.

Fabiani, M., Kazmerski, V.A., Cycowicz, Y.M., & Friedman, D. (1996). Naming norms for brief environmental sounds: Effects of age and dementia. *Psychophysiology, 33*, 462–475.

Gaeta, H., Friedman, D., & Hunt, G. (2003). Stimulus characteristics and task category dissociate the anterior and posterior aspects of the novelty P3. *Psychophysiology, 40*, 198–208.

Giard, M.H., & Peronnet, F. (1999). Auditory-visual integration during multimodal object recognition in humans: A behavioral and electrophysiological study. *Journal of Cognitive Neuroscience, 11*, 473–490.

Greene, A.J., Easton, R.D., & LaShell, L.S. (2001). Visual-auditory events: Cross-modal perceptual priming and recognition memory. *Consciousness and Cognition, 10*, 425–435.

Henson, R.N. (2003). Neuroimaging studies of priming. *Progress in Neurobiology, 70*(1), 53–81.

Hine, J., & Debener, S. (2007). Late auditory evoked potentials asymmetry revisited. *Clinical Neurophysiology, 118,* 1274–1285.

Lang, P.J., Bradley, M.M., & Cuthbert, B.N. (1998). Emotion, motivation, and anxiety: Brain mechanisms and psychophysiology. *Biological Psychiatry, 44*, 1248–1263.

Lang, P.J., Öhman, A., & Vaitl, D. (1988). *The international affective picture system* [Photographic slides]. Gainesville, FL: Center for Research in Psychophysiology, University of Florida.

McClelland, A.G.R., & Pring, L. (1991). An investigation of cross-modality effects in implicit and explicit memory. *The Quarterly Journal of Experimental Psychology A,* 43(1), 19–33.

Michalewski, H.J., Starr, A., Nguyen, T.T., Kong, Y.Y., & Zeng, F.G. (2005). Auditory temporal processes in normal-hearing individuals and in patients with auditory neuropathy. *Clinical Neurophysiology, 116*, 669–680.

Molholm, S., Ritter, W., Javitt, D.C., & Foxe, J.J. (2004). Multisensory visual-auditory object recognition in humans: A high-density electrical mapping study. *Cerebral Cortex, 14*, 452–465.

Newell, F.N. (2004). Cross-modal object recognition. In G.A. Calvert, C. Spence, & B.E. Stein (Eds.), *The handbook of multisensory processes* (pp. 123–139). Cambridge, MA: MIT Press.

Rossion, B., & Pourtois, G. (2004). Revisiting Snodgrass and Vanderwart's object set: The role of surface detail in basic-level object recognition. *Perception, 33*, 217–236.

Saygin, A.P., Dick, F., & Bates, E. (2005). An online task for contrasting auditory processing in the verbal and nonverbal domains and norms for younger and older adults. *Behavior Research Methods, 37*(1), 99–110.

Schroeder, C.E., & Foxe, J.J. (2004). Multisensory convergence in early cortical processing. In G.A. Calvert, C. Spence, & T.R. Stanford (Eds.), *The handbook of multisensory processes* (pp. 295–309). Cambridge, MA: Massachusetts Institute of Technology.

Schroeder, C.E., & Foxe, J. (2005). Multisensory contributions to low-level, "unisensory" processing. *Current Opinion Neurobiology, 15*, 454–458.

Senkowski, D., Talsma, D., Herrmann, C.S., & Woldorff, M.G. (2005). Multisensory processing and oscillatory gamma responses: Effects of spatial selective attention. *Experimental Brain Research, 166*, 411–426.

Shafiro, V., & Gygi, B. (2004). How to select stimuli for environmental sound research and where to find them. *Behavior Research Methods, Instruments and Computers, 36*, 590–598.

Snodgrass, J.G., & Vanderwart, M. (1980). A standardized set of 260 pictures: Norms for name agreement, image agreement, familiarity, and visual complexity. *Journal of Experimental Psychology: Human Learning and Memory, 6*, 174–215.

Stein, B.E., & Meredith, M.A. (1993). *The merging of the senses*. Cambridge, MA: MIT Press.

Stoffregen, T.A., & Bardy, B.G. (2001). On specification and the senses. *The Behavioral and Brain Sciences, 24*, 195–213.

Stuart, G.P., & Jones, D.M. (1995). Priming the identification of environmental sounds. *The Quarterly Journal of Experimental Psychology A, 48*, 741–761.

Stuart, G.P., & Jones, D.M. (1996). From auditory image to auditory percept: Facilitation through common processes? *Memory and Cognition, 24*, 296–304.

Thorpe, S., Fize, D., & Marlot, C. (1996). Speed of processing in the human visual system. *Nature, 381*, 520–522.

Tulving, E., & Schacter, D.L. (1990). Priming and human memory systems. *Science, 247*, 301–306.

Welch, R.B., & Warren, D.H. (1980). Immediate perceptual response to intersensory discrepancy. *Psychological Bulletin, 88*, 638–667.

Woods, A.T., & Newell, F.N. (2004). Visual, haptic and cross-modal recognition of objects and scenes. *Journal of Physiology, Paris, 98*(1–3), 147–159.

Zajonc, R.B. (1968). Attitudinal effects of mere exposure. *Journal of Personality and Social Psychology, Monograph, Supplement, 9*, 1–27.

Received April 3, 2006
Revision received March 9, 2007
Accepted March 12, 2007

Till R. Schneider

Department of Neurophysiology and Pathophysiology
University Medical Center Hamburg-Eppendorf
Martinistraße 52
D-20246 Hamburg
Germany
Tel. +49 40 42803-3188
Fax +49 40 42803-7752
E-mail t.schneider@uke.uni-hamburg.de

